

Schizophrenia risk from complex variation of complement component 4

Aswin Sekar¹⁻³, Allison R. Bialas^{4,5}, Heather de Rivera^{1,2}, Avery Davis^{1,2}, Timothy R. Hammond⁴, Nolan Kamitaki^{1,2}, Katherine Tooley^{1,2}, Jessy Presumey⁵, Matthew Baum¹⁻⁴, Vanessa Van Doren¹, Giulio Genovese^{1,2}, Samuel A. Rose², Robert E. Handsaker^{1,2}, Schizophrenia Working Group of the Psychiatric Genomics Consortium⁶, Mark J. Daly^{2,7}, Michael C. Carroll⁵, Beth Stevens^{2,4}, Steven A. McCarroll^{1,2}

¹ Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.

² Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

³ MD-PhD Program, Harvard Medical School, Boston, Massachusetts, USA

⁴ Department of Neurology, F.M. Kirby Neurobiology Center, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA.

⁵ Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, Massachusetts, USA.

⁶ A full list of collaborators is in Supplementary Information.

⁷ Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA.

Please address correspondence to mccarroll@genetics.med.harvard.edu.

Abstract

Schizophrenia is a heritable brain illness with unknown pathogenic mechanisms. Schizophrenia's strongest genetic association at a population level involves variation in the Major Histocompatibility Complex (MHC) locus, but the genes and molecular mechanisms accounting for this have been challenging to recognize. We show here that schizophrenia's association with the MHC locus arises in substantial part from many structurally diverse alleles of the complement component 4 (*C4*) genes. We found that these alleles promoted widely varying levels of *C4A* and *C4B* expression and associated with schizophrenia in proportion to their tendency to promote greater expression of *C4A* in the brain. Human C4 protein localized at neuronal synapses, dendrites, axons, and cell bodies. In mice, C4 mediated synapse elimination during postnatal development. These results implicate excessive complement activity in the development of schizophrenia and may help explain the reduced numbers of synapses in the brains of individuals affected with schizophrenia.

Introduction

Schizophrenia is a heritable psychiatric disorder involving impairments in cognition, perception and motivation that usually manifest late in adolescence or early in adulthood. The pathogenic mechanisms underlying schizophrenia are unknown, but observers have repeatedly noted pathological features involving excessive loss of gray matter^{1,2} and reduced numbers of synaptic structures on neurons³⁻⁵. While treatments exist for the psychotic symptoms of schizophrenia, there is no mechanistic understanding of, nor effective therapies to prevent or treat, the cognitive impairments and deficit symptoms of schizophrenia, its earliest and most constant features. An important goal in human genetics is to find the biological processes that underlie such disorders.

More than 100 loci in the human genome contain SNP haplotypes that associate with risk of schizophrenia⁶; the functional alleles and mechanisms at these loci remain to be discovered. By far the strongest such genetic relationship is schizophrenia's unexplained association with genetic markers across the Major Histocompatibility Complex (MHC) locus, which spans several megabases of chromosome 6⁶⁻¹⁰. The MHC locus is best known for its role in immunity, containing 18 highly polymorphic human leukocyte antigen (*HLA*) genes that encode a vast suite of antigen-presenting molecules. In some autoimmune diseases, genetic associations at the MHC locus arise from alleles of *HLA* genes^{11,12}; however, schizophrenia's association to the MHC is not yet explained.

Though the functional alleles that give rise to genetic associations have in general been challenging to find, the schizophrenia-MHC association has been particularly challenging, as schizophrenia's complex pattern of association to markers in the MHC locus spans hundreds of genes and does not correspond to the linkage disequilibrium (LD) around any known variant^{6,10}.

This prompted us to consider cryptic genetic influences that might generate unconventional genetic signals. The most strongly associated markers in several large case/control cohorts were near a complex, multi-allelic, and only partially characterized form of genome variation that affects the *C4* gene encoding complement component 4 (**Extended Data Fig. 1**). The association of schizophrenia to *CSMD1*^{6,10}, which encodes a regulator of *C4*¹³, further motivated us to consider *C4*.

Results

***C4* structures and MHC SNP haplotypes**

Human *C4* exists as two functionally distinct genes (isotypes), *C4A* and *C4B*; both vary in structure and copy number. One to three *C4* genes (*C4A* and/or *C4B*) are commonly present as a tandem array within the MHC class III region (**Fig. 1a, Extended Data Fig. 1g**)¹⁴⁻¹⁸. The protein products of *C4A* and *C4B* bind different molecular targets^{19,20}. *C4A* and *C4B* segregate in both long and short genomic forms (*C4AL*, *AS*, *BL* and *BS*), distinguished by the presence or absence (in intron 9) of a human endogenous retroviral (HERV) insertion that lengthens *C4* from 14 to 21 kb without changing the *C4* protein sequence¹⁶ (**Fig. 1b**).

We developed a way (**Extended Data Fig. 2**) to identify the “structural haplotypes” of *C4* – the copy number of *C4A* and *C4B* and the long/short (HERV) status of each *C4A* and *C4B* copy – present on 222 copies of human chromosome 6. Using droplet digital PCR (ddPCR), we found that genomes contained 0-5 *C4A* genes, 0-3 *C4B* genes, 1-5 long (L) *C4* genes, and 0-3 short (S) *C4* genes (**Extended Data Fig. 2a, b**). We also developed assays to determine the long/short status of each *C4A* and *C4B* gene copy (**Extended Data Fig. 2c**), thus revealing copy number of *C4AL*, *C4BL*, *C4AS*, and *C4BS* in each genome (**Methods**).

We analyzed inheritance in father-mother-offspring trios (**Extended Data Fig. 2d**) to identify the *C4A* and *C4B* contents of individual alleles (**Extended Data Fig. 2e**). We found that 4 common *C4* structural haplotypes (AL-BL, AL-BS, AL-AL, and BS) were collectively present on 90% of the 222 independent chromosomes sampled; 11 uncommon *C4* haplotypes comprised the other 10% (**Fig. 1c**).

The series of many SNP alleles along a genomic segment (the SNP haplotype) can be used to identify chromosomal segments that come from shared common ancestors. We identified the SNP haplotype(s) on which each *C4* locus structure was present (**Fig. 2**). The three most common *C4* locus structures were each present on multiple MHC SNP haplotypes (**Fig. 2**). For example, the *C4* AL-BS structure (frequency 31%) was present on five common haplotypes (frequencies 4%, 4%, 4%, 8%, and 6%) and many rare haplotypes (collective frequency 5%, **Fig. 2**). Reflecting this haplotype diversity, each of these *C4* structures exhibited real but only partial correlation to individual SNPs (**Extended Data Fig. 3**). The relationship between *C4* structures and SNP haplotypes was generally one-to-many: a *C4* structure might be present on many haplotypes, but a given SNP haplotype tended to have one characteristic *C4* structure (**Fig. 2**).

***C4* expression variation in the brain**

Since *C4A* and *C4B* vary in both copy number and *C4*-HERV status (**Fig. 1**), and because other HERVs can function as enhancers²¹⁻²³, *C4* variation might affect *C4* genes' expression. We assessed how *C4* structural variation related to RNA expression of *C4A* and *C4B* in eight panels of *post mortem* human adult brain samples (674 samples from 245 distinct donors in 3 cohorts, **Methods**).

The results of this expression analysis were consistent across all five brain regions

analyzed. First, RNA expression of *C4A* and *C4B* increased proportionally with copy number of *C4A* and *C4B* respectively (**Fig. 3a, b; Extended Data Fig. 4**). (These observations mirror earlier observations in human serum²⁴.) Second, expression levels of *C4A* were 2-3 times greater than expression levels of *C4B*, even after controlling for relative copy number in each genome (**Fig. 3c**). Third, copy number of the *C4*-HERV sequence increased the ratio of *C4A* to *C4B* expression ($p < 10^{-7}$, $p < 10^{-2}$, $p < 10^{-3}$) (**Fig. 3c, Extended Data Fig. 4**).

We used the above data to create genetic predictors of *C4A* and *C4B* expression levels in the brain (**Methods**). If *C4A* or *C4B* expression levels influence a phenotype, then the aggregate genetic predictor might associate to schizophrenia more strongly than individual variants do.

***C4* structural variation in schizophrenia**

Schizophrenia cases and controls from 22 countries have been analyzed genome-wide for SNPs, implicating the MHC locus as the strongest of more than 100 genome-wide-significant associations⁶. Our analysis above showed that long haplotypes defined by many SNPs carry characteristic *C4* alleles (**Fig. 2**), potentially making it possible to infer *C4* alleles by statistical imputation²⁵ from combinations of many SNPs. We used our 222 integrated haplotypes of MHC SNPs and *C4* alleles (**Fig. 2**) as reference chromosomes for imputation. We found that the four most common structural forms of the *C4A/C4B* locus (BS, AL-BS, AL-BL, and AL-AL) could be inferred with reasonably high accuracy (generally $0.70 < r^2 < 1.00$).

We then analyzed SNP data from 28,799 schizophrenia cases and 35,986 controls, from 40 cohorts in 22 countries contributing to the Psychiatric Genomics Consortium (PGC)⁶. We evaluated association to 7,751 SNPs across the extended MHC locus (chr6: 25-34Mb), to *C4*

structural alleles (**Fig. 1c**), and to *HLA* sequence polymorphisms imputed from the SNP data. We also predicted levels of *C4A* and *C4B* expression from the imputed *C4* structural alleles.

The association of schizophrenia to these genetic variants exhibited two prominent features (**Fig. 4a, b**). One feature involved a large set of similarly-associating SNPs spanning 2 Mb across the distal end of the extended MHC region. (Below we use this set's most strongly associating SNP, rs13194504, as its genetic proxy.) The other peak of association centered at *C4*, where schizophrenia associated most strongly with the genetic predictor of *C4A* expression levels ($p = 3.6 \times 10^{-24}$) (**Fig. 4a, Extended Data Fig. 5**). In the region near *C4* (chromosome 6, 31-33 Mb), the more strongly a SNP correlated with predicted *C4A* expression, the more strongly it associated with schizophrenia (**Fig. 4b, bottom panel**).

Although the variation at *C4* and in the distal extended MHC region associated to schizophrenia with similar strengths ($p = 3.6 \times 10^{-24}$ and 5.5×10^{-28} , respectively), their correlation with each other was low ($r^2 = 0.18$, **Fig. 4b**), suggesting that they reflect distinct genetic influences. Conditional analysis confirmed this: in analyses controlling for either rs13194504 or genetically predicted *C4A* expression, the other genetic variable still defined a genome-wide-significant association peak ($p = 7.8 \times 10^{-10}$ and 8.0×10^{-14} , **Fig. 4c, d**). Controlling for both genetic variables revealed a third association signal just proximal to the MHC locus (**Fig. 4e**) involving SNPs around *BAKI* and *SYNGAPI*, the latter of which encodes a major component of the postsynaptic density; *de novo* loss-of-function mutations in *SYNGAPI* associate with autism²⁶. In joint analysis, all three genetic signals remained significant ($p = 8.0 \times 10^{-14}$, 2.8×10^{-8} , and 1.7×10^{-8} , respectively) and no additional genome-wide significant signals remained in the MHC locus (**Fig. 4f**).

In some autoimmune diseases with genetic associations in the MHC locus, alleles of *HLA* genes associate more strongly than do other variants in the MHC locus, appearing to explain the associations^{11,12}. In contrast, in schizophrenia, classical *HLA* alleles associated to schizophrenia less strongly than other genetic variants in the MHC region did (**Extended Data Fig. 6**). We further considered the strongest schizophrenia associations to classical *HLA* alleles at distinct loci (involving *HLA-B*0801*, *HLA-DRB1*0301*, and *HLA-DQB1*02*); conditional analysis indicated that each could be explained by LD to the stronger signals at *C4* and rs13194504 (**Extended Data Fig. 7**).

If each *C4* allele affects schizophrenia risk via its effect on *C4A* expression, then this relationship should be visible across specific *C4* alleles. We measured schizophrenia risk levels for the common *C4* structural alleles (BS, AL-BS, AL-BL, and AL-AL); these alleles showed relative risks ranging from 1.00 to 1.27 (**Fig. 5a**). We also estimated (from the *post mortem* brain samples) the *C4A* expression levels generated by these four alleles (**Fig. 5b**). Schizophrenia risk and *C4A* expression levels yielded the same ordering of the *C4* allelic series (**Fig. 5a, b**).

We sought an even more stringent test. If this allelic series of relationships to schizophrenia risk (**Fig. 5a**) arises from *C4* locus structure – rather than from other genetic variation in the MHC locus – then a given *C4* structure should exhibit the same schizophrenia risk regardless of the MHC haplotype on which it appears. We measured the schizophrenia association of all 13 common combinations of *C4* structure and MHC SNP haplotype (**Fig. 5c**). Across this allelic series, each *C4* allele exhibited a characteristic level of schizophrenia risk, regardless of the haplotype on which it appeared (**Fig. 5c**).

***C4A* RNA expression in schizophrenia**

These genetic findings (**Fig. 5a,c**) predict that *C4A* expression might be elevated in brain tissue from schizophrenia patients. We measured *C4A* RNA expression levels in brain tissue from 35 schizophrenia patients and 70 individuals without schizophrenia. The median expression of *C4A* in brain tissues from schizophrenia patients was 1.4-fold greater ($p = 2 \times 10^{-5}$ by Mann-Whitney test; **Fig. 5d**) and was elevated in each of the five brain regions assayed (**Extended Data Fig. 8**). This relationship did not meaningfully change in analyses adjusted for age or *post mortem* interval. The relationship remained significant after correcting for the higher average *C4A* copy number among the brain donors affected with schizophrenia (1.3-fold greater, $p = 0.002$). Some earlier studies have also reported elevated levels of complement proteins in serum of schizophrenia patients^{27,28}.

C4 in the central nervous system

C4 is a critical component of the classical complement cascade, an innate-immune-system pathway that rapidly recognizes and eliminates pathogens and cellular debris. In the brain, other genes in the classical complement cascade have been implicated in the elimination or “pruning” of synapses²⁹⁻³¹.

To evaluate the distribution of C4 in human brain, we performed immunohistochemistry on sections of the prefrontal cortex and hippocampus. We observed C4+ cells in the gray and white matter, with the greatest number of C4+ cells detected in the hippocampus. Co-staining with cell-type-specific markers revealed C4 in subsets of NeuN⁺ neurons (**Fig. 6a**; antibody specificity further evaluated in **Extended Data Fig. 9a**) and a subset of astrocytes. Much of the C4 immunoreactivity was punctate (**Fig. 6b**), colocalizing with synaptic puncta identified by co-

immunostaining for the pre- and postsynaptic markers VGLUT1/2 and PSD95 (**Fig. 6b**). These results suggest that C4 is produced by, or deposited on, neurons and synapses.

To further characterize neuronal C4, we cultured human primary cortical neurons and evaluated C4 expression, localization and secretion. Neurons expressed *C4* mRNA and secreted C4 protein (**Extended Data Fig. 9c**). Neurons exhibited C4-immunoreactive puncta along their processes and cell bodies (**Fig. 6c, d**; antibody specificity further evaluated in **Extended Data Fig. 9b**). About 75% of C4 immunoreactivity localized to neuronal processes (**Fig. 6c**); of the C4 in neuronal processes, approximately 65% was observed in dendrites (MAP2+, NF+ processes) and 35% in axons (MAP2-, NF+ processes). Punctate C4 immunoreactivity was observed at 48% of structural synapses as defined by co-localized synaptotagmin and PSD-95 (**Fig. 6d**).

The association of increased C4 with schizophrenia (**Fig. 4, 5**), the presence of C4 at synapses (**Fig. 6b, d**), the involvement of other complement proteins in synapse elimination²⁹⁻³¹, and earlier reports of decreased synapse numbers in schizophrenia patients³⁻⁵, together suggested that C4 might work with other components of the classical complement cascade to promote synaptic pruning. To test this hypothesis, we moved to a mouse model. (*C4A* and *C4B* appear to have functionally specialized outside the rodent lineage, but the mouse genome contains a *C4* gene that shares features with both *C4A* and *C4B*, **Extended Data Fig. 10a, b**). Impairments in schizophrenia tend to affect higher cognitive functions and recently-expanded brain regions for which analogies in mice are uncertain³². However, waves of postnatal synapse elimination occur in many brain regions, and strong experimental models have been established in several mammalian visual systems in which synaptic projections from retinal ganglion cells (RGCs) onto thalamic relay neurons within the dorsal lateral geniculate nucleus (dLGN) of the visual thalamus undergo activity-dependent synaptic refinement^{29-31,33-35}. We found that *C4* RNA was expressed in the LGN and in RGCs purified from retina (**Extended Data Fig. 10c**).

In the immune system, C4 promotes C3 activation, allowing C3 to covalently attach onto its targets and promote their engulfment by phagocytic cells. In the developing mouse brain, C3 targets subsets of synapses and is required for synapse elimination by microglia, the principal CNS cells expressing receptors for complement^{29,30}. We found that in mice deficient in *C4*³⁶, C3 immunostaining in the dLGN was greatly reduced compared to WT littermates (**Fig. 7a, b**), with fewer synaptic inputs being C3-positive in the absence of C4 (**Fig. 7c**). These data demonstrate a role for C4 in complement deposition on synaptic inputs.

We then evaluated whether mice deficient in *C4* had defects in synaptic remodeling, as has been described for *C3*-deficient mice²⁹. Mice lacking functional *C4* exhibited greater overlap between RGC inputs from the two eyes ($p < 0.001$) than wild-type littermate controls, suggesting reduced synaptic pruning (**Fig. 7d; Extended Data Fig. 10d, e; Methods**). The degree of deficit in *C4*^{-/-} mice was similar to that previously reported for *C1q*^{-/-} and *C3*^{-/-} mice^{29,31}. Heterozygous *C4*^{+/-} mice, with one wild-type copy of *C4*, had an intermediate phenotype (**Fig. 7d**). These data provide direct evidence that C4 mediates synaptic refinement in the developing brain.

Discussion

We developed ways to analyze a complex form of genome structural variation (**Fig. 1-2**) and discovered that schizophrenia's association with variation in the MHC locus involves many common, structurally distinct *C4* alleles that affect expression of *C4A* and *C4B* in the brain; each allele associated with schizophrenia risk in proportion to its effect on *C4A* expression (**Fig. 3-5**). We found that C4 is expressed by neurons, localized to dendrites, axons, and synapses, and

secreted (**Fig. 6**); and that C4 promotes synapse elimination during the developmentally timed maturation of a neuronal circuit (**Fig. 7**).

In humans, adolescence and early adulthood bring extensive elimination of synapses in distributed association regions of cerebral cortex, such as the prefrontal cortex, that have greatly expanded in recent human evolution³⁷⁻⁴⁰. Synapse elimination in human association cortex appears to continue from adolescence into the third decade of life³⁹. This late phase of cortical maturation, which may distinguish humans even from some other primates³⁷, corresponds to the period during which schizophrenia most often becomes clinically apparent and patients' cognitive function declines, a temporal correspondence that others have also noted⁴¹.

Principal pathological findings in schizophrenia brains involve loss of cortical gray matter without cell death: affected individuals exhibit abnormal cortical thinning^{1,2} and abnormally reduced numbers of synaptic structures on cortical pyramidal neurons³⁻⁵. The possibility that neuron-microglia interactions via the complement cascade contribute to schizophrenia pathogenesis – for example, that schizophrenia arises or intensifies from excessive or inappropriate synaptic pruning during adolescence and early adulthood – would offer a potential mechanism for these longstanding observations about age of onset and synapse loss. Many other genetic findings in schizophrenia involve genes that encode synaptic proteins^{6,42-44}. Diverse synaptic abnormalities might interact with the complement system and other pathways^{45,46} to cause excessive stimulation of microglia and/or elimination of synapses.

The two human *C4* genes (*C4A* and *C4B*) exhibited distinct relationships with schizophrenia risk, with increased risk associating most strongly with variation that increases expression of *C4A*. Human *C4A* and *C4B* proteins, whose functional specialization appears to be evolutionarily recent (**Extended Data Fig. 10a**), show striking biochemical differences: *C4A*

more readily forms amide bonds with proteins, while C4B favors binding to carbohydrate surfaces^{19,20}, differences with an established basis in C4 protein sequence and structure^{47,48}. An intriguing possibility is that C4A and C4B differ in affinity for an unknown binding site at synapses.

To date, few GWAS associations have been explained by specific functional alleles. An unexpected finding at *C4* involves the large number of common, functionally distinct forms of the same locus that appear to contribute to schizophrenia risk. The human genome contains hundreds of other genes with complex, multi-allelic forms of structural variation⁴⁹. It will be important to learn the extent to which such variation contributes to brain diseases and indeed to all human phenotypes.

Figure 1. Structural variation of the complement component 4 (*C4*) gene

(a) Location of the *C4* genes within the Major Histocompatibility Complex (MHC) locus on human chromosome 6.

(b) Human *C4* exists as two paralogous genes (isotypes), *C4A* and *C4B*; the encoded proteins are distinguished at a key site that determines which molecular targets they bind^{19,20}. Both *C4A* and *C4B* also exist in both long (L) and short (S) forms distinguished by an endogenous retroviral (*C4*-HERV) sequence in intron 9.

(c) Structural forms of the *C4* locus and their frequencies among a European-ancestry population sample (222 chromosomes from 111 genetically unrelated individuals, HapMap CEU), inferred as described in **Extended Data Fig. 2**. Asterisks indicate allele frequencies too low to be well-estimated.

Figure 2. Haplotypes formed by *C4* structures and SNPs

SNP haplotype(s) on which common *C4* structures were present. Each thin horizontal line represents the series of SNP alleles (haplotype) along a 250-kilobase chromosomal segment. Each column represents a SNP; gray and black indicate which allele is present on each haplotype. The SNP haplotypes are grouped into 13 sets of haplotypes associating with each of the four most common *C4* structures. Three *C4* structures (AL-BS, AL-BL, and AL-AL) each segregated on multiple SNP haplotypes (numbered at right).

Figure 3. Brain RNA expression of *C4A* and *C4B* in relation to copy numbers of *C4A*, *C4B*, and the *C4*-HERV

mRNA expression of *C4A* (a) and *C4B* (b) was measured (by ddPCR) in brain tissue from 244 individuals. Copy number of *C4A*, *C4B*, and the *C4*-HERV were measured (by ddPCR analysis of genomic DNA) in the brain donors. The results were consistent across 8 panels of brain tissue representing 5 brain regions and 3 distinct sets of donors (one set shown here, with data from 101 individuals; all panels in **Extended Data Fig. 4**; a few outlier points are beyond the range of these plots but are shown in **Extended Data Fig. 4**.) P-values were obtained by a Spearman rank correlation test. In panel c, expression of *C4A* (per genomic copy) is normalized to expression of *C4B* (per genomic copy) to control for trans-acting influences shared by *C4A* and *C4B*.

Figure 4. Association of schizophrenia to *C4* and the extended MHC locus

Association of schizophrenia to 7,751 SNPs across the MHC locus and to genetically predicted expression levels of *C4A* and *C4B* in the brain (represented in the genomic location of the *C4* gene). The data shown are based on analysis of 28,799 schizophrenia cases and 35,986 controls of European ancestry from the Psychiatric Genomics Consortium. The height of each point represents the statistical strength ($-\log_{10}(p)$) of association with schizophrenia.

(a, b) Association of schizophrenia to SNPs in the MHC locus and to genetically predicted expression of *C4A* and *C4B*. In **(b)**, genetic variants are colored by their levels of correlation to rs13194504 (upper panel) or by their levels of correlation to genetically predicted brain *C4A* expression levels (lower panel).

(c-f) Conditional association analysis. The red dashed line indicates the statistical threshold for genome-wide significance ($p = 5 \times 10^{-8}$).

See also **Extended Data Fig. 5-7** for detailed association analyses involving *C4* locus structures and *HLA* alleles.

Figure 5. C4 structures, C4A expression, and schizophrenia risk

(a) Schizophrenia risk associated with four common structural forms of *C4* in analysis of 28,799 schizophrenia cases and 35,986 controls.

(b) Brain *C4A* RNA expression levels associated with four common structural forms of *C4*. β was calculated from fitting *C4A* RNA expression (in brain tissue) to the number of chromosomes (0, 1, or 2) carrying each *C4* structure (across 120 individuals sampled).

(c) Schizophrenia risk associated with 13 combinations of *C4* structural allele and MHC SNP haplotype. The numbers on the y-axis adjacent to the *C4* structures indicate the “haplogroup”, the MHC SNP haplotype background on which the *C4* structure segregates, and correspond to **Fig. 2**. Statistical tests of heterogeneity yielded $p = 0.55$ for *AL-AL* alleles; $p = 0.93$ for *AL-BL* alleles; $p = 0.06$ for *AL-BS* alleles; and $p = 5.7 \times 10^{-5}$ across the overall allelic series.

(d) Expression levels of *C4A* RNA were directly measured (by RT-ddPCR) in *post mortem* brain samples from 35 schizophrenia patients and 70 individuals not affected with schizophrenia.

Measurements for all five brain regions analyzed exhibited the same relationship (**Extended Data Fig. 8**). Horizontal lines show the median value for each group. P-values were derived by a (non-parametric) one-sided Mann-Whitney test.

Error bars shown in **a-c** represent 95% confidence intervals around the effect size estimate.

Figure 6. C4 protein at neuronal cell bodies, processes and synapses

(a) C4 protein localization in human brain tissue. Two representative confocal images (drawn from immunohistochemistry performed on samples from five individuals with schizophrenia and two unaffected individuals) within the hippocampal formation demonstrate localization of C4 in a subset of NeuN⁺ neurons.

(b) High-resolution structured illumination microscopy (SIM) imaging of tissue in the hippocampal formation reveals colocalization of C4 with the presynaptic terminal marker Vglut1/2 and the postsynaptic marker PSD95.

(c) Confocal images of primary human cortical neurons show colocalization of C4, MAP2, and neurofilament along neuronal processes.

(d) Confocal image of primary cortical neurons stained for C4, presynaptic marker synaptotagmin, and postsynaptic marker PSD95.

Scale bar for **a**, **c**, and **d** = 25 μm ; **b** = 5 μm ; **b (inset)** = 1 μm . **Extended Data Fig. 9** contains additional data on antibody specificity

Figure 7. *C4* in retinogeniculate synaptic refinement

(a) Representative confocal images of immunohistochemistry for C3 in the P5 dLGN showed reduced C3 deposition in the dLGN of *C4*^{-/-} mice compared to WT littermates.

(b) Quantification confirmed reduced C3 immunoreactivity in the dLGN (N = 3 mice/group, $p < 0.05$, *t*-test; y-axis: mean fluorescence intensity, normalized to WT).

(c) Co-localization analysis revealed a reduction in the fraction of VGLUT2+ puncta that were C3+ in *C4*-deficient mice relative to their WT littermates (N = 3 mice/group, $p = 0.0011$, two-sided *t*-test).

(d) Synaptic refinement in mice with 0, 1, or 2 copies of *C4*. These images represent the segregation of ipsilateral and contralateral RGC projections to the dLGN; two analysis methods were used. (Top) Projections from the ipsilateral (green) and contralateral (red) eyes show minimal overlap (yellow) in WT mice. The overlapping area is significantly increased in *C4*^{-/-} mice (N = 6 mice/group, $p < 0.01$, ANOVA with Bonferroni post tests). (Bottom) Threshold-independent analysis using the *R*-value⁵⁰ ($R = \log_{10}[F_{\text{ipsi}}/F_{\text{contra}}]$). Pixels are pseudocolored with an *R*-value heat map (red indicates areas having only contralateral inputs; purple, only ipsilateral inputs). Compared to their WT littermates, *C4*-deficient mice exhibited lower *R*-value variance, indicating defects in synaptic refinement (N = 6 mice/group, $p < 0.001$, ANOVA with Bonferroni post tests). Control experiments analyzing total dLGN size, dLGN area receiving ipsilateral input, and number of RGCs are shown in **Extended Data Fig. 10f-h**, respectively.

Error bars in (b), (c), and (d) represent S.E.M.

Acknowledgements. We thank Steve Hyman, Eric Lander, Cori Bargmann, and Chris Patil for helpful conversations about the project and comments on drafts of the manuscript; Maree Webster for expert advice on immunohistochemistry; Brian Browning for expert advice on imputation; the Stanley Medical Research Institute Brain Collection and the NHGRI Gene and Tissue Expression (GTEx) Project for access to RNA and tissue samples; Cheta Emba for assistance with experiments; and Christina Usher for contributions to manuscript figures. This work was supported by R01 HG 006855 (to S.A.M), by the Stanley Center for Psychiatric Research (to S.A.M and B.S.), by R01 MH077139 (to the PGC), and by T32 GM007753 (to A.S. and M.B.).

Author contributions. S.A.M. and A.S. conceived the genetic studies. A.S. performed the laboratory experiments and computational analyses to understand the molecular and population genetics of the *C4* locus (Fig.1 and Fig. 2). A.S., K.T., N.K., and V.V.D. analyzed *C4* expression variation in human brain (Fig. 3, Fig 5b,d.). G.G., R.E.H., and S.A.R. contributed to genetic analyses. A.S. and A.D. did the imputation and association analysis (Fig. 4., and Fig 5a, c). M.J.D. provided valuable advice on the association analyses. Investigators in the Schizophrenia Working Group of the Psychiatric Genomics Consortium collected and phenotyped cohorts and contributed genotype data for analysis. B.S. and M.C contributed expertise and reagents for experiments described in Fig. 6 and 7. H.d.R and T.H. performed the *C4* immunocytochemistry and immunohistochemistry experiments respectively, with valuable advice from A.B. (Fig. 6). A.B. and J.P. analyzed *C4*'s role in synaptic refinement in the mouse visual system (Fig. 7). M.B. analyzed *C4* expression in mice. S.A.M and A.S. wrote the manuscript with contributions from all authors.

Author information. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to mccarroll@genetics.med.harvard.edu.

References

- 1 Cannon, T. D. *et al.* Cortex mapping reveals regionally specific patterns of genetic and disease-specific gray-matter deficits in twins discordant for schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 3228-3233, doi:10.1073/pnas.052023499 (2002).
- 2 Cannon, T. D. *et al.* Progressive reduction in cortical thickness as psychosis develops: a multisite longitudinal neuroimaging study of youth at elevated clinical risk. *Biological psychiatry* **77**, 147-157, doi:10.1016/j.biopsych.2014.05.023 (2015).
- 3 Garey, L. J. *et al.* Reduced dendritic spine density on cerebral cortical pyramidal neurons in schizophrenia. *J Neurol Neurosurg Psychiatry* **65**, 446-453 (1998).
- 4 Glantz, L. A. & Lewis, D. A. Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia. *Arch Gen Psychiatry* **57**, 65-73 (2000).
- 5 Glausier, J. R. & Lewis, D. A. Dendritic spine pathology in schizophrenia. *Neuroscience* **251**, 90-107, doi:10.1016/j.neuroscience.2012.04.044 (2013).
- 6 Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).
- 7 Shi, J. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753-757, doi:10.1038/nature08192 (2009).
- 8 Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744-747, doi:10.1038/nature08186 (2009).
- 9 International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752, doi:10.1038/nature08185 (2009).
- 10 Schizophrenia Psychiatric Genome-Wide Association Study Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nature genetics* **43**, 969-976, doi:10.1038/ng.940 (2011).
- 11 Howson, J. M., Walker, N. M., Clayton, D. & Todd, J. A. Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A. *Diabetes Obes Metab* **11 Suppl 1**, 31-45, doi:10.1111/j.1463-1326.2008.01001.x (2009).
- 12 Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nature genetics* **44**, 291-296, doi:10.1038/ng.1076 (2012).
- 13 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 journal : official publication of the Federation of American Societies for Experimental Biology* **27**, 5083-5093, doi:10.1096/fj.13-230706 (2013).
- 14 Carroll, M. C., Campbell, R. D., Bentley, D. R. & Porter, R. R. A molecular map of the human major histocompatibility complex class III region linking complement genes C4, C2 and factor B. *Nature* **307**, 237-241 (1984).
- 15 Carroll, M. C., Belt, T., Palsdottir, A. & Porter, R. R. Structure and organization of the C4 genes. *Philos Trans R Soc Lond B Biol Sci* **306**, 379-388 (1984).
- 16 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).

- 17 Horton, R. *et al.* Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* **60**, 1-18, doi:10.1007/s00251-007-0262-2 (2008).
- 18 Banlaki, Z., Doleschall, M., Rajczyk, 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, doi:10.1038/gene.2012.29 (2012).
- 19 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).
- 20 Isenman, D. E. & Young, J. R. The molecular basis for the difference in immune hemolysis activity of the Chido and Rodgers isotypes of human complement component C4. *J Immunol* **132**, 3019-3027 (1984).
- 21 Illarionova, A. E., Vinogradova, T. V. & Sverdlov, E. D. Only those genes of the KIAA1245 gene subfamily that contain HERV(K) LTRs in their introns are transcriptionally active. *Virology* **358**, 39-47, doi:10.1016/j.virol.2006.06.027 (2007).
- 22 Nakamura, A., Okazaki, Y., Sugimoto, J., Oda, T. & Jinno, Y. Human endogenous retroviruses with transcriptional potential in the brain. *Journal of human genetics* **48**, 575-581, doi:10.1007/s10038-003-0081-8 (2003).
- 23 Suntsova, M. *et al.* Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 19472-19477, doi:10.1073/pnas.1318172110 (2013).
- 24 Yang, Y. *et al.* Diversity in intrinsic strengths of the human complement system: serum C4 protein concentrations correlate with C4 gene size and polygenic variations, hemolytic activities, and body mass index. *J Immunol* **171**, 2734-2745 (2003).
- 25 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, doi:10.1086/521987 (2007).
- 26 Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221, doi:10.1038/nature13908 (2014).
- 27 Mayilyan, K. R., Arnold, J. N., Presanis, J. S., Soghoyan, A. F. & Sim, R. B. Increased complement classical and mannan-binding lectin pathway activities in schizophrenia. *Neurosci Lett* **404**, 336-341, doi:10.1016/j.neulet.2006.06.051 (2006).
- 28 Hakobyan, S., Boyajyan, A. & Sim, R. B. Classical pathway complement activity in schizophrenia. *Neurosci Lett* **374**, 35-37, doi:10.1016/j.neulet.2004.10.024 (2005).
- 29 Stevens, B. *et al.* The classical complement cascade mediates CNS synapse elimination. *Cell* **131**, 1164-1178, doi:10.1016/j.cell.2007.10.036 (2007).
- 30 Schafer, D. P. *et al.* Microglia sculpt postnatal neural circuits in an activity and complement-dependent manner. *Neuron* **74**, 691-705, doi:10.1016/j.neuron.2012.03.026 (2012).
- 31 Bialas, A. R. & Stevens, B. TGF-beta signaling regulates neuronal C1q expression and developmental synaptic refinement. *Nat Neurosci* **16**, 1773-1782, doi:10.1038/nn.3560 (2013).
- 32 Kaiser, T. & Feng, G. Modeling psychiatric disorders for developing effective treatments. *Nat Med* **21**, 979-988, doi:10.1038/nm.3935 (2015).
- 33 Shatz, C. J. & Kirkwood, P. A. Prenatal development of functional connections in the cat's retinogeniculate pathway. *J Neurosci* **4**, 1378-1397 (1984).
- 34 Sretavan, D. W. & Shatz, C. J. Prenatal development of retinal ganglion cell axons: segregation into eye-specific layers within the cat's lateral geniculate nucleus. *J Neurosci* **6**, 234-251 (1986).
- 35 Chen, C. & Regehr, W. G. Developmental remodeling of the retinogeniculate synapse. *Neuron* **28**, 955-966 (2000).

- 36 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).
- 37 Huttenlocher, P. R. & Dabholkar, A. S. Regional differences in synaptogenesis in human
cerebral cortex. *J Comp Neurol* **387**, 167-178 (1997).
- 38 Huttenlocher, P. R. Synaptic density in human frontal cortex - developmental changes
and effects of aging. *Brain Res* **163**, 195-205 (1979).
- 39 Petanjek, Z. *et al.* Extraordinary neoteny of synaptic spines in the human prefrontal
cortex. *Proceedings of the National Academy of Sciences of the United States of America*
108, 13281-13286, doi:10.1073/pnas.1105108108 (2011).
- 40 Buckner, R. L. & Krienen, F. M. The evolution of distributed association networks in the
human brain. *Trends Cogn Sci* **17**, 648-665, doi:10.1016/j.tics.2013.09.017 (2013).
- 41 Feinberg, I. Schizophrenia: caused by a fault in programmed synaptic elimination during
adolescence? *Journal of psychiatric research* **17**, 319-334 (1982).
- 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,
doi:10.1038/mp.2011.154 (2012).
- 43 Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks.
Nature **506**, 179-184, doi:10.1038/nature12929 (2014).
- 44 Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia.
Nature **506**, 185-190, doi:10.1038/nature12975 (2014).
- 45 Datwani, A. *et al.* Classical MHCI molecules regulate retinogeniculate refinement and
limit ocular dominance plasticity. *Neuron* **64**, 463-470, doi:10.1016/j.neuron.2009.10.015
(2009).
- 46 Lee, H. *et al.* Synapse elimination and learning rules co-regulated by MHC class I H2-
Db. *Nature* **509**, 195-200, doi:10.1038/nature13154 (2014).
- 47 van den Elsen, J. M. *et al.* X-ray crystal structure of the C4d fragment of human
complement component C4. *J Mol Biol* **322**, 1103-1115 (2002).
- 48 Dodds, A. W., Ren, X. D., Willis, A. C. & Law, S. K. The reaction mechanism of the
internal thioester in the human complement component C4. *Nature* **379**, 177-179,
doi:10.1038/379177a0 (1996).
- 49 Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nature
genetics* **47**, 296-303, doi:10.1038/ng.3200 (2015).
- 50 Torborg, C. L. & Feller, M. B. Unbiased analysis of bulk axonal segregation patterns. *J
Neurosci Methods* **135**, 17-26, doi:10.1016/j.jneumeth.2003.11.019 (2004).

Extended Data Figure 2

- 51 Fernando, M. M. *et al.* Assessment of complement C4 gene copy number using the
paralog ratio test. *Hum Mutat* **31**, 866-874, doi:10.1002/humu.21259 (2010).

Extended Data Figure 5

- 52 Rudduck, C., Beckman, L., Franzen, G., Jacobsson, L. & Lindstrom, L. Complement
factor C4 in schizophrenia. *Hum Hered* **35**, 223-226 (1985).
- 53 Schroers, R. *et al.* Investigation of complement C4B deficiency in schizophrenia. *Hum
Hered* **47**, 279-282 (1997).
- 54 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,
doi:10.1080/15622970701227803 (2008).

Extended Data Figure 7

- 55 Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS
One* **8**, e64683, doi:10.1371/journal.pone.0064683 (2013).

Extended Data Figure 10

- 56 Nonaka, M., Nakayama, K., Yeul, Y. D. & Takahashi, M. Complete nucleotide and derived amino acid sequences of sex-limited protein (Slp), nonfunctional isotype of the fourth component of mouse complement (C4). *J Immunol* **136**, 2989-2993 (1986).

(Note that all references after #50 are cited for the first time in the legends of Extended Data Figures.)

Extended Data Figure 1. Association of schizophrenia to common variants in the MHC locus in individual case-control cohorts, and schematic of the repeat module containing *C4*.

(a-f) Data for several schizophrenia case-control cohorts that were genome-scanned before we began this work (a-d) exhibits peaks of association near chr6:32Mb (blue vertical line) on the human genome reference sequence (GRCh37/hg19). Note that association patterns vary from cohort to cohort, reflecting statistical sampling fluctuations and potentially fluctuations in allele frequencies of the (unknown) causal variants in different cohorts. Cohorts such as in (b), (e) and (f) suggest the existence of effects at multiple loci within the MHC region. Even in the cohorts with simpler peaks (a, c, d), the pattern of association across the individual SNPs at chr6:32 Mb does not correspond to the linkage disequilibrium (LD) around any known variant. This motivated the focus in the current work on cryptic genetic influences in this region that could cause unconventional association signals that do not resemble the LD patterns of individual variants.

(g) A complex form of genome structural variation resides near chr6:32 Mb. Shown here are three of the known alternative structural forms of this genomic region. The most prominent feature of this structural variation is the tandem duplication of a genomic segment that contains a *C4* gene, 3' fragments of the *STK19* and *TNXB* genes, and a pseudogenized copy of the *CYP21A2* gene. (This cassette is present in 1-3 copies on the three alleles depicted above; the boundaries below each haplotype demarcate the sequence that is duplicated.) Haplotypes with multiple copies of this module (middle and bottom) contain multiple functional copies of *C4*, whereas the additional gene fragments or copies denoted *STK19P*, *CYP21A2P*, and *TNXA* are typically pseudogenized. (Rare haplotypes with a gain or loss of intact *CYP21A2* have also been observed¹⁸.) Note that although *C4A* and *C4B* contain multiple sequence variants, they are defined based on the differences encoded by exon 26, which determine the relative affinities of C4A and C4B for distinct molecular targets^{19,20} (Fig. 1). Many additional forms of this locus appear to have arisen by non-allelic homologous recombination and gene conversion (ref¹⁸ and Fig. 1).

Extended Data Figure 2. Schematic of strategy for identifying the segregating structural forms of the *C4* locus. **(a)** Molecular assays for measuring copy number of the key, variable *C4* structural features – the length polymorphism (HERV insertion) that distinguishes the long (L) from the short (S) genomic form of *C4*, and the *C4A/C4B* isotypic difference. Each primer-probe assay is represented with the combination of arrows (primers) and asterisk (probe) in its approximate genomic location (though not to scale). **(b)** Measurement of copy number of *C4* gene types in the genomes of 162 individuals (from HapMap CEU sample). The absolute, integer copy number of each *C4* gene type in each genome is precisely inferred from the resulting data. To ensure high accuracy, the data are further evaluated for a checksum relationship ($A + B = L + S$) and for concordance with earlier data from Southern blotting of 89 of the same HapMap individuals⁵¹. **(c)** To measure the copy number of compound structural forms of *C4* (involving combinations of L/S and A/B), we perform long-range PCR followed by quantitative measurement of the A/B isotype-distinguishing sequences in droplets. **(d)** Analysis of transmissions in father-mother-offspring trios enables inference of the *C4* gene contents of individual copies (alleles) of chromosome 6. Three example trios are shown in this schematic. **(e)** Examples of the inferred structural forms of the *C4* locus (more shown in **Fig. 1c**). For the common *C4* structures (AL-BL, AL-BS, AL-AL, and BS), genomic order of the *C4* gene copies is known from earlier assemblies of sequence contigs in individuals homozygous for MHC haplotypes due to consanguinity¹⁷ and other molecular analyses of the *C4* locus¹⁸. For the rarer *C4* structures, genomic order of *C4* gene copies is hypothesized or provisional.

Extended Data Figure 3. Linkage disequilibrium relationships (r^2) of MHC SNPs to forms of *C4* structural variation. Correlations of SNPs in the MHC locus with (a) copy number of *C4* gene types and (b) larger-scale structural forms (haplotypes) of the *C4* locus. Dashed, vertical lines indicate the genomic location of the *C4* locus. Note that *C4* structural forms show only partial correlation (r^2) to the allelic states of nearby SNPs, reflecting the relationship shown in **Fig. 2**, in which a structural form of the *C4* locus often segregates on multiple different SNP haplotypes.

Extended Data Figure 4. RNA expression of *C4A* and *C4B* in relation to copy number of *C4A*, *C4B*, and the *C4*-HERV (long form of *C4*), in eight panels of *post mortem* brain tissue. Copy number of *C4* structural features was measured by ddPCR; RNA expression levels were measured by RT-ddPCR. Panels **a-e** show data for tissues from the Stanley Medical Research Institute (SMRI) Array Consortium and consist of **(a)** anterior cingulate cortex, **(b)** cerebellum, **(c)** corpus callosum, **(d)** orbital frontal cortex, and **(e)** parietal cortex. Panel **f** shows data for the frontal cortex samples from the NHGRI Genes and Tissues Expression (GTEx) Project. Panels **g** and **h** show data for tissues from the SMRI Neuropathology Consortium (anterior cingulate cortex and cerebellum, respectively). These data were then used to inform (by linear regression) the derivation of a linear model for predicting each individual's RNA expression of *C4A* and *C4B* as a function of the numbers of copies of AL, BL, AS, and BS. The derivation of this model, and the regression coefficients induced, are described in **Methods**.

In the rightmost plot of each panel, expression of *C4A* (per genomic copy) is normalized to expression of *C4B* (per genomic copy) to more specifically visualize the effect of the *C4*-HERV by controlling for genomic copy number and for any trans-acting influences shared by *C4A* and *C4B*; the inferred regression coefficients (**Methods**) suggest that the observed effect is mostly due to increased expression of *C4A*.

Extended Data Figure 5. Detailed analysis of the association of schizophrenia to genetic variation at and around *C4*, in data from 28,799 schizophrenia cases and 35,986 controls (Psychiatric Genomics Consortium, ref⁶). SCZ, schizophrenia; β , estimated effect size per copy of the genomic feature or allele indicated; SE, standard error. Detailed association analyses of *HLA* alleles are in **Extended Data Fig. 6-7**. (*) We specifically tested *C4B*-null status because a 1985 study⁵² reported an analysis of 165 schizophrenia patients and 330 controls in which rare *C4B*-null status associated with elevated risk of schizophrenia, though two subsequent studies^{53,54} found no association of schizophrenia to *C4B*-null genotype. We sought to evaluate this using the large data set in this study, finding no association to *C4B*-null status. (**) Total copy number of *C4* is also strongly correlated to copy number of the *CYP21A2P* pseudogene, which is present on duplicated copies of the sequence shown in **Extended Data Fig. 1g**.

Extended Data Figure 6. Evaluation of the association of schizophrenia with *HLA* alleles and coding-sequence polymorphisms. **(a-e)** Associations to *HLA* alleles and coding-sequence polymorphisms are shown in black; to provide the context of levels of association to nearby SNPs, associations to other SNPs are shown in gray. The series of conditional analyses shown **(b-e)** parallels the analyses in **Fig. 4**. Further detail on the most strongly associating *HLA* alleles (including conditional association analysis) is provided in **Extended Data Fig. 7**.

Extended Data Figure 7. Detailed association analysis for the most strongly associating classical *HLA* alleles. The most strongly associating *HLA* loci were *HLA-B* (in primary analyses, **Fig. 4a, Extended Data Fig. 6a**) and *HLA-DRB1* and *-DQB1* (in analyses controlling for the signal defined by rs13194504, **Fig. 4c, Extended Data Fig. 6b**). At these loci, the most strongly associating classical *HLA* alleles were *HLA-B*0801*, *HLA-DRB1*0301*, and *HLA-DQB*02*, respectively. These *HLA* alleles are all in strong but partial LD with *C4* BS, the most protective of the *C4* alleles; they are also in partial LD with the low-risk allele at rs13194505, representing the distinct signal several megabases to the left (**Fig. 4**). In joint analyses with each of these *HLA* alleles, genetically predicted *C4A* expression and rs13194505 continued to associate strongly with schizophrenia, while the *HLA* alleles did not.

In further joint analyses with rs13194504 and genetically predicted *C4A* expression, 0 of 2,514 tested *HLA* SNP, amino-acid and classical-allele polymorphisms (from ref⁵⁵, including all variants with MAF > 0.005) associated to schizophrenia as strongly as rs13194504 or predicted *C4A* expression did.

Extended Data Figure 8. Expression of *C4A* RNA in brain tissue (five brain regions) from 35 schizophrenia cases and 70 non-schizophrenia controls, from the Stanley Medical Research Institute Array Consortium. *C4A* RNA expression levels were measured by ddPCR.

Extended Data Figure 9. Secretion of C4, and specificity of the monoclonal anti-C4 antibody for C4 protein in human brain tissue and cultured primary cortical neurons.

(a) Brain tissue (from an individual affected with schizophrenia) was stained with a fluorescent secondary antibody, C4 antibody, or C4 antibody that was pre-adsorbed with purified C4 protein. Confocal images demonstrate the loss of immunoreactivity in the secondary-only and pre-adsorbed conditions.

(b) Primary human neurons were stained with a fluorescent secondary antibody, C4 antibody, or C4 antibody that was pre-adsorbed with purified C4 protein. Confocal images demonstrate the loss of immunoreactivity in the secondary-only and pre-adsorbed conditions. Scale bar for all images = 25 μm .

(c) Secretion of C4 protein by cultured primary neurons. Western blot for C4 protein analysis. (+) Purified human C4 protein. (–) Unconditioned medium, a negative control. (HN-conditioned) shows the same medium after conditioning by cultured human neurons at days 7 (d7) and 30 (d30). Details of Western blot protocol, antibody catalog numbers and concentrations used are in **Methods**. C4 molecular weight \sim 210 kDa.

Extended Data Figure 10. Mouse *C4* genes and additional analyses of the dLGN eye segregation phenotype in *C4* mutant mice and wild-type and heterozygous littermate controls.

(a) The functional specialization of *C4* into *C4A* and *C4B* in humans does not have an analogy in mice. Although the mouse genome contains both a *C4* gene and a *C4*-like gene (classically called *Slp*), and these genes are also present as a tandem duplication within the mouse MHC locus, analysis of the encoded protein sequences indicates a distinct specialization, as illustrated by the protein phylogenetic tree. Above, mouse *Slp* is indicated in gray to reflect its potential pseudogenization: *Slp* is already known to have mutations at a C1s cleavage site, which are thought to abrogate activation of the protein through the classical complement pathway⁵⁶; and the *M. musculus* reference genome sequence (mm10) at *Slp* shows a 1-bp deletion (relative to *C4*) within the coding region at chr17:34815158, which would be predicted to cause a premature termination of the encoded protein. (In some genome data resources, mouse *Slp* and *C4* have been annotated respectively as “*C4a*” (e.g. NM_011413.2) and “*C4b*” (e.g. NM_009780.2) based on synteny with the human *C4A* and *C4B* genes, but the above sequence analysis indicates that they are not paralogous to *C4A* and *C4B*.)

(b) Sequence differences between *C4A* and *C4B* – which are otherwise 99.5% identical at an amino acid level – are concentrated at the “isotypic site” where they shape each isotype's relative affinity for different molecular targets^{19,20}. At the isotypic site, mouse *C4* contains a combination of the residues present in human *C4A* and *C4B*.

(c) Expression of mouse *C4* mRNA in whole retina and lateral geniculate nucleus (LGN) from P5 animals and in purified retinal ganglion cells (RGCs) from P5 and P15 animals. These time points were chosen as P5 is a time of more robust synaptic refinement in the retinogeniculate system compared to P15. The same assays detected no *C4* RNA in control RNA isolated from *C4*^{-/-} mice (not shown).

N = 3 samples for p5 retina, LGN, and P15 RGCs, N = 4 samples for P5 RGCs; * $p < 0.05$ by ANOVA with post hoc Tukey-Kramer multiple-comparisons test.

(d) Representative images of dLGN innervation by contralateral projections (red in bottom image), ipsilateral projections (green in bottom image), and their overlap (yellow in bottom image). Scale bar = 100 μ m

(e) Quantification of the percentage of total dLGN area receiving both contralateral and ipsilateral projections shows a significant increase in *C4*^{-/-} compared to WT littermates (ANOVA, N = 5 mice/group, $p < 0.01$). These data are consistent with results using *R*-value analysis as shown in **Fig. 7**.

(f) Quantification of total dLGN area showed no significant difference between WT and *C4*^{-/-} mice (ANOVA, N = 5 per group, $p > 0.05$).

(g) Quantification of dLGN area receiving ipsilateral innervation showed a significant increase in ipsilateral territory in the *C4*^{-/-} mice compared to WT littermates (ANOVA, N = 5 mice/group, $p > 0.01$). This result is consistent with defects in eye specific segregation. Scale bar = 100 μ m

(h) The number of RGCs in the retina was estimated by counting the number of Brn3a⁺ cells in WT and *C4*^{-/-} mice. No differences were observed between WT and *C4*^{-/-} (*t*-test, N = 4 mice/group, $p > 0.05$). Scale bar = 100 μ m.