Schizophrenia risk from complex variation of complement component 4

Aswin Sekar1,2,3, Allison R. Bialas4,5, Heather de Rivera1,2, Avery Davis1,2, Timothy R. Hammond4, Nolan Kamitaki1,2, Katherine Tooley1,2, Jessy Presumey3, Matthew Bauml1,2,3,4, Vanessa Van Doren1, Giulio Genovese1,2, Samuel A. Rose2, Robert E. Handsaker1,2, Schizophrenia Working Group of the Psychiatric Genomics Consortium6, Mark J. Daly2,6, Michael C. Carroll5, Beth Stevens2,4 & Steven A. McCarroll1,2

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 identify. Here we show that this association arises in part from many structurally diverse alleles of the complement component 4 (C4) genes. We found that these alleles generated widely varying levels of C4A and C4B expression in the brain, with each common C4 allele associating with schizophrenia in proportion to its tendency to generate greater expression of C4A. Human C4 protein localized to 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 with schizophrenia.

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 grey matter1,2, and reduced numbers of synaptic structures on neurons3–5. Although 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, which are the earliest and most constant features of the disorder. 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 single nucleotide polymorphism (SNP) haplotypes that associate with risk of schizophrenia6; however, the functional alleles and mechanisms at these loci remain to be discovered. By far the strongest such genetic relationship is schizophrenia’s association with genetic markers across the major histocompatibility complex (MHC) locus, which spans several megabases (Mb) of chromosome 6 (refs 6–10). 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 genes11,12; however, schizophrenia’s association to the MHC has not yet been 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 because 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 variant5,6,10. 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 (refs 6, 10), which encodes a regulator of C4 (ref. 13), further motivated us to consider C4.

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 and Extended Data Fig. 1g)14–18. The protein products of C4A and C4B bind different molecular targets19,20. C4A and C4B segregate in both long (L) and short (S) genomic forms (C4AL, C4AS, C4BL and C4BS), 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 (HERV) 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 (Supplementary Methods).

We analysed 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 four common C4 structural haplotypes (AL–BL, AL–BS, AL–AL, and BS) were collectively present

1Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.
2Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA.
3MD–PhD Program, Harvard Medical School, Boston, Massachusetts 02115, USA.
4Department of Neurology, F.M. Kirby Neurobiology Center, Boston Children’s Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA.
5Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
6Program in Cellular and Molecular Medicine, Boston Children’s Hospital, Boston, Massachusetts 02115, USA.
7Lists of participants and their affiliations appear in the Supplementary Information.

© 2016 Macmillan Publishers Limited. All rights reserved
The results of this expression analysis were consistent across all five brain regions analysed. First, RNA expression of C4A and C4B increased proportionally with copy number of C4A and C4B, respectively (Fig. 3a, b and Extended Data Fig. 4). These observations mirror earlier observations in human serum15. Second, expression levels of C4A were two to three 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 ($P < 10^{-7}$, $P < 10^{-2}$, $P < 10^{-3}$, respectively, in the three cohorts examined, by Spearman rank correlation) (Fig. 3c and Extended Data Fig. 4).

We used the above data to create genetic predictors of C4A and C4B expression levels in the brain (Supplementary 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 analysed genome-wide for SNPs, implicating the MHC locus as the strongest of more than 100 genome-wide-significant associations6. Our analysis indicated 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, Supplementary Table 2).

We then analysed 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: 25–34 Mb), 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 (we use this set’s most strongly associating SNP, rs13194504, as its genetic proxy). The other peak of association in C4, where schizophrenia associated most strongly with the genetic predictor of C4A expression levels ($P = 3.6 \times 10^{-25}$) (Fig. 4a and 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 with schizophrenia with similar strengths ($P = 3.6 \times 10^{-24}$ and $5.5 \times 10^{-28}$, 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 $8.0 \times 10^{-14}$, respectively, Fig. 4c, d).

Controlling for both genetic variables revealed a third association signal just proximal to the MHC locus (Fig. 4e) involving SNPs around BAK1 and SYNGAP1, the latter of which encodes a major component of the postsynaptic density; de novo loss-of-function mutations in SYNGAP1 associate with autism26. In joint analysis, all three genetic signals remained significant ($P = 8.0 \times 10^{-14}$, $2.8 \times 10^{-10}$, and $1.7 \times 10^{-8}$, 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 associations14,15,12. In contrast, in schizophrenia, classical HLA alleles
associated with 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 with 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 was present (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 U-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 synapses.\(^{29-31}\)

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 grey 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...
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 35,896 controls of European ancestry from the Psychiatric Genomics Consortium. The height of each point represents the statistical significance score 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 colour-coded by their levels of correlation to rs13194504 (upper panel) or by their levels of correlation to genetically predicted brain C4 expression (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 Figs 5–7 for detailed association analyses involving C4 locus structures and HLA alleles.

Extended Data Fig. 9a) and a subset of astrocytes. Much of the C4 immunoreactivity was punctate (Fig. 6b), co-localizing with synaptic puncta identified by co-immunostaining for the pre- and postsynaptic markers VGLUT1/2 (also known as SLC17A7 and SLC17A6, respectively) and PSD-95 (also known as DLG4) (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+; neurofilament+ processes) and 35% in axons (MAP2+, neurofilament+ 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 (Figs 4 and 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 patients4–5, 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 uncertain52. However, waves of postnatal synapse elimination occur in many brain regions, and strong experimental models have been established in several mammalian visual systems. In these systems, 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 refinement29–31,33–35. We found that C4 RNA was expressed in the LGN and in RGCs purified from the retina during the period of synaptic pruning (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 complement29,30. We found that in mice deficient in C4 (ref. 36), C3 immunostaining in the dLGN was greatly reduced compared to wild-type 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 remodelling, 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,
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 markers VGLUT1/2 and the postsynaptic marker PSD-95. 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 PSD-95. Scale bars, 25 μm (a, c, and d); 5 μm (b, left); and 1 μm (b, right). Extended Data Fig. 9 contains additional data on antibody specificity.

suggesting reduced synaptic pruning (Fig. 7d, Extended Data Fig. 10d, e, and Supplementary Methods). The degree of deficit in C4−/− mice was similar to that previously reported for C1q−/− and C3−/− mice29,31. 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.

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 wild-type (WT) littermates. b, Quantification confirmed reduced C3 immunoreactivity in the dLGN (n = 3 mice per group, P < 0.05, t-test; y axis: mean fluorescence intensity, normalized to wild type). 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 per 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 wild-type mice. The overlapping area is significantly increased in C4−/− mice (n = 6 mice per group, P < 0.01, ANOVA with Bonferroni post-hoc tests). Bottom, threshold-independent analysis using the R value36 (R = log10(F disput/F control)). Pixels are pseudocoloured with an R value heat map (red indicates areas having only contralateral inputs; purple, only ipsilateral inputs). Compared to their wild-type littermates, C4-deficient mice exhibited lower R value variance, indicating defects in synaptic refinement (n = 6 mice per group, P < 0.001, ANOVA with Bonferroni post-hoc tests). Control experiments analysing 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–d represent s.e.m.
Discussion

We developed ways to analyse a complex form of genome structural variation (Figs 1 and 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 (Figs 3–5). We found that C4 is expressed by neurons, localized to dendrites, axons, and synapses, and is secreted (Fig. 6). In mice, C4 promoted 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 evolution and appear to become impaired in schizophrenia35–40. 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.

The principal pathological findings in brains of those affected with schizophrenia involve loss of cortical grey matter without cell death: affected individuals exhibit abnormal cortical thinning1,2 and reduced numbers of synaptic structures on cortical pyramidal neurons3–5. In the brain, complement receptors are expressed primarily by microglia, the phagocytic immune cells of the central nervous system. 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 by microglia 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 proteins42–44. Diverse synaptic abnormalities could in principle interact with the complement system and other pathways45,46 to cause excessive stimulation of microglia and elimination of synapses during adolescence and early adulthood.

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 favours binding to carbohydrate surfaces49,50, differences with an established basis in C4 protein sequence and structure47,48. An intriguing possibility is that C4A and C4B differ in affinity for an unknown binding site at synapses.

To date, few genome-wide association study (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 to all human phenotypes.

Online Content Methods, along with any additional Extended Data display items and Online Content all human phenotypes.

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 to all human phenotypes.
Supplementary Information is available in the online version of the paper.

Acknowledgements The authors would like to remember the late T. Stanley with appreciation and express their gratitude for his support. We thank S. Hyman, E. Lande, C. Bargmann, and C. Patil for conversations about the project and comments on drafts of the manuscript; M. Webster for expert advice on immunohistochemistry; B. 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; C. Emba for assistance with experiments; and C. 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 U01 MH105641 (to S.A.M.), 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. and B.S. performed the laboratory experiments and computational analyses to understand the molecular and population genetics of the C4 locus (Figs 1 and 2). A.S. and M.D. performed the C4 immunocytochemistry and classical pathway complement. A.S. and A.D. did the imputation and association analysis (Figs 4 and 5a, c). M.B. analysed expression variation in human brain (Figs 3 and 5b). M.B. and S.A.M. performed the C4 immunohistochemistry experiments respectively, with advice from A.R.B. H.d.R and T.R.H. performed the C4 immunocytochemistry and complement factor C4 in schizophrenia. A.S. and A.D. contributed expertise and reagents for experiments described in Fig. 6 and 7. H.d.R and T.R.H. performed the C4 immunocytochemistry and immunohistochemistry experiments respectively, with advice from A.R.B. A.S. and A.D. performed the laboratory experiments and computational analyses to understand the molecular and population genetics of the C4 locus (Figs 1 and 2). A.S. and M.D. performed the C4 immunocytochemistry and classical pathway complement. A.S. and A.D. did the imputation and association analysis (Figs 4 and 5a, c). M.B. analysed expression variation in human brain (Figs 3 and 5b). M.B. and S.A.M. performed the C4 immunohistochemistry experiments respectively, with advice from A.R.B. H.d.R and T.R.H. performed the C4 immunocytochemistry and classical pathway complement. A.S. and A.D. contributed expertise and reagents for experiments described in Fig. 6 and 7. H.d.R and T.R.H. performed the C4 immunocytochemistry and immunohistochemistry experiments respectively, with advice from A.R.B. A.S. and A.D. performed the laboratory experiments and computational analyses to understand the molecular and population genetics of the C4 locus (Figs 1 and 2). A.S. and M.D. performed the C4 immunocytochemistry and classical pathway complement. A.S. and A.D. did the imputation and association analysis (Figs 4 and 5a, c). M.B. analysed expression variation in human brain (Figs 3 and 5b).
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: 32 Mb (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 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. Haplotype with multiple copies of this module (middle and bottom) contains 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 observed. 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 (Fig. 1). Many additional forms of this locus appear to have arisen by non-allelic homologous recombination and gene conversion (ref. 18 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 each C4 gene type (A or B isotype; long or short form) in each genome sampled. Molecular assays to measure the copy number of each C4 gene type (A or B isotype; long or short form) in each genome sampled. b, Measurement of copy number of C4 gene types in the genomes of 162 individuals (from HapMap). 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, Inference of the C4 gene contents of the founder chromosomes in each trio. Where possible, use existing sequence data to also infer gene order.

© 2016 Macmillan Publishers Limited. All rights reserved
Extended Data Figure 3 | Linkage disequilibrium relationships ($r^2$) of MHC SNPs to forms of C4 structural variation. a, b, Correlations of SNPs in the MHC locus with copy number of C4 gene types (a) and larger-scale structural forms (haplotypes) (b) of the C4 locus. Dashed, vertical lines indicate the genomic location of the C4 locus. 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. a–e, Data for tissues from the Stanley Medical Research Institute (SMRI) Array Consortium consisting of anterior cingulate cortex (a), cerebellum (b), corpus callosum (c), orbital frontal cortex (d), and parietal cortex (e). f, Data for the frontal cortex samples from the NHGRI Genes and Tissues Expression (GTEx) Project. g, h, 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 Supplementary 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 (Supplementary 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, ref. 6.) 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 Figs 6 and 7.

The single asterisk (*) indicates that 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 studies53,54 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. The double asterisk (***) indicates 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.

| Expression predictors (derived from post mortem brain RNA analysis) | Association of SCZ with this variant (unconditioned analysis) | Association of SCZ with this variant (joint analysis with predicted C4A expression as a covariate) | Association of SCZ with genetically predicted C4A expression (joint analysis with this variant as a covariate) |
|---|---|---|---|
| C4A expression (genetic prediction from imputed C4 locus structure) | 3.60E-24 | 0.247 | 0.024 | NA | NA | NA | NA | NA | NA | NA | NA | 2.35E-18 | 0.253 | 0.029 |
| C4B expression (genetic prediction from imputed C4 locus structure) | 2.31E-07 | -0.092 | 0.018 | 0.71 | 0.008 | 0.021 | 5.40E-11 | 0.247 | 0.038 |

| Copy number of C4 structural features | Association of SCZ with this variant (unconditioned analysis) | Association of SCZ with this variant (joint analysis with predicted C4A expression as a covariate) | Association of SCZ with genetically predicted C4A expression (joint analysis with this variant as a covariate) |
|---|---|---|---|
| C4A genes | 6.31E-19 | 0.112 | 0.013 | 0.45 | -0.023 | 0.030 | 6.54E-07 | 0.287 | 0.058 |
| C4B genes | 1.45E-04 | -0.066 | 0.017 | 0.63 | 0.009 | 0.019 | 4.58E-21 | 0.252 | 0.027 |
| Long C4 genes (with HERV) | 7.09E-23 | 0.084 | 0.009 | 0.17 | 0.029 | 0.021 | 4.89E-03 | 0.170 | 0.060 |
| Short C4 genes (no HERV) | 3.94E-14 | -0.086 | 0.011 | 0.32 | -0.015 | 0.015 | 7.38E-12 | 0.225 | 0.033 |
| Total C4 copy number | 9.41E-15 | 0.123 | 0.016 | 1.00 | 0.000 | 0.025 | 5.40E-11 | 0.247 | 0.038 |

| Specific C4 locus structures (haplotypes of one or more C4 genes) | Association of SCZ with this variant (unconditioned analysis) | Association of SCZ with this variant (joint analysis with predicted C4A expression as a covariate) | Association of SCZ with genetically predicted C4A expression (joint analysis with this variant as a covariate) |
|---|---|---|---|
| BS | 2.29E-19 | -0.171 | 0.019 | 0.16 | -0.045 | 0.032 | 1.10E-06 | 0.200 | 0.041 |
| AL-BS | 0.03 | -0.027 | 0.013 | 0.62 | 0.007 | 0.013 | 3.39E-23 | 0.250 | 0.025 |
| AL-BL | 6.16E-07 | 0.058 | 0.012 | 0.41 | 0.011 | 0.013 | 6.34E-19 | 0.238 | 0.027 |
| AL-AL | 2.09E-06 | 0.093 | 0.020 | 0.56 | -0.013 | 0.023 | 2.49E-19 | 0.256 | 0.028 |

| Combinations of C4 locus structure and flanking SNP haplotype | Association of SCZ with this variant (unconditioned analysis) | Association of SCZ with this variant (joint analysis with predicted C4A expression as a covariate) | Association of SCZ with genetically predicted C4A expression (joint analysis with this variant as a covariate) |
|---|---|---|---|
| AL-BS-4 | 0.21 | -0.027 | 0.022 | 0.95 | -0.002 | 0.022 | 7.96E-24 | 0.247 | 0.025 |
| AL-BS-2 | 0.35 | -0.031 | 0.033 | 0.80 | -0.008 | 0.033 | 5.45E-24 | 0.247 | 0.024 |
| AL-BS-3 | 0.34 | -0.028 | 0.029 | 0.91 | -0.003 | 0.030 | 5.69E-24 | 0.247 | 0.024 |
| AL-BS-5 | 0.07 | 0.053 | 0.029 | 0.01 | 0.077 | 0.029 | 5.73E-25 | 0.252 | 0.024 |
| AL-BS-1 | 0.04 | -0.064 | 0.031 | 0.19 | -0.040 | 0.031 | 1.32E-23 | 0.245 | 0.024 |
| AL-BS-other | 0.30 | -0.023 | 0.022 | 0.92 | 0.002 | 0.023 | 6.19E-24 | 0.247 | 0.025 |
| AL-BL-2 | 0.03 | 0.036 | 0.017 | 0.77 | 0.005 | 0.017 | 3.52E-23 | 0.246 | 0.025 |
| AL-BL-3 | 2.15E-03 | 0.042 | 0.014 | 0.55 | 0.008 | 0.014 | 3.28E-22 | 0.243 | 0.025 |
| AL-BL-1 | 0.35 | 0.032 | 0.034 | 0.92 | 0.003 | 0.035 | 5.55E-24 | 0.247 | 0.024 |
| AL-BL-other | 0.23 | 0.029 | 0.024 | 0.95 | 0.002 | 0.024 | 7.37E-24 | 0.247 | 0.025 |
| AL-AL-1 | 0.01 | 0.074 | 0.028 | 0.38 | -0.026 | 0.029 | 9.42E-23 | 0.255 | 0.026 |
| AL-AL-2 | 3.33E-04 | 0.097 | 0.027 | 0.94 | -0.002 | 0.029 | 2.31E-21 | 0.248 | 0.026 |

| Alternative hypotheses | Association of SCZ with this variant (unconditioned analysis) | Association of SCZ with this variant (joint analysis with predicted C4A expression as a covariate) | Association of SCZ with genetically predicted C4A expression (joint analysis with this variant as a covariate) |
|---|---|---|---|
| C4B nulls (for whom total C4B copy number = 0) * | 0.36 | 0.061 | 0.066 |
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 grey. The series of conditional analyses shown in 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 and Extended Data Fig. 6a) and HLA-DRB1 and HLA-DQB1 (in analyses controlling for the signal defined by rs13194504, Fig. 4c and Extended Data Fig. 6b). At these loci, the most strongly associating classical HLA alleles were HLA-B*0801, HLA-DRB1*0301, and HLA-DQB1*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 ref. 55, including all variants with minor allele frequency (MAF) >0.005) associated with schizophrenia as strongly as rs13194504 or predicted C4A expression did.

| Variants included in analysis | rs13194504 | Genetically predicted C4A expression | B*0801 | DRB1*0301 | DQB1*02 |
|-------------------------------|------------|--------------------------------------|--------|-----------|--------|
| rs13194504                    | 5.50E-28   |                                      |        |           |        |
| C4Aexp                        |            | 3.60E-24                             |        |           |        |
| B*0801                        |            |                                      | 1.20E-21 |           |        |
| DRB1*0301                     |            |                                      |        | 6.40E-19  |        |
| DQB1*02                       |            |                                      |        |           | 6.25E-17 |
| rs13194504, C4Aexp            | 8.00E-14   | 7.80E-10                             |        |           |        |
| rs13194504, B*0801            | 5.75E-11   | 2.00E-04                             |        |           |        |
| rs13194504, DRB1*0301         | 2.89E-15   |                                      | 6.20E-06 |           |        |
| rs13194504, DQB1*02           | 9.60E-19   |                                      |        |           | 1.82E-07 |
| rs13194504, C4Aexp, B*0801    | 6.94E-11   | 8.23E-07                             | 0.96   |           |        |
| rs13194504, C4Aexp, DRB1*0301 | 1.39E-12   | 2.18E-05                             | 0.41   |           |        |
| rs13194504, C4Aexp, DQB1*02   | 3.06E-13   | 1.03E-04                             |        |           | 0.03   |
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. P values are derived from Mann–Whitney U-test.

Anterior cingulate cortex

Orbital frontal cortex

Parietal cortex

Cerebellum

Corpus callosum
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 bars, 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 catalogue numbers and concentrations used are in Supplementary Methods. C4 molecular weight, ~210 kDa.
Article Research

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. Top, mouse Slp is indicated in grey 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 pathway, 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’ (for example, NM_011413.2) and ‘C4b’ (for example, 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, 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; *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 μm. e, Quantification of the percentage of total dLGN area receiving both contralateral and ipsilateral projections shows a significant increase in C4−/− mice compared to wild-type littermates (ANOVA, n = 5 mice per 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 wild-type 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 wild-type littermates (ANOVA, n = 5 mice per group, P > 0.01). This result is consistent with defects in eye specific segregation. Scale bar, 100 μm. h, The number of RGCs in the retina was estimated by counting the number of Brn3a+ cells in wild-type and C4−/− mice. No differences were observed between wild-type and C4−/− mice (t test, n = 4 mice per group, P > 0.05). Scale bar, 100 μm.