Identification of novel neuroprotective loci modulating ischemic stroke volume in a cross between wild-derived inbred mouse strains

Han Kyu Lee*, Samuel J. Widmayer†, David L. Aylor†, Douglas A. Marchuk*

* Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, North Carolina, USA

† Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, USA

Running title: Novel neuroprotective loci for stroke

Key words: Ischemic stroke, wild-derived mouse strains, Neuroprotection, quantitative trait mapping

Correspondence should be addressed to:

Douglas A. Marchuk, PhD

213 Research Drive

Duke University Medical Center, BOX 3175

Durham, NC 27710

Phone / Fax: +1-919-684-1945 / +1-919-684-2790

Email: douglas.marchuk@duke.edu

Abstract

To identify genes involved in cerebral infarction we have employed a forward genetic approach using quantitative trait locus (QTL) mapping for cerebral infarct volume after middle cerebral artery occlusion, using common inbred mouse strains. This work has demonstrated that infarct volume is inversely correlated with cerebral collateral vessel density, since the collateral circulation provides a compensatory shunt for reperfusion of the ischemic territory. To overcome the limited genetic diversity among domesticated inbred mouse strains, we have expanded the pool of allelic variation by utilizing the eight parental strains of the Collaborative Cross. We found that one wild-derived strain, WSB/EiJ, breaks general rule that collateral vessel density inversely correlates with infarct volume. WSB/EiJ and another wild-derived strain, CAST/EiJ, show the highest collateral vessel densities of any inbred strain, but infarct volume of WSB/EiJ is 8.7-fold larger than CAST/EiJ. QTL mapping between these strains identified 4 new neuroprotective loci modulating cerebral infarct volume while not affecting collateral vessel phenotypes. To identify causative variants in genes within the loci, we surveyed non-synonymous coding SNPs between the strains and also performed RNA sequencing of brain tissue to determine strain-specific gene expression differences. Combining the lists of genes harboring coding SNPs predicted to be damaging with genes showing strain-specific expression differences leaves only 15 genes mapping in any of the four intervals. The identification of the genes underlying these neuroprotective loci will provide new understanding of genetic risk factors of ischemic stroke which may provide novel targets for future therapeutic intervention of human ischemic stroke.

Introduction

Stroke, the sudden death of the brain tissue occurring when the blood flow to the brain is lost by blockage or rupture, is a leading cause of death in the United States (the fourth) and the world (the second) (Roger *et al.* 2012; Johnson *et al.* 2016). The majority (over 80%) of strokes are ischemic in origin and results in irreversible death of brain tissue (infarction). Decades of studies have identified individual risk factors that causes stroke (eg, smoking, diabetes, obesity, etc.). Furthermore, numerous epidemiologic studies have estimated that two-thirds of the risk for ischemic stroke is due to genetic factors (Jamrozik *et al.* 1994; Dichgans 2007). For example, a recent study has reported that a documented parental ischemic stroke by the age of 65 years was associated with a 3-fold increase in ischemic stroke risk in offspring (Mozaffarian *et al.* 2008; Mozaffarian *et al.* 2015). However, to date, identification of a significant component of the genetic factors for ischemic stroke remains elusive.

Animal models of focal cerebral ischemia have been established to investigate the pathophysiologic events occurring after ischemic stroke. Most stroke models induce cerebral ischemia within the middle cerebral artery (MCA) territory as most relevant for thrombo-embolic stroke. MCA occlusion models vary both in the extent of occlusion (permanent vs. transient occlusion) and the site of occlusion (proximal vs. distal portion of the vessel). Several reports have determined that permanent occlusion of distal MCA method produces more restricted and reproducible damage to the cerebral hemisphere (Majid *et al.* 2000; Lambertsen *et al.* 2002; Carmichael 2005). Furthermore, previous studies have also demonstrated that different inbred mouse strains show robust differences in stroke outcomes, providing evidence that the innate response to permanent

focal cerebral ischemia is under strong genetic control (Barone *et al.* 1993; Majid *et al.* 2000; Lambertsen *et al.* 2002; Sugimori *et al.* 2004).

Therefore, we have taken a forward genetic approach (quantitative trait locus - QTL - mapping) to identify novel genes/pathways (genetic components) involved in modulating infarct volume in ischemic stroke using the well-established model of permanent occlusion of distal MCA (pMCAO). Previously, we uncovered several genetic loci that are involved in regulating ischemic stroke and identified several genes modulating infarct volume within the genetic loci (Keum and Marchuk 2009; Chu *et al.* 2013; Keum *et al.* 2013; Lee *et al.* 2016; Lee *et al.* 2018). For example, as a major determinant of infarct volume after pMCAO, we identified a locus on distal chromosome 7 (cerebral infarct volume QTL 1, *Civq1*) using multiple F2 crosses between inbred mouse strains. Interestingly, this locus shares the same location on chromosome 7 with a locus modulating cerebral collateral vessel number (*Canq1*) (Keum and Marchuk 2009; Zhang *et al.* 2010; Keum *et al.* 2013; Lee *et al.* 2016). Due to the strong effects of this locus, infarct volume is inversely correlated with pial collateral vessel density across a broad range of inbred mouse strains. However, we also found that a small number of inbred strains break this correlation. One such strain is C3H/HeJ (C3H) that exhibits a similar number of vessel connections compared to C57BL/6J (B6), but these two strains show significantly different infarct volumes (Chu *et al.* 2013). Thus, in the common inbred mouse strains, infarct volume is mainly modulated by collateral circulation, but can also regulated by a collateral-independent (eg, innate neuroprotection) mechanism. Although we have identified both collateral-dependent and –independent loci that modulate infarct volume, genetic variation is limited within the domesticated inbred strains. Therefore, in

this study, we utilize eight parental mouse strains of the collaborative cross (CC) mouse to expand genetic variations/differences in the mouse genome and through F2 intercross between two such strains, identify novel genetic loci modulating ischemic stroke via a collateral-independent manner (ie, neuroprotection).

Materials and methods

Animals. All inbred mouse strains were obtained from the Jackson Laboratory (Bar Harbor, ME) or bred locally from breeding pairs of each strain. Mice (both male and female animals) were age matched (P21 for collateral vessel perfusion and 12 ± 1 week for pMCAO) for all experiments. All animal procedures were conducted under protocols approved by the Duke University IACUC in accordance with NIH guidelines.

Collateral vessel density measurement. As collateral vessel traits are determined by 3 weeks of age and remain constant for many months (Clayton *et al.* 2008), the collateral vessel phenotype was measured at P21 as previously described (Lee *et al.* 2016; Lee *et al.* 2018). Mice were anesthetized with ketamine (100 mg/kg) and xylazine (5 mg/kg), and the ascending thoracic aorta was cannulated. The animals were perfused with freshly made buffer (1 mg/ml adenosine, 40 g/ml papaverine, and 25 mg/ml heparin in PBS) to remove the blood. The pial circulation was then exposed after removal of the dorsal calvarium and adherent dura mater. The cardiac left ventricle was cannulated and a polyurethane solution with a viscosity sufficient to minimize capillary transit (1:1 resin to 2-butanone, PU4ii, VasQtec) was slowly infused; cerebral circulation was visualized under a stereomicroscope during infusion. The brain surface was topically rinsed with 10% PBS-buffered formalin and the dye solidified for 20 min. After post-fixation with 10% PBS-buffered formalin, pial circulation was imaged. All collaterals interconnecting the anterior- and middle cerebral artery trees of both hemispheres were counted.

Permanent MCAO. Focal cerebral ischemia was induced by direct permanent occlusion of the distal MCA as previously described (Lee *et al.* 2016; Lee *et al.* 2018). Briefly, adult mice were anesthetized with ketamine (100 mg/kg) and xylazine (5 mg/kg), and then 0.5%

bupivacaine (5 mg/ml) was also administrated by injection at the incision site. The right MCA was exposed by a 0.5 cm vertical skin incision midway between the right eye and ear under a dissecting microscope. After the temporalis muscle was split, a 2-mm burr hole was made with a high-speed micro drill at the junction of the zygomatic arch and the squamous bone through the outer surface of the semi-translucent skull. The MCA was clearly visible at the level of the inferior cerebral vein. The inner layer of the skull was removed with fine forceps, and the dura was opened with a 32-gauge needle. While visualizing under an operating microscope, the right MCA was electrocauterized. The cauterized MCA segment was then transected with microscissors to verify permanent occlusion. The surgical site was closed with 6-0 sterile nylon sutures. The temperature of each mouse was maintained at 37°C with a heating pad during the surgery and then was placed in an animal recovery chamber until the animal was fully recovered from the anesthetic. Mice were then returned to their cages and allowed free access to food and water in an air-ventilated room with the ambient temperature set to 25°C.

Infarct volume measurement. Cerebral infarct volumes were measured 24 h after surgery because the size of the cortical infarct is largest and stable at 24 h after distal permanent MCA occlusion (Lambertsen *et al.* 2005). Twenty-four hours after pMCAO surgery, the animals were euthanized, and the brains were carefully removed. The brains were placed in a brain matrix and sliced into 1 mm coronal sections after being chilled at -80°C for 4 min to slightly harden the tissue. Each brain slice was placed in 1 well of a 24-well plate and incubated for 20 min in a solution of 2% 2,3,5-triphenyltetrazolium chloride (TTC) in PBS at 37°C in the dark. The sections were then washed once with PBS and fixed with 10% PBS-buffered formalin at 4°C. Then, 24 h after fixation, the caudal face of

each section was scanned using a flatbed color scanner. The scanned images were used to determine infarct volume (Wexler *et al.* 2002). Image-Pro software (Media Cybernetics) was used to calculate the infarcted area of each slice by subtracting the infarcted area of the hemisphere from the non-infarcted area of the hemisphere to minimize error introduced by edema. The total infarct volume was calculated by summing the individual slices from each animal.

SNP genotyping. Genomic DNA was isolated from tails of F2 intercross between CAST/EiJ (CAST) and WSB/EiJ (WSB) mice using DNeasy Tissue kit (Qiagen, Hilden, Germany). Genome-wide SNP genotyping was performed with a Mouse Universal Genotyping Array (MUGA, 8K SNPs). Array hybridization including sample preparation was performed by Neogen/GeneSeek, Inc (Lincoln, NE).

Linkage (QTL) analysis. Genome-wide scans were performed using R/qtl software. Genotypes from MUGA were prepared for QTL mapping as follows. Total 277 informative markers for CAST and WSB spaced at approximately 5 cM intervals across the mouse genome were used for genetic mapping. The significance thresholds for LOD scores were determined by 1000 permutations using all informative markers. Each QTL was indicated as a significant when its LOD score exceeded 95% ($p < 0.05$) of the permutation distribution and QTL intervals exceeded 99% ($p < 0.01$) of the permutation distribution were considered as candidate intervals. The 95% confidence interval (CI) of each peak was determined by 1.5-LOD support interval. The physical (Mb) map positions based on the genomic sequence from the GRCm38/mm10 were calculated using Mouse Map Converter tool of the Jackson Laboratory (http://cgd.jax.org/mousemapconverter/).

Haplotype analysis. For the intervals of the identified 4 neuroprotective loci (*Civq8* on chromosome 1, *Civq9* on chromosome 6, *Civq10* on chromosome 13, and *Civq11* on chromosome 17), SNP data were obtained from the Mouse Phenome Database (http://phenome.jax.org/). An amino acid change affected detrimental to a protein was examined by 3 independents *in silico* prediction algorithms, Polymorphism Phenotyping v2 (PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/index.shtml), Sorting Intolerant From Tolerant (SIFT, http://sift.jcvi.org), and Protein Variation Effect Analyzer (PROVEAN, http://provean.jcvi.org).

RNA sequencing analysis. Paired-end, 150 bp sequencing reads were generated from adult (8 – 12 week) brain tissue mRNA of 6 CAST (3 males and 3 females) and 6 WSB (3 males and 3 females) mice on the Illumina HiSeq 2500 platform. Adapters for all paired-end sequencing reads were trimmed using Cutadapt (Martin 2011). Differential gene expression analysis is occasionally confounded by differences in alignment rate of reads obtained from strains that are highly diverged from the reference genome. Furthermore, there are millions of variant sites between the CAST and WSB strains, which would make these effects substantially pronounced. To alleviate these effects, we performed our alignment to pseudo-reference genomes, produced by altering the mouse reference genome (GRCm38.p4) using variant calls for the CAST and WSB strains. Reads from each strain were aligned to their respective pseudo-reference genome using TopHat (v. 2.1.1) (Trapnell *et al.* 2009) under default settings and re-mapped to reference coordinates using the Lapels pipeline (Holt *et al.* 2013; Huang *et al.* 2014).

Differential gene expression. Gene expression counts for each sample were obtained using HTseq (Anders *et al.* 2015) under default settings and subsequently imported into

DESeq2 (Love *et al.* 2014), where gene counts for each mouse were normalized for differences in sequencing effort and dispersions were calculated for each sample. Differential expression was determined using a two-sided Wald test comparing two negative binomial distributions. *p*-values were adjusted using a false-discovery rate of 5% by Benjamini-Hochberg correction (Benjamin and Hochberg 1995).

Statistics analysis. Results were represented as the mean ± SEM. Significant differences between datasets were determined using *p*-values $<.05$ were considered significant.

Data availability. Strain used in this study are available through the Jackson Laboratory catalog. Supplemental data contain one figure (Figure S1) and seven tables (Table S1 – S7). Figure S1 shows both the number of collateral vessel connections and infarct volume after pMCAO for all individual F2 animals. Table S1 contains detailed information of all coding SNPs including *in silico* prediction of three independent algorithms. Table S2 contains detailed information of RNA sequencing analysis within identified neuroprotective loci. Table S3 contains genotype and phenotype information of 251 F2 (CAST x WSB) animals for QTL mapping. Table S4, S5, S6, and S7 contain genotype and phenotype information for fine-mapping of *Civq8* on chromosome 1, *Civq9* on chromosome 6, *Civq10* on chromosome 13, and *Civq11* on chromosome 17, respectively. RNA sequencing data are available at GEO with the accession number: *RNA sequencing data has been submitted*. Supplemental material available at Figshare:

Results

Analysis of pial collateral vessel density and infarct volume after permanent MCAO for all eight parental strains of Collaborative Cross mouse strain.

Using the eight parental mouse strains of the CC, we first examined the number of collateral vessel connections between the anterior cerebral artery (ACA) and middle cerebral artery (MCA) (Figure 1A and B). Strains B6, NOD/ShiLtJ (NOD), NZO/HILtJ (NZO), and 129S1/SvlmJ (129S1), and a wild-derived strain, PWK/PhJ (PWK) show similar level of vessel connections (B6 (20.4), NOD (21.0), NZO (21.4), 129S1 (22.4), and PWK (20.4)), whereas A/J (10.2) showed a lower level. Interestingly, two wild-derived strains, CAST and WSB, showed the highest number of collateral vessel connections (CAST (28.8) and WSB (27.3)) that we have ever observed in any inbred strain.

To determine whether collateral vessel density is inversely correlated with infarct volume in these new strains, we next examined ischemic infarct volume in the permanent occlusion model. We performed pMCAO and measured infarct volume for the parental mouse strains (Figure 1C and D). Infarct volume in these parental mouse strains displayed various levels after ischemic stroke induction; A/J (33.4 mm^3), B6 (8.6 mm^3), PWK (8.7 mm^3), NOD (12.3 mm^3), NZO (6.8 mm^3), 129S1 (9.0 mm^3), WSB (22.6 mm^3), and CAST (2.6 mm^3). As expected, most parental mouse strains showed an inverse correlation between collateral vessel connections and infarct volume. However, one strain, WSB, breaks this inverse correlation. Although two wild-derived mouse strains, CAST and WSB, exhibit the highest number of vessel connections we have seen in any inbred strain (Figure 1B), WSB consistently displayed a dramatically larger infarct volume (8.7-fold) after pMCAO compared to CAST (22.6 mm^3 vs. 2.6 mm^3) (Figure 1D). This

suggests that, at least in part, WSB may modulate infarct volume after cerebral ischemia through a collateral-independent mechanism.

Four novel loci contribute infarct volume differences between CAST and WSB.

In order to identify the genetic loci modulating infarct volume between these two strains, we generated F1 and F2 progeny between CAST and WSB and examined both collateral vessel density and infarct volume. Consistent with the number of collateral vessel connections between CAST and WSB, collateral vessel connections in the F1 and F2 generations exhibited a very tight distribution; WSB (27.3), CAST (28.8), F1 (27.7), and F2 (27.6) (Figure 2A). Moreover, the number of collateral vessel connections in individual F2 animals showed a similar range of connections (Figure S1A). Thus, genetic variation within the two strains does not appear to change the collateral vessel phenotype. Next, using F1 and F2 progeny, we measured infarct volume after pMCAO. In contrast to the vascular phenotype, infarct volume in the F2 generation was widely distributed, ranging from 1.1 to 38.9 mm^3 (Figure 2B and Figure S1B) and covering nearly the entire phenotypic spectrum observed across all inbred strains (Keum and Marchuk 2009; Keum *et al.* 2013).

To discover genetic loci that modulate infarct volume we performed genome-wide linkage analysis. A total of 251 F2 mice (Figure 2B and Supplemental Figure 1B) were used for genome-wide SNP genotyping using a Mouse Universal Genotyping Array (MUGA, 8K SNPs). For genome-wide linkage analysis, 277 informative SNP markers spaced at approximately 5 cM intervals across the mouse genome were selected. From this linkage mapping, we identified four new linkage peaks that display highly significant linkage to the infarct volume trait. These were located on chromosomes 1, 6, 13, and 17 (Figure 3).

These new loci were designated as *Civq8* through *Civq11* (Cerebral infarct volume QTL) based on our previous work that identified seven distinct loci for this trait (Keum and Marchuk 2009; Chu *et al.* 2013). The novel loci are *Civq8* (chromosome 1, LOD 5.09 (JAX00250952)), *Civq9* (chromosome 6, LOD 7.68 (UNC_rs51474193)), *Civq10* (chromosome 13, LOD 6.23 (UNC130410312)), and *Civq11* (chromosome 17, LOD 5.45 (UNC170143495)).

Fine-mapping uncovers potential candidate genes in the four new loci.

To better define the linkage intervals, we fine-mapped each locus by adding back additional SNP markers from the MUGA arrays and extended the interval an additional 1 Mb flanking each arm of 1.5-LOD support interval. Forty-five informative SNP markers obtained from the MUGA array used for fine-mapping of the *Civq8*, mapping to chromosome 1. The resulting candidate interval for *Civq8* is from 39 Mb to 70 Mb (Figure 4A and E). For *Civq9*, mapping to chromosome 6, forty-six informative SNP markers were used and its candidate interval is from 21 Mb to 29 Mb (Figure 4B and E). For *Civq10*, mapping to chromosome 13, thirty-eight informative SNP markers were used and its candidate interval is from 108 Mb to 120.4 Mb (Figure 4C and E). For *Civq11*, mapping to chromosome 17, forty informative SNP markers were used and its candidate interval is from 54 Mb to 65 Mb (Figure 4D and E).

To identify possible candidate gene(s) modulating infarct volume in these loci, we first surveyed candidate genes within the candidate interval of each QTLs and then sought the presence of non-synonymous coding SNPs (hereafter, coding SNP) in these genes. We identified a total of 330 coding SNPs in 90 coding genes in *Civq8*, 77 coding SNPs in 20 coding genes in *Civq9*, 109 coding SNPs in 34 coding genes in *Civq10*, and 157 coding

genes in 47 coding genes in *Civq11*. Within these intervals, the vast majority of these genes harbor multiple coding SNPs (Figure 4E and Table S1). To determine whether any of these non-synonymous amino acid substitutions might affects protein function, all the coding SNPs were subjected to three independent *in silico* algorithms that predict their functional consequences; SIFT, PolyPhen-2, and PROVEAN. In *Civq8*, 46 genes have a coding SNP (or SNPs) predicted to be damaging by at least one of the prediction algorithms. However, only 11 genes harbor a SNPs predicted to be damaging by all three programs and another 15 harbor damaging SNPs if the threshold is relaxed that only two of the three algorithms predict damaging consequences. In *Civq9*, 7 genes have coding SNPs predicted to be damaging by at least one algorithm, with only one predicted to be damaging by all three programs and another 3 genes by only two algorithms. In *Civq10*, a total 16 genes have coding SNPs predicted to be damaging with only 4 predicted to be damaging by all three programs and another 1 gene by only two algorithms. In *Civq11*, a total 27 genes have coding SNPs predicted to be damaging with only 3 predicted to be damaging by all three programs and another 7 genes by only two algorithms (Table 1 and Table S1). Although genes containing coding SNPs predicted to be damaging by all 3 independents *in silico* algorithms are the best candidate genes, even genes containing coding SNPs predicted to be damaging by only one *in silico* algorithm are still potential candidates. The details of all coding SNPs and functional predictions are listed in Table S1.

Strain-specific differential gene expression discovers candidate genes in the four new loci.

To identify other candidate genes underlying these loci, we sought evidence of strain-specific differential transcript levels between two wild-derived mouse strains, CAST and WSB. Strain-specific differences in transcript levels could potentially be caused by regulatory sequence variation acting in cis, but of any number of potential molecular mechanisms. We employed RNA sequencing data analysis using adult brain tissues between CAST and WSB. As shown in Figure 5A, a total 9,563 genes were differentially expressed between strains but only 220 of these genes are located within one of the four new QLTs, with 135 in *Civq8*, 10 in *Civq9*, 47 in *Civq10* and 28 in *Civq11*. Assuming that the most likely candidate genes would show at least a 2-fold difference in strain-specific expression reduces the list to 47 genes in *Civq8*, 3 genes in *Civq9*, 19 genes in *Civq10*, and 10 genes in *Civq11* (Figure 4E). All 220 genes are displayed in Table S2 which includes information on direction of difference, that is, which inbred strain of the pair exhibits the higher expression.

Upon the assumption genes that exhibit both strain-specific expression differences and coding SNPs predicted to be damaging would be the most robust candidates, we compared genes with at least a 2-fold difference in expression between CAST and WSB within the loci, and asked whether any of these harbored coding SNPs with functional consequences (Figure 5B). Eight genes meet this combined threshold in *Civq8*, one gene in *Civq9*, four genes in *Civq10*, and two genes in *Civq11*. As a starting point, these genes are considered our strongest candidates to modulate infarct volume after ischemic stroke via a collateral vessel-independent mechanism.

Discussion

Occlusion of the blood vessel in the brain leads to ischemic stroke that results in infarction – irreversible death of the brain tissue. Through decades of research to identify risk factors causing ischemic stroke, our understanding of the molecular events occurring during infarction has dramatically increased. However, although studies of genetic risk factors for stroke susceptibility in the human have been attempted using both family-based linkage (Gretarsdottir *et al.* 2003; Helgadottir *et al.* 2004) and population-based genome-wide association studies (GWAS) (Matarin *et al.* 2007; Matarin *et al.* 2008a; Matarin *et al.* 2008b), the use of these studies to discover druggable targets to treat stroke outcome have been largely disappointing. Furthermore, despite an explosion of the human GWAS for disease phenotypes, no published GWAS for infarct volume in ischemic stroke has been published. Genetic studies in the human related to infarct volume are intrinsically problematic due to uncontrollable variation in the extent and location of the occluded vessel, and especially, variation in the critical time between first recognized symptoms of stroke and medical intervention. Instead, current understanding of the mechanisms underlying infarct damage is primarily based on experimental animal models.

We have employed a forward genetic mapping approach (QTL analysis) to discover novel genes modulating infarction using surgical occlusion mouse model of cerebral ischemia. Through surgical occlusion of the distal middle cerebral artery (pMCAO) in 32 commonly used mouse inbred strains, we have shown that inbred mouse strains demonstrate robust differences in infarct volume (Keum and Marchuk 2009; Keum *et al.* 2013). Exploiting these strain differences, we have mapped several genetic loci (QTLs) that modulate infarct volume in ischemic stroke (Keum and Marchuk 2009; Chu *et al.* 2013; Keum *et al.*

2013). One of these loci located on distal chromosome 7 (cerebral infarct volume QTL 1 (*Civq1*)) is highly significant in regulating infarct volume in multiple pairwise crosses of inbred mouse strains (Keum and Marchuk 2009; Keum *et al.* 2013). Interestingly, Civq1 is identically overlapped with a locus on chromosome 7 that modulates cerebral collateral vessel number (collateral artery number QTL 1 (*Canq1*)) (Keum and Marchuk 2009; Zhang *et al.* 2010; Keum *et al.* 2013; Lee *et al.* 2016). In general, collateral vessels can serve as a compensatory circulatory shunt when a larger vessel is blocked, enabling reperfusion of the ischemic cerebral territory. Thus, it is conceivable that cerebral infarct volume is correlated with the number of collateral vessels in most inbred mouse strains. Canq1 has recently been shown to be due to variation in the RABEP2 gene (Lucitti *et al.* 2016), encoding a protein that plays a role in endosomal recycling of VEGFR2 (Kofler *et al.* 2018). This gene also strongly influences the infarct volume phenotype via the reperfusion mechanism discussed above, although we have identified 3 other genes mapping within *Civq1*, *Il21r*, *Il4r*α, and *Itgal*, that also modulate infarct volume mainly via a collateral-independent manner (Keum *et al.* 2013; Lee *et al.* 2016; Lee *et al.* 2018).

Since the collateral vasculature is developmentally established and consequently, may be difficult to modify therapeutically, we have attempted to identify novel genes modulating infarct volume in ischemic stroke through a collateral-independent manner (ie. via neuroprotection). Given the strong influence that the collateral circulation plays in modulating infarct volume across most inbred mouse strains, we sought strains that appear to break the inverse correlation between collateral vessel density and infarct volume. Previously, we found one such strain, C3H, that breaks this inverse correlation. C3H mice show a similar number of collateral vessel connections compared to the B6

strain, while paradoxically exhibiting a 4-fold increase in infarct volume after pMCAO (Chu *et al.* 2013). Genome-wide linkage for infarct volume after pMCAO in F2 progeny of a cross between B6 and C3H identified a neuroprotective locus, *Civq4*, on chromosome 8 (Chu *et al.* 2013). *Civq4* was identified in a cross between two domesticated inbred mouse strains that overall exhibit relatively low sequence variation compared to that found within the entire species. These and other commonly used inbred mouse strains contain 94% of their genetic background from *Mus musculus domesticus*, 5% from *Mus musculus musculus*, and less than 1% from *Mus musculus castaneus* (Yang *et al.* 2011). Thus, genetic mapping using these inbred mouse strains may systemically miss potential candidate loci that might cause phenotypic variation within the entire species.

Here, to increase the genetic diversity of our mapping approach, we utilized eight founder inbred mouse strains of the Collaborative Cross mapping population. These strains include five domesticated inbred strains (A/J, B6, NOD, NZO, and 129S1) and three wild-derived strains from Mus musculus subspecies (PWK, WSB, and CAST). Specifically, these wild-derived strains, PWK (*musculus*), WSB (*domesticus*), and CAST (*castaneous*), are more genetically diverse than the classical inbred strains (Churchill *et al.* 2004; Threadgill 2004; Aylor *et al.* 2011; Collaborative Cross Consortium 2012). Using all eight founder strains of the CC mouse, we surveyed both the number of collateral vessel connections as well as infarct volume after pMCAO. The number of collateral vessel connection of two wild-derived strains, CAST and WSB, were the highest we have observed in any inbred mouse strains. Consistent with the role of collateral vessels in reperfusion of the ischemic territory following pMCAO, CAST mice show the lowest infarct volume that we have ever measured. However, although CAST and WSB have the largest

number of collateral vessel connections, unlike CAST, upon pMCAO, WSB shows a large infarct volume. Therefore, we may postulate that this wild-derived WSB strain breaks the inverse correlation between collateral vessel connections and infarct volume. Supportive evidence for a non-collateral mechanism is the observation that infarct volume in the F2 animals shows a broad range of values while the collateral vessel connections are nearly invariant among F2 animals. Based on the ability to reperfuse the ischemic tissue due to more than sufficient collateral vessel connections, but a much larger infarct volume than would be predicted, WSB likely contains loci that upon ischemic insult, contribute to neuronal death. One compelling hypothesis is that most strains contain natural neuroprotective loci that are defective in WSB. In this F2 cross, we identified the most significant four novel neuroprotective loci (*Civq8* on chromosome 1, *Civq9* on chromosome 6, *Civq10* on chromosome 13, and *Civq11* on chromosome 17). Importantly, these loci identified in an F2 cross between CAST and WSB did not overlap with other loci that we previously identified with other F2 crosses. More importantly, the *Civq1* / *Canq1* locus, the strongest locus controlling collateral vessel number, is absent in this cross.

From the linkage analysis for infarct volume, we identified four candidate intervals regulating infarct volume through a collateral-independent mechanism. In order to identify candidate genes within the 1.5 LOD-support intervals, we searched for coding SNP differences between CAST and WSB focusing on genes where the functional consequence of the variants were predicted using 3 different *in silico* prediction algorithms (Ng and Henikoff 2003; Adzhubei *et al.* 2010; Choi *et al.* 2012). Coding SNPs predicted

as “damaging” by all 3 algorithms were considered the highest priority candidates, with full realization that any of the other genes harboring coding SNPs cannot be ignored.

As the genes underlying these four loci might harbor cis-regulatory sequence variation modulating differences in mRNA levels, we determined strain-specific differential transcript levels between CAST and WSB using RNA sequencing analysis. Genes exhibiting a statistically significant, 2-fold or greater difference in transcript levels are considered our highest priority candidate genes. Finally, to identify a short list of candidate genes for initial consideration for each of the loci, we sought the intersection of the predicted functional coding SNPs and those exhibiting strain-specific transcript levels. Through this association, functional studies can be focused on only 15 genes mapping within any of the intervals (Figure 5B).

In summary, we identified a wild-derived mouse strain, WSB, that breaks inverse correlation between a collateral vessel connections and infarct volume after pMCAO. By selecting two wild-derived mouse strains, CAST and WSB, that contain similarly high numbers of collateral vessel connections but exhibit large differences in infarct volume after pMCAO, we identified a strain pair that might uncover novel loci involved in neuroprotetion. In a cross between these strains, we discovered four novel neuroprotective genetic loci. Using RNA sequencing of brain tissue and *in silico* analyses of coding SNPs, we further prioritized the genes mapping within these intervals. The identification of such neuroprotective genes may provide novel targets for future therapeutic intervention for human ischemic stroke.

Acknowledgments

The authors thank Dr. Sena Bae for helpful discussion concerning data sorting and analysis. This work was supported by a grant from NIH grant 5R01HL097281 and the Foundation Leducq Transatlantic Network of Excellence in Neurovascular Disease (17 CVD 03).

References

Adzhubei, I.A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, *et al.*, 2010 A method and server for predicting damaging missense mutations. Nat Methods. 7(4):248-9.

Anders, S., P. T. Pyl, and W. Huber, 2015 HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 31(2):166-9.

Aylor, D. L., W. Valdar, W. Foulds-Mathes, R. J. Buus, R. A. Verdugo, *et al.*, 2011 Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. 21(8):1213–22.

Barone, F.C., D. J. Knudsen, A. H. Nelson, G. Z. Feuerstein, and R. N. Willette, 1993 Mouse strain differences in susceptibility to cerebral ischemia are related to cerebral vascular anatomy. J Cereb Blood Flow Metab. 13(4):683-92.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B. 57(1): 289-300.

Carmichael, S.T., 2005 Rodent models of focal stroke: size, mechanism, and purpose. NeuroRx. 2(3), 396-409.

Choi, Y., G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, 2012 Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One. 7(10):e46688.

Chu, P.L., S. Keum, and D. A. Marchuk, 2013 A novel genetic locus modulates infarct volume independently of the extent of collateral circulation. Physiol Genomics. 45(17):751-63.

Churchill, G. A., D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie *et al.*, 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat. Genet. 36(11):1133-7.

Clayton, J. A., D. Chalothorn, and J. E. Faber, 2008 Vascular Endothelial Growth Factor-A Specifies Formation of Native Collaterals and Regulates Collateral Growth in Ischemia. Circ Res. 103(9):1027-36.

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics. 190(2):389-401.

Dichgans, M., 2007 Genetics of ischaemic stroke. Lancet Neurol. 6(2):149-61.

Gretarsdottir, S., G. Thorleifsson, S. T. Reynisdottir, A. Manolescu, S. Jonsdottir, *et al.*, 2003 The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. Nat Genet. 35(2):131-138.

Helgadottir, A., A. Manolescu, G. Thorleifsson, S. Gretarsdottir, H. Jonsdottir, *et al.*, 2004 The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. Nat Genet. 36(3):233-9.

Holt, J., S. Huang, L. McMillan, and W. Wang, 2013 Read annotation pipeline for high-throughput sequencing data, pp. 605-612 in International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, edited by ACM, New York.

Huang, S., J. Holt, C. Y. Kao, L. McMillan, and W. Wang, 2014 A novel multi-alignment pipeline for high-throughput sequencing data. Database (Oxford). bau057.

Jamrozik, K., R. J. Broadhurst, C. S. Anderson, and E. G. Stewart-Wynne. 1994 The role of lifestyle factors in the etiology of stroke. A population-based case-control study in Perth, Western Australia. Stroke. 25(1):51-9.

Johnson, W., O. Onuma, M. Owolabi, and S. Sachdev, 2016 Stroke: a global response is needed. Bull. World Health Organ. 94(9):634-634A.

Keum, S., and D. A. Marchuk, 2009 A locus mapping to mouse chromosome 7 determines infarct volume in a mouse model of ischemic stroke. Circ Cardiovasc Genet. 2(6):591-8.

Keum, S., H. K. Lee, P. L. Chu, M. J. Kan, M. N. Huang, *et al.*, 2013 Natural genetic variation of integrin alpha L (Itgal) modulates ischemic brain injury in stroke. PLoS Genet. 9(10):e1003807.

Kofler, N., F. Corti, F. Rivera-Molina, Y. Deng, D. Toomre, *et al.*, 2018 The Rab-effector protein RABEP2 regulates endosomal trafficking to mediate vascular endothelial growth factor receptor-2 (VEGFR2)-dependent signaling. J Biol Chem. 293(13):4805-4817.

Lambertsen, K. L., R. Gregersen, and B. Finsen, 2002 Microglial-macrophage synthesis of tumor necrosis factor after focal cerebral ischemia in mice is strain dependent. J Cereb Blood Flow Metab. 22(7), 785-97.

Lambertsen, K.L., M. Meldgaard, R. Ladeby, and B. Finsen, 2005 A quantitative study of microglial-macrophage synthesis of tumor necrosis factor during acute and late focal cerebral ischemia in mice. J Cereb Blood Flow Metab. 25(1):119-35.

Lee, H.K., S. Keum, H. Sheng, D. S. Warner, D. C. Lo, *et al.*, 2016 Natural allelic variation of the IL-21 receptor modulates ischemic stroke infarct volume. J Clin Invest. 126(8):2827-38.

Lee, H.K., S. Koh, D. C. Lo, and D. A. Marchuk, 2018 Neuronal IL-4Rα modulates neuronal apoptosis and cell viability during the acute phases of cerebral ischemia. FEBS J. 285(15):2785-2798.

Love, M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12):550.

Lucitti, J.L., R. Sealock, B. K. Buckley, H. Zhang, L. Xiao, *et al.*, 2016 Variants of Rab GTPase-Effector Binding Protein-2 Cause Variation in the Collateral Circulation and Severity of Stroke. Stroke. 47(12):3022-3031.

Majid, A., Y. Y. He, J. M. Gidday, S. S. Kaplan, E. R. Gonzales, *et al.*, 2000 Differences in vulnerability to permanent focal cerebral ischemia among 3 common mouse strains. Stroke. 31(11):2707-14.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10-12.

Matarin, M., W. M. Brown, S. Scholz, J. Simon-Sanchez, H. C. Fung, *et al.*, 2007 A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. Lancet Neurol. 6(5):414-20.

Matarin, M., W. M. Brown, A. Singleton, J. A. Hardy, and J. F. Meschia, 2008a Whole genome analyses suggest ischemic stroke and heart disease share an association with polymorphisms on chromosome 9p21. Stroke. 39(5):1586-9.

Matarin, M., J. Simon-Sanchez, H. C. Fung, S. Scholz, J. R. Gibbs, *et al.*, 2008b Structural genomic variation in ischemic stroke. Neurogenetics. 9(2):101-8.

Mozaffarian, D., C. D. Furberg, B. M. Psaty, and D. Siscovick, 2008 Physical activity and incidence of atrial fibrillation in older adults: the Cardiovascular Health Study. Circulation. 118(8):800–7.

Mozaffarian, D., E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, *et al.*, 2015 Heart Disease and Stroke Statistics--2015 update: a report from the American Heart Association. Circulation. 131(4):e29-322.

Ng, P. C., and S. Henikoff, 2003 SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31(13):3812–4.

Roger, V.L., A. S. Go, D. M. Lloyd-Jones, E. J. Benjamin, J. D. Berry, *et al.*, 2012 Heart disease and stroke statistics--2012 update: a report from the American Heart Association. Circulation. 125(1):e2-e220.

Sugimori, H., H. Yao, H. Ooboshi, S. Ibayashi, and M. Iida, 2004 Krypton laser-induced photothrombotic distal middle cerebral artery occlusion without craniectomy in mice. Brain Res Brain Res Protoc. 13(3):189-96.

Threadgill, D. W., 2006 Meeting report for the 4th annual Complex Trait Consortium meeting: from QTLs to systems genetics. Mamm. Genome. 17(1): 2–4.

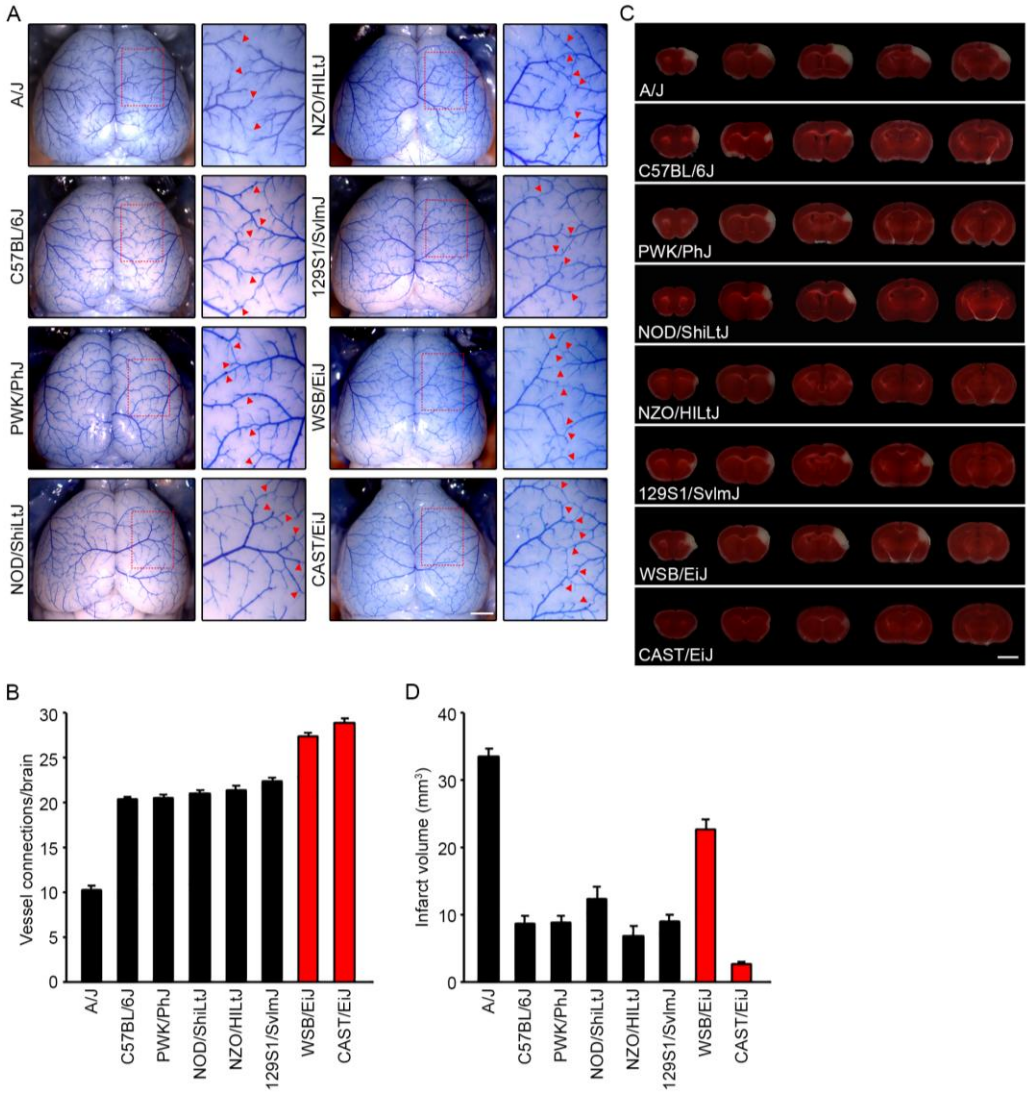
Trapnell, C., L. Pachter, and S. L. Salzberg, 2009 TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 25(9):1105-11.

Wexler, E.J., E. E. Peters, A. Gonzales, M. L. Gonzales, A. M. Slee, and J. S. Kerr, 2002 An objective procedure for ischemic area evaluation of the stroke intraluminal thread model in the mouse and rat. J Neurosci Methods. 113(1):51-8.

Yang, H., J. R. Wang, J. P. Didion, R. J. Buus, T. A. Bell, *et al.*, 2011. Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet. 43(7): 648–55.

Zhang, H., P. Prabhakar, R. Sealock, and J. E. Faber, 2010 Wide genetic variation in the native pial collateral circulation is a major determinant of variation in severity of stroke. J Cereb Blood Flow Metab. 30(5):923-34.

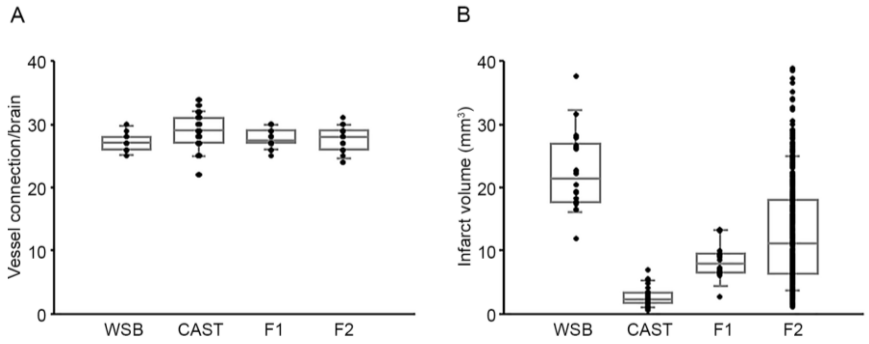
Figure 1



Collateral vessel density and infarct volume after pMCAO in 8 parental strains of the Collaborative Cross mouse. (A) Representative images of the brains for the 8 parental strains of the CC recombinant inbred mapping panel; A/J, B6, PWK, NOD, NZO, 129S1, WSB, and CAST. For each strain, the area outlined in a red box in the leftmost image is 3-fold magnified in the image to the right, where the red arrowheads indicate collateral vessel connections between the ACA and the MCA. Scale bar: 1mm. **(B)** The

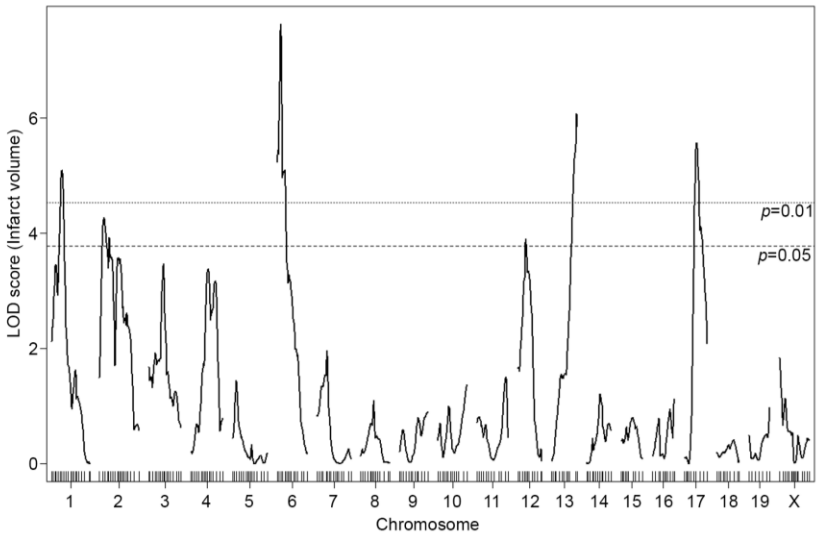
graph indicates the average number of collateral vessel connections between ACA and MCA in the brain. The total number of animals for; A/J, B6, PWK, NOD, NZO, 129S1, WSB, and CAST were 21, 37, 24, 9, 14, 31, 13, and 32 mice, respectively. Data represent the mean ± SEM. **(C)** Serial brain sections (1 mm) for each parental strain of the CC mouse 24 h after pMCAO. The infarct appears as white tissue after 2% TTC staining. Scale bar: 5 mm. **(D)** The graph shows the infarct volume for each parental strain of the CC mouse. The total number of animals for; A/J, B6, PWK, NOD, NZO, 129S1, WSB, and CAST were 19, 32, 27, 9, 10, 32, 18, and 32 animals, respectively. Data represent the mean ± SEM.

Figure 2



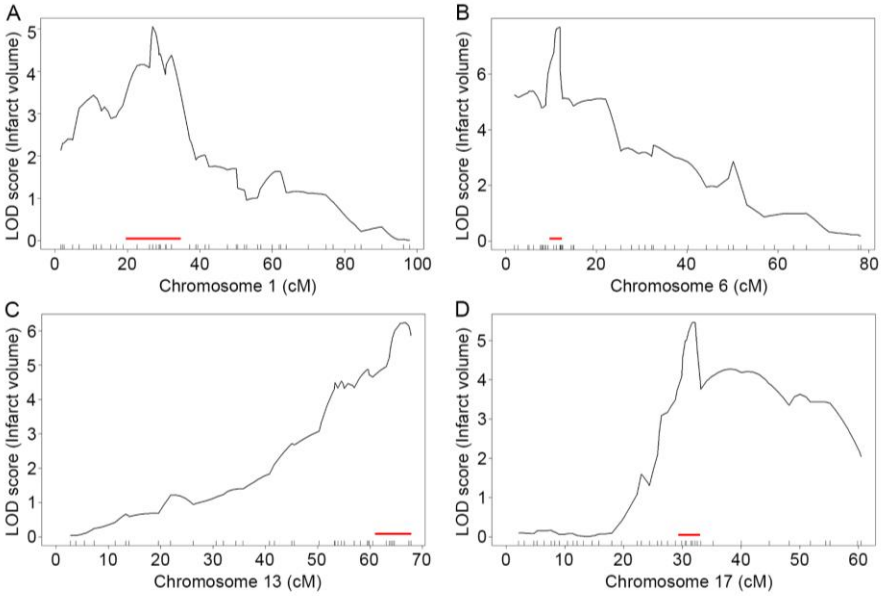
F2 intercross animals between CAST and WSB exhibit a tight distribution of collateral vessel connections but a wide distribution for pMCAO-induced infarct volume. (A) The dot plot graph shows the number of collateral vessel connections between the ACA and MCA for parental strains, WSB and CAST, and their F1 and F2 intercross progeny. The box plot indicates the degree of dispersion and skewness. The number of animals for WSB, CAST, F1 and F2 was 13, 32, 18, 24 animals, respectively. Data represent the mean ± SEM. **(B)** The dot and box plot graph show the distribution of pMCAO-induced infarct volume for WSB, CAST, and their F1 and F2 intercross progeny. The number of animals for the infarct volume measurements was 18, 32, 14, and 251 animals, respectively. Data represent the mean ± SEM.

Figure 3



Identification of four novel *Civq* loci modulating infarct volume in the F2 intercross between CAST and WSB. The graph represents the analysis of a genome-wide linkage scan for infarct volume measured 24 h after pMCAO using 251 F2 progeny (CAST and WSB). Chromosomes 1 through X are indicated numerically on the x axis. The y axis shows a logarithm of the odds (LOD) score and the significant levels ($p < 0.05$ and 0.01) were determined by 1000 permutation tests. Four regions of the genome mapping to chromosomes 1, 6, 13, and 17 display highly significant linkage to the infarct volume trait with LOD scores of 5.09, 7.68, 6.23, and 5.45, respectively.

Figure 4



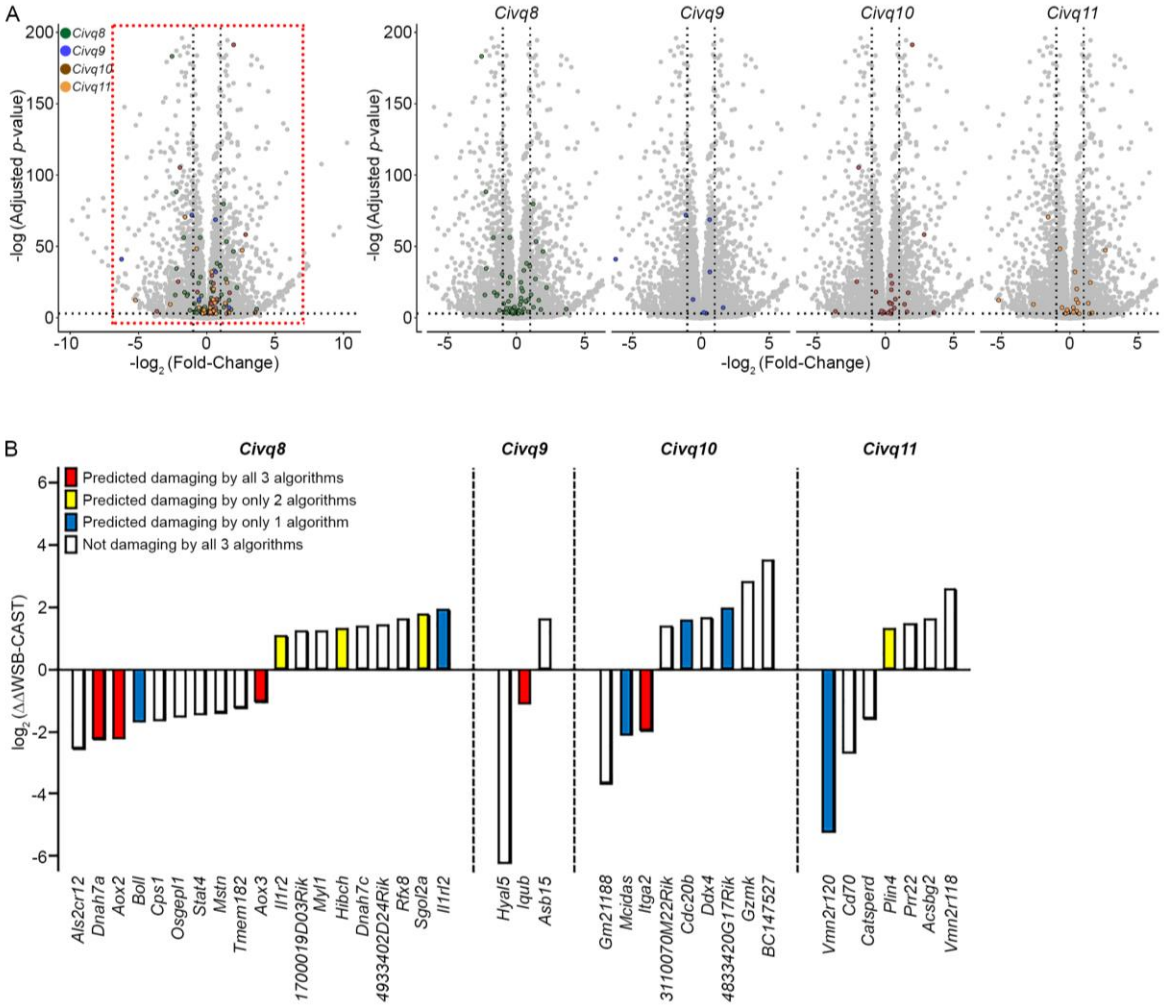
E

QTL	Chromosome interval (GRCm38/mm10)	Different non-synonymous SNPs between CAST and WSB	Differential gene expressions between CAST and WSB (>2-*fold*)
Civq8	1:39,000,000-70,000,000 (31 Mb)	330 coding SNPs in 90 coding genes	47 genes
Civq9	6:21,000,000-29,000,000 (8 Mb)	77 coding SNPs in 20 coding genes	3 genes
Civq10	13:108,000,000-120,421,639 (22.4 Mb)	109 coding SNPs in 34 coding genes	19 genes
Civq11	17:54,000,000-65,000,000 (11 Mb)	157 coding SNPs in 47 coding genes	10 genes

Fine-mapping and potential candidate genes for the four novel *Civq* loci. (A) The graph shows the linkage map across chromosome 1 using 45 informative SNP markers, highlighting the *Civq8* locus. The LOD score at the peak is 5.09 (JAX00250952), and the 1.5-LOD support interval is from 19.627 to 34.627 cM, indicated by red bar on the graph. **(B)** The graph shows the linkage map across chromosome 6 using 46 informative SNP markers, highlighting the *Civq9*. The LOD score at the peak is 7.68 (UNC_rs51474193), and the 1.5-LOD support interval is from 9.80 to 11.95 cM, indicated by red bar on the graph. **(C)** The graph shows the linkage map across chromosome 13 using 38 informative SNP markers, highlighting the *Civq10* locus. The LOD score at the peak is 6.23 (UNC130410312), and the 1.5-LOD support interval is from 61.50 to 67.83 cM, indicated

by red bar on the graph. **(D)** The graph shows the linkage map across chromosome 17 using 40 informative SNP markers, highlighting the *Civq11* locus. The LOD score at the peak is 5.45 (UNC170143495), and the 1.5-LOD support interval is from 29.50 to 33.10 cM, indicated by red bar on the graph. **(E)** The table provides information for each of the four loci including confidence intervals, the number of genes harboring coding SNP differences between CAST and WSB, and the number of genes showing at least 2-fold expression difference between brain cortex of CAST and WSB.

Figure 5



RNA sequencing data identifies genes showing strain-specific differences in cerebral cortex gene expression. (A) The volcano plot shows differential gene expression between CAST and WSB from brain cortex determined by RNA sequencing analysis. Each dot represents a different gene with the $\log_2$ fold change plotted against $\log_{10}$ *p*-value. The total number of significantly differential expressed genes across the mouse genome is 9,563 genes but only 220 genes map within one of the four QTL intervals. Differentially-expressed genes in the red box in the leftmost plot are separated

for each interval on the right; green dots indicate 135 genes mapping within *Civq8*, blue dots indicate 10 genes mapping within *Civq9*, brown dots indicate 47 genes mapping within *Civq10*, and orange dots indicate 28 genes mapping within *Civq11*. **(B)** For each locus, this graph identifies genes that show at least a 2-fold difference in expression between CAST and WSB and also harbor a potentially damaging coding SNP. The bar goes downward (negative value) for genes with higher gene expression of the CAST allele and the bar goes upward (positive value) for genes with higher gene expressions of the WSB allele. The different colors represent how many of the three *in silico* algorithms used (PolyPhen-2, SIFT, and PROVEAN) predicted that at least one coding SNP in the gene was damaging.

Table 1

Candidate genes harboring coding SNPs that are predicted to be damaging by three different *in silico* algorithms.

	Predicted damaging by all 3 *in silico* algorithms			Predicted damaging by only 2 *in silico* algorithms			Predicted damaging by only 1 *in silico* algorithm		
Civq8	*Aox2*	*Aox3*	*Aox4*	*Aox1*	*Bmpr2*	*Ccdc150*	*Boll*	*Carf*	*Ccdc168*
	Casp8	*Crygb*	*Dnah7a*	*Ercc5*	*Gm973*	*Hibch*	*Clfar*	*Col5a2*	*Cyp20a1*
	Dytn	*Il1r1*	*Mpp4*	*Il18rap*	*Il1r2*	*Il1rl1*	*D630023F18Rik*	*Dnah7b*	*Il18r1*
	Nabp1	*Nbeal1*		*Pard3b*	*Pms1*	*Sgol2a*	*Il1rl2*	*Map4k4*	*Mett21c*
				Timem237	*Trak2*	*Wdr75*	*Mettl21e*	*Mfsd9*	*Plcl1*
							Rftn2	*Slc39a10*	*Spag16*
							Spats2l	*Zdbf2*	
Civq9	*Iqub*			*Hyal4*	*Pax4*	*Ptprz1*	*Cadps2*	*Cped1*	*Pot1a*
Civq10	*Gzma*	*Itga1*	*Itga2*	*Mrps30*			*4833420G17Rik*	*Cdc20b*	*Dhx29*
	Map3k1						*Emb*	*Gapt*	*Hcn1*
							Il31ra	*Mcidas*	*Nnt*
							Pde4d	*Tcstv3*	
Civq11	*Adgre1*	*Arrdc5*	*Ticam1*	*Cntnap5c*	*Dennd1c*	*Pdzph1*	*1700061G19Rik*	*4930583I09Rik*	*Acer1*
				Plin4	*Pspn*	*Shd*	*Adgre4*	*Chaf1a*	*Clpp*
				Zfp119b			*Ebi3*	*Fer*	*Fsd1*
							Mpnd	*Plin5*	*Safb2*
							Stap2	*Tnfsf9*	*Vmn2r120*
							Zfp119a	*Zfp959*	

Detailed coding SNP information for each gene is available in Table S1.

Table 2

Differential gene expression between CAST and WSB determined by RNA sequencing analysis.

Chr	Symbol	Description	Gene type	log_2 FC (fold change)
1	*Gm37373*	Predicted gene, 37373	TEC	-4.36
	Gm15759	Predicted gene 15759	antisense_RNA	-3.07
	Gm28802	Predicted gene 28802	unprocessed_pseudogene	-2.75
	Als2cr12	Amyotrophic lateral sclerosis 2 chromosome region 12	protein_coding	-2.55
	4930558J18Rik	RIKEN cDNA 4930558J18 gene	lincRNA	-2.31
	Gm37915	Predicted gene, 37915	antisense_RNA	-2.28
	Dnah7a	Dynein, axonemal, heavy chain 7A	protein_coding	-2.24
	Aox2	Aldehyde oxidase 2	protein_coding	-2.23
	Gm26813	Predicted gene, 26813	lincRNA	-1.80
	Gm8241	Predicted gene 8241	processed_pseudogene	-1.72
	Boll	Boule homolog, RNA binding protein	protein_coding	-1.69
	Gm8419	Predicted gene 8419	processed_pseudogene	-1.65
	Cps1	Carbamoyl-phosphate synthetase 1	protein_coding	-1.64
	C2cd6b	C2 calcium dependent domain containing 6B	lincRNA	-1.60
	E330011M16Rik	RIKEN cDNA E330011M16 gene	TEC*	-1.59
	Gm27607	Predicted gene, 27607	snoRNA	-1.55
	Osgepl1	O-sialoglycoprotein endopeptidase-like 1	protein_coding	-1.52
	Stat4	Signal transducer and activator of transcription 4	protein_coding	-1.46
	Mstn	Myostatin	protein_coding	-1.40
	Gm35801	Predicted gene, 35801	antisense_RNA	-1.28
	Tmem182	Transmembrane protein 182	protein_coding	-1.24
	Gm11600	Predicted gene 11600	processed_pseudogene	-1.24
	Aox3	Aldehyde oxidase 3	protein_coding	-1.05
6	*Hyal5*	Hyaluronoglucosaminidase 5	protein_coding	-6.24
	Iqub	IQ motif and ubiquitin domain containing	protein_coding	-1.10
13	*Gm47776*	Predicted gene, 47776	processed_pseudogene	-3.67
	Gm21188	Predicted gene, 21188	protein_coding	-3.66
	Mcidas	Multiciliate differentiation and DNA synthesis associated cell cycle protein	protein_coding	-2.10
	Itga2	Integrin alpha 2	protein_coding	-1.97
	Gm6270	Predicted gene 6270	lincRNA	-1.84
	Gm15290	Predicted gene 15290	lincRNA	-1.69
	Gm15488	Predicted gene 15488	unitary_pseudogene	-1.69
	Gm15285	Predicted gene 15285	processed_pseudogene	-1.25
17	*Vmn2r120*	Vomeronasal 2, receptor 120	protein_coding	-5.23
	Nudt12os	Nudix (nucleoside diphosphate linked moiety X)-type motif 12, opposite strand	antisense_RNA	-2.98
	Cd70	CD70 antigen	protein_coding	-2.68
	Catsperd	Cation channel sperm associated auxiliary subunit delta	protein_coding	-1.59

* To be experimentally confirmed

(A) The table lists all genes showing at least 2-fold differential expression where the CAST allele is higher than WSB.

Chr	Symbol	Description	Gene type	log_2 FC (fold change)
1	*Rps27a-ps1*	Ribosomal protein S27A, pseudogene 1	processed_pseudogene	3.82
	1700066B17Rik	RIKEN cDNA 1700066B17 gene	lincRNA	3.64
	Gm553	Predicted gene 553	lincRNA	3.07
	Gm29018	Predicted gene 29018	antisense_RNA	2.77
	Gm28177	Predicted gene 28177	processed_transcript	2.49
	BC055402	CDNA sequence BC055402	lincRNA	2.19
	Il1rl2	Interleukin 1 receptor-like 2	protein_coding	1.94
	D930019O06Rik	RIKEN cDNA D930019O06	antisense_RNA	1.91
	Sgo2a	Shugoshin 2A	protein_coding	1.78
	Rfx8	Regulatory factor X 8	protein_coding	1.63
	9530026F06Rik	RIKEN cDNA 9530026F06 gene	lincRNA	1.60
	4933402D24Rik	RIKEN cDNA 4933402D24 gene	protein_coding	1.43
	Dnah7c	Dynein, axonemal, heavy chain 7C	protein_coding	1.39
	Gm15832	Predicted gene 15832	processed_transcript	1.39
	Hibch	3-hydroxyisobutyryl-Coenzyme A hydrolase	protein_coding	1.30
	Myl1	Myosin, light polypeptide 1	protein_coding	1.24
	1700019D03Rik	RIKEN cDNA 1700019D03 gene	protein_coding	1.22
	Gm28826	Predicted gene 28826	lincRNA	1.21
	Il1r2	Interleukin 1 receptor, type II	protein_coding	1.07
	Gm28449	Predicted gene 28449	lincRNA	1.07
	Gm28982	Predicted gene 28982	lincRNA	1.05
	2810408I11Rik	RIKEN cDNA 2810408I11 gene	processed_transcript	1.04
	Gm18849	Predicted gene, 18849	unprocessed_pseudogene	1.03
	C230029F24Rik	RIKEN cDNA C230029F24 gene	lincRNA	1.02
6	*Asb15*	Ankyrin repeat and SOCS box-containing 15	protein_coding	1.63
13	*BC147527*	CDNA sequence BC147527	protein_coding	3.53
	Gzmk	Granzyme K	protein_coding	2.83
	Gm15323	Predicted gene 15323	lincRNA	2.28
	4833420G17Rik	RIKEN cDNA 4833420G17 gene	protein_coding	1.96
	Ddx4	DEAD (Asp-Glu-Ala-Asp) box polypeptide 4	protein_coding	1.66
	Cdc20b	Cell division cycle 20B	protein_coding	1.57
	Gm47914	Predicted gene, 47914	lincRNA	1.52
	3110070M22Rik	RIKEN cDNA 3110070M22 gene	protein_coding	1.39
	Gm15327	Predicted gene 15327	processed_transcript	1.16
	4930467J12Rik	RIKEN cDNA 4930467J12 gene	lincRNA	1.07
	Gm15287	Predicted gene 15287	antisense_RNA	1.03
17	*Vmn2r118*	Vomeronasal 2, receptor 118	protein_coding	2.58
	Acsbg2	Acyl-CoA synthetase bubblegum family member 2	protein_coding	1.62
	Prr22	Proline rich 22	protein_coding	1.49
	Gm16712	Predicted gene, 16712	lincRNA	1.44
	Plin4	Perilipin 4	protein_coding	1.34
	Gm17168	Predicted gene 17168	antisense_RNA	1.13

(B) The table lists all genes showing at least 2-fold differential expression where the WSB allele is higher than CAST.

Additional information for all 220 genes is available in Table S2.