scientific reports

OPEN

Extensive variation in the intelectin gene family in laboratory and wild mouse strains

Faisal Almalki[1,2,6], Eric B. Nonnecke[3,6], Patricia A. Castillo[3], Alex Bevin-Holder[1], Kristian K. Ullrich[4], Bo Lönnerdal[5], Linda Odenthal-Hesse[4], Charles L. Bevins[3]✉ & Edward J. Hollox[1]✉

Intelectins are a family of multimeric secreted proteins that bind microbe-specific glycans. Both genetic and functional studies have suggested that intelectins have an important role in innate immunity and are involved in the etiology of various human diseases, including inflammatory bowel disease. Experiments investigating the role of intelectins in human disease using mouse models are limited by the fact that there is not a clear one-to-one relationship between intelectin genes in humans and mice, and that the number of intelectin genes varies between different mouse strains. In this study we show by gene sequence and gene expression analysis that human intelectin-1 (*ITLN1*) has multiple orthologues in mice, including a functional homologue *Itln1*; however, human intelectin-2 has no such orthologue or homologue. We confirm that all sub-strains of the C57 mouse strain have a large deletion resulting in retention of only one intelectin gene, *Itln1*. The majority of laboratory strains have a full complement of six intelectin genes, except CAST, SPRET, SKIVE, MOLF and PANCEVO strains, which are derived from different mouse species/subspecies and encode different complements of intelectin genes. In wild mice, intelectin deletions are polymorphic in *Mus musculus castaneus* and *Mus musculus domesticus*. Further sequence analysis shows that *Itln3* and *Itln5* are polymorphic pseudogenes due to premature truncating mutations, and that mouse *Itln1* has undergone recent adaptive evolution. Taken together, our study shows extensive diversity in intelectin genes in both laboratory and wild-mice, suggesting a pattern of birth-and-death evolution. In addition, our data provide a foundation for further experimental investigation of the role of intelectins in disease.

Intelectins (*inte*stinal *lectins*) are a family of calcium-dependent multimeric, secreted proteins that selectively bind microbial carbohydrates[1–3]. In mammals, intelectins were initially identified in the mouse small intestine, but are found throughout vertebrates, and have a variety of roles including host-microbe interactions[3–9]. In humans, there are two intelectin proteins (intelectin-1 and intelectin-2) encoded by the genes *ITLN1* and *ITLN2* that are tandemly arranged on chromosome 1q23.3. Human ITLN1, also known as omentin-1, is present in visceral adipose as well as the intestine[1,10–12]. ITLN1 binds to exocyclic vicinal 1,2-diols, a chemical moiety present in microbial carbohydrate-containing structures such as β-ᴅ-galactofuranose, which is a galactose isomer synthesised by microorganisms, including protozoa, fungi, and bacteria, but not by mammalian cells[5]. The tissue expression and lectin binding properties of ITLN2 appear different[13] but are yet to be delineated. Intelectins have been implicated in several diseases including inflammatory bowel disease (IBD), obesity, non-insulin-dependent diabetes mellitus, and asthma[4,8,12,13].

Expression in the small intestine, and the ability to recognise bacteria-specific glycans, suggests that ITLN1 has a key role in the innate immune response in the gut. Furthermore, genome-wide association studies (GWAS) have identified a number of associations with inflammatory bowel diseases. In particular, an early GWAS identified a common single nucleotide variant allele rs2274910-C associated with increased risk of Crohn's disease[14,15]. Subsequent GWASs have identified other variants within and surrounding *ITLN1* associated with Crohn's disease

[1]Department of Genetics and Genome Biology, University of Leicester, Leicester, UK. [2]Medical Laboratories Technology Department, College of Applied Medical Sciences, Taibah University, Almadinah Almunwarah, Saudi Arabia. [3]Department of Microbiology and Immunology, School of Medicine, University of California, Davis, CA, USA. [4]Max Planck Institute for Evolutionary Biology, Plön, Germany. [5]Department of Nutrition, University of California, Davis, CA, USA. [6]These authors contributed equally: Faisal Almalki and Eric B. Nonnecke. ✉email: clbevins@ucdavis.edu; ejh33@le.ac.uk

and ulcerative colitis[16–18]. It is likely that these associations are driven by the same causative variant, but that causative variant has not yet been identified[19].

The link between intelectins and diseases such as IBD suggests that it will be informative to pursue further research on intelectin function in vivo using animal models. Nevertheless, analysis of mouse intelectins is challenging as mice show strain-specific variation in intelectin gene copy number and tissue-specific expression patterns. The laboratory mouse strain 129S7/Sv encodes six intelectin genes on chromosome 1, generated by recurrent inversion and duplication, whereas the C57BL/6J sub-strain encodes a single intelectin gene, *Itln1*, remaining from a large 420 kb deletion[8]. The assignment of particular mouse intelectin genes as orthologues of human *ITLN1* remains unclear.

We therefore considered it important to fully characterise the intelectin gene locus in both laboratory strains and wild mice and investigate the complement of intelectin genes across these mice, as well as sequence and expression variation. We characterised expression patterns of the intelectin genes in different tissues of different strains, with a particular focus on the gastrointestinal tract, where intelectins appear to be commonly expressed in vertebrates[3]. This characterisation will provide a firm grounding for experiments using mouse models to identify or study intelectins in gastrointestinal, lung, metabolic, and infectious diseases. Importantly, because analyses of the genomes of laboratory mice and wild mice relied on mapping to a reference genome from the C57BL/6J sub-strain, all members of the intelectin family except *Itln1* appear to be absent from these genomes. In 15 laboratory strains and 29 wild mice we mapped publicly available short sequence read data to a sequence contig of the intelectin gene region derived previously from a 129S7/Sv strain mouse. We confirmed and characterised the deletion in C57BL/6J, show it is present also in three progenitor C57 strains (i.e., C58/J, C57L/J, C57BR/cdJ), and found novel deletions in both wild and laboratory strains of mice. Moreover, we characterised the sequence variation of intelectin genes, in the context of the repeated nature of this region.

Results

Human ITLN1, but not ITLN2, has multiple orthologues in mice.

We selected 20 coding sequences of representative intelectin genes from six primates and nine rodents. A phylogenetic tree shows that Muridae rodents have one intelectin gene except in mice where a recent burst of duplications generated six intelectin genes as previously noted[8] (Fig. 1a). *Itln1*, *Itln2*, *Itln5* and *Itln6* are annotated as full-length proteins of 313 amino acids. Protein products from *Itln3* and *Itln4* genes are less clear, and indeed *Itln3* is annotated with an early truncation mutation predicting a shortened peptide of 142 amino acids. Nevertheless, both *Itln3* and *Itln4* genes exhibit very high sequence identity to the other mouse intelectin genes, and therefore have arisen by recent duplication.

In primates, an independent duplication event has generated two intelectin genes, *ITLN1* and *ITLN2*, encoding proteins of 313 and 325 amino acids, respectively. This event may have occurred early in primate evolution, with the tree suggesting a duplication before the divergence of New World monkeys and Old World monkeys. In the lemur lineage, an independent duplication of *ITLN1* has generated another intelectin gene, named *ITLN2,* though very distinct from other primate *ITLN2* sequences. These results suggest there is not a simple one-to-one relationship between human and mouse intelectin genes. In addition to length, inspection of the N-terminal protein sequence alignment (Fig. 1b) confirms that ITLN2-like proteins are distinguished from ITLN1-like proteins by this distinct sequence of amino acids at the N-terminus of the mature peptide, where all mouse intelectins are ITLN1-like.

We addressed the possibility that a functional homologue—an intelectin that is not necessarily orthologous but fulfils the same function in human and mouse—exists for both human genes. To explore this, we noted the constitutive expression pattern of *ITLN1* and *ITLN2* in humans and examined the pattern of expression of the intelectin genes in the intestine of the 129S2/SvPasCrl strain, which, like 129S7/Sv, carries the full complement of six intelectin genes (Supplementary table 2). Inspection of the Gtex, FANTOM5 and Human Protein atlas expression data show that human *ITLN1* and *ITLN2* have distinct expression patterns (Supplementary Fig. 1). Across the datasets, for the intestine, *ITLN1* is expressed in both small intestine and colon, while *ITLN2* expression appears narrower.

In the mouse, using RT-qPCR, we confirmed that *Itln1* in the C57BL/6NCrl sub-strain is primarily expressed in the ileum, and, to a lesser extent, in the colon (Fig. 2a). In 129S2/SvPasCrl, which has the full complement of intelectin genes, simple RT-qPCR is useful to detect tissue expression patterns (Fig. 2b) but is unable to unambiguously distinguish the transcripts from the individual intelectin genes due to the high sequence similarity across the intelectin genes. Therefore, we used a complementary strategy using high-throughput sequencing (Supplementary Fig. 2) to determine which orthologues were expressed in the small intestine and colon of the 129S2/SvPasCrl mice. In the small intestine we observed *Itln1* expression consistent with previous reports[4,6], but also detected trace levels of *Itln2* and *Itln6* transcripts (0.22% and 0.15% of total intelectin transcripts, respectively) (Supplementary Table 1). In the colon we observed *Itln6* expression (99.94%) and trace levels of *Itln2* (0.06%); however, neither *Itln1* nor any other orthologue was detected (Supplementary Table 1). Despite genetic differences in encoded intelectin genes between C57 and 129 strains, the highest mRNA expression levels in mice appear to be in the small intestine, whereas extra-intestinal baseline expression is at much lower levels (Fig. 2b)—a contrast to human *ITLN1* that is highly expressed in multiple tissue types, including visceral adipose. Higher expression levels in mouse esophagus, stomach, ovary, and uterus (Fig. 2b) may be due to other intelectin genes being expressed in those tissues.

Looking across other mouse strains, we confirmed a higher level of intelectin expression in ileum compared to colon (Fig. 3). Using the next generation sequencing strategy described above, we analysed the intelectin transcripts in wild-derived strains PANCEVO/EiJ and SKIVE/EiJ and confirmed that, as in 129S2/SvPasCrl, only *Itln6*, and not *Itln1*, was expressed in the colon, with the small intestine expressing mostly *Itln1*; although in these

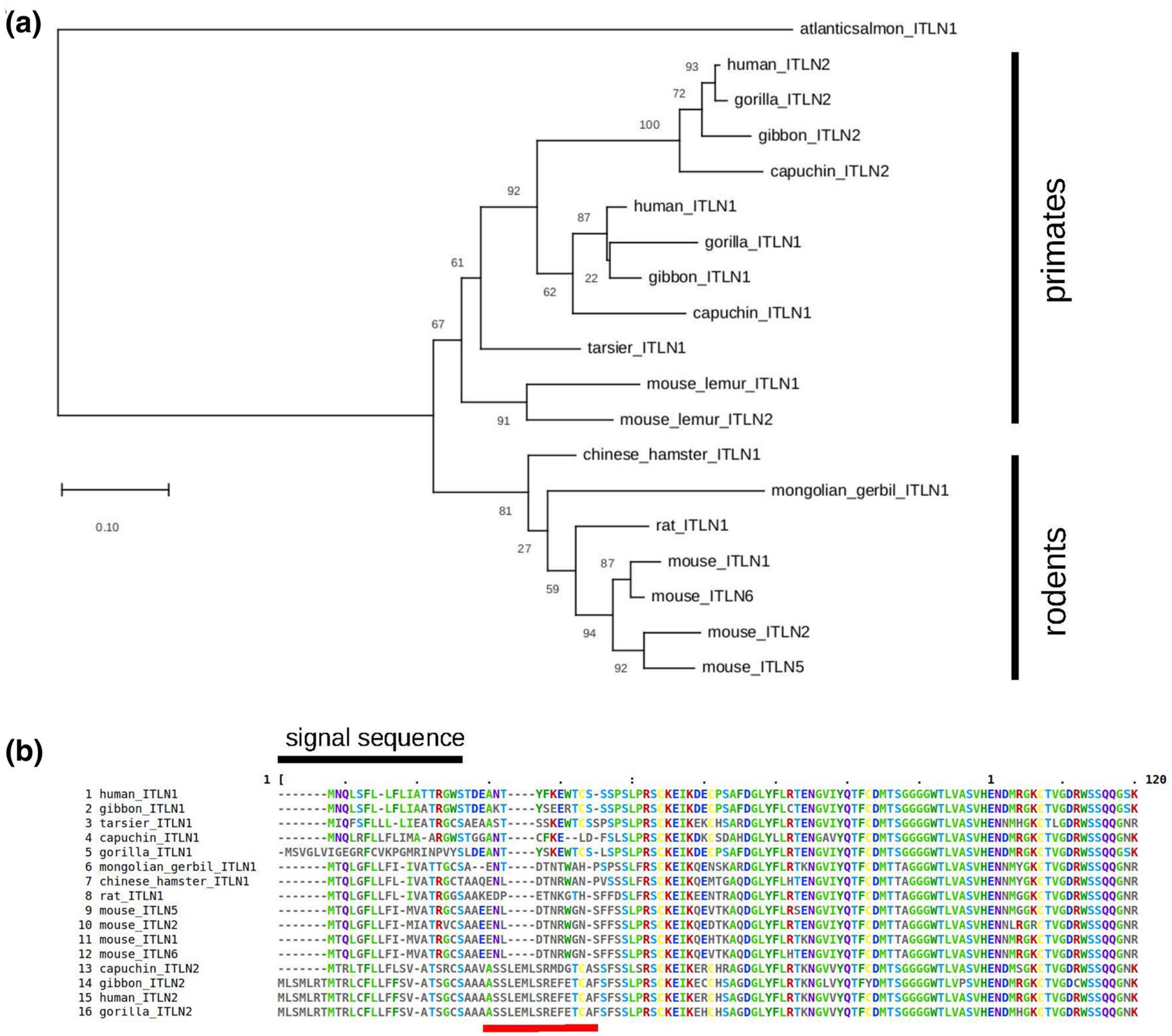
(a)
atlanticsalmon_ITLN1
human_ITLN2
gorilla_ITLN2
gibbon_ITLN2
capuchin_ITLN2
human_ITLN1
gorilla_ITLN1
gibbon_ITLN1
capuchin_ITLN1
tarsier_ITLN1
mouse_lemur_ITLN1
mouse_lemur_ITLN2
chinese_hamster_ITLN1
mongolian_gerbil_ITLN1
rat_ITLN1
mouse_ITLN1
mouse_ITLN6
mouse_ITLN2
mouse_ITLN5
93
72
100
92
87
61
22
62
67
91
81
27
87
59
94
92
0.10
primates
rodents
(b)
signal sequence
1 [. . . . : . . . . 1 . . 120
1 human_ITLN1 -------MNQLSFL-LFLIATTRGWSTDEANT----YFKEWTCS-SSPSLPRSCKEIKDECPSAFDGLYFLRTENGVIYQTFCDMTSGGGGWTLVASVHENDMRGKCTVGDRWSSQQGSK
2 gibbon_ITLN1 -------MNQLSFL-LFLIAATRGWSTDEAKT----YSEERTCS-SSPSLPRSCKEIKDECPSAFDGLYFLCTENGVIYQTFCDMTSGGGGWTLVASVHENDMRGKCTVGDRWSSQQGSK
3 tarsier_ITLN1 -------MIQFSFLLL-LIEATRGCSAEAAST----SSKEWTCSSPSPSLPRSCKEIKEKCHSARDGLYFLRTENGVIYQTFCDMTSGGGGWTLVASVHENNMHGKCTLGDRWSSQQGNR
4 capuchin_ITLN1 -------MNQLRFLLFLIMA-ARGWSTGGANT----CFKE--LD-FSLSLPRSCKEIKDKCSDAHDGLYLLRTENGAVYQTFCDMTSGGGGWTLVASVHENDMRGKCTVGDRWSSQQGNK
5 gorilla_ITLN1 -MSVGLVIGEGRFCVKPGMRINPVYSLDEANT----YSKEWTCS-LSPSLPRSCKEIKDECPSAFDGLYFLRTENGVIYQTFCDMTSGGGGWTLVASVHENDMRGKCTVGDRWSSQQGSK
6 mongolian_gerbil_ITLN1 -------MTQLGFLLFI-IVATTGCSA--ENT----DTNTWAH-PSPSSLFRSCKEIKQENSKARDGLYFLRTKNGVIYQTFCDMTTAGGGWTLVASVHENNMYGKCTVGDRWSSQQGNR
7 chinese_hamster_ITLN1 -------MTQLGFLLFL-IVATRGCTAAQENL----DTNRWAN-PVSSSLFRSCKEIKQEMTGAQDGLYFLHTENGVIYQTFCDMTTAGGGWTLVASVHENNMYGKCTVGDRWSSQQGNR
8 rat_ITLN1 -------MTQLGFLLFL-IVATRGGSAAKEDP----ETNKGTH-SFFDSLFRSCKEIKEENTRAQDGLYFLRTENGVIYQTFCDMTTAGGGWTLVASVHENNMGGKCTVGDRWSSQQGNR
9 mouse_ITLN5 -------MTQLGFLLFI-MVATRGCSAAEENL----DTNRWGN-SFFSSLPRSCKEIKQEVTKAQDGLYFLRSENGVIYQTFCDMTTAGGGWTLVASVHENNMGGKCTVGDRWSSQQGNR
10 mouse_ITLN2 -------MTQLGFLLFI-MIATRVCSAAEENL----DTNRWGN-SFFSSLPRSCKEIKQEDTKAQDGLYFLRTENGVIYQTFCDMTTAGGGWTLVASVHENNLRGRCTVGDRWSSQQGNR
11 mouse_ITLN1 -------MTQLGFLLFI-MVATRGCSAAEENL----DTNRWGN-SFFSSLPRSCKEIKQEHTKAQDGLYFLRTKNGVIYQTFCDMTTAGGGWTLVASVHENNMRGKCTVGDRWSSQQGNR
12 mouse_ITLN6 -------MTQLGFLLFI-MVATRGCSAAEENL----DTNRWGN-SFFSSLPRSCKEIKQEVTKAQDGLYFLHTENGVIYQTFCDMTTAGGGWTLVASVHENNMRGKCTVGDRWSSQQGNR
13 capuchin_ITLN2 -------MTRLTFLLFLSV-ATSRCSAAVASSLEMLSRMDGTCASSFSSLSRSCKEIKERCHRAGDGLYFLRTKNGVVYQTFCDMTSGGGGWTLVASVHENDMSGKCTVGDRWSSQQGNK
14 gibbon_ITLN2 MLSMLRTMTRLCFLLFFSV-ATSGCSAAAASSLEMLSREFETCAFSFSSLPRSCKEIKECCHSAGDGLYFLRTKNGLVYQTFYDMTSGGGGWTLVPSVHENDMRGKCTVGDCWSSQQGNK
15 human_ITLN2 MLSMLRTMTRLCFLLFFSV-ATSGCSAAAASSLEMLSREFETCAFSFSSLPRSCKEIKERCHSAGDGLYFLRTKNGVVYQTFCDMTSGGGGWTLVASVHENDMRGKCTVGDRWSSQQGNK
16 gorilla_ITLN2 MLSMLRTMTRLCFLLFFSV-ATSGCSAAAASSLEMLSREFETCAFSFSSLPRSCKEIKEHCHSAGDGLYFLRTKNGVVYQTFCDMTSGGGGWTLVASVHENDMHGKCTVGDRWSSQQGNK

Figure 1. Intelectin genes in rodents and primates. **(a)** A phylogenetic tree based on amino acid sequences is drawn to scale, with branch lengths measured in the number of substitutions per site. Numbers at each node indicate bootstrap support for that node. **(b)** Amino acid sequence alignment of the N-terminal region of intelectin.

cases about 20% of the intelectin transcript was *Itln6*. This strongly suggests that after duplication *Itln6* and *Itln1* evolved to become more tissue specific, and that the subsequent deletion in C57 strains removed this tissue specificity and resulted in *Itln1* being expressed in both tissues again—perhaps by removing key *cis*-acting elements.

Mouse intelectin genes show extensive copy number variation. To detect whether all sub-strains of C57 only had *Itln1*, we designed several paralogue ratio tests (PRTs) to test for the presence of *Itln1*, *Itln2*, *Itln4* and *Itln6* genes in mouse genomic DNA (Fig. 4). Analysis of twelve C57BL/6 substrains showed that these, together with the C57 progenitor strains (i.e., C57L/J, C57BR/cdJ, and C58/J), only have *Itln1* but not *Itln2*, *Itln4* or *Itln6*, suggesting they share the same deletion (Supplementary table 2). Deletion-specific PCR primers were designed across the deletion breakpoint in C57BL/6J, which yielded PCR amplification products of identical size in the strains with the deletion. Sanger sequencing confirmed that the breakpoint was identical in these strains (chr1:173447174–173447330), within a SINE, and therefore identical by descent. Other laboratory mouse strains derived from *Mus musculus musculus* and *Mus musculus domesticus* showed the full complement of six intelectin genes (Supplementary table 2). Using optical mapping, we confirmed the presence of the full region in PWD/PhJ, an inbred mouse strain of the subspecies *M. m. musculus*, and expected deletion in C57BL/6J (Fig. 5).

It was previously suggested that the genome of the CAST/EiJ strain has a different deletion from that observed in C57BL/6J[8], although neither deletion was characterised. Since CAST/EiJ is a strain derived from wild-caught *Mus musculus castaneus*, we tested the hypothesis that in wild-derived mouse strains other copy number variants

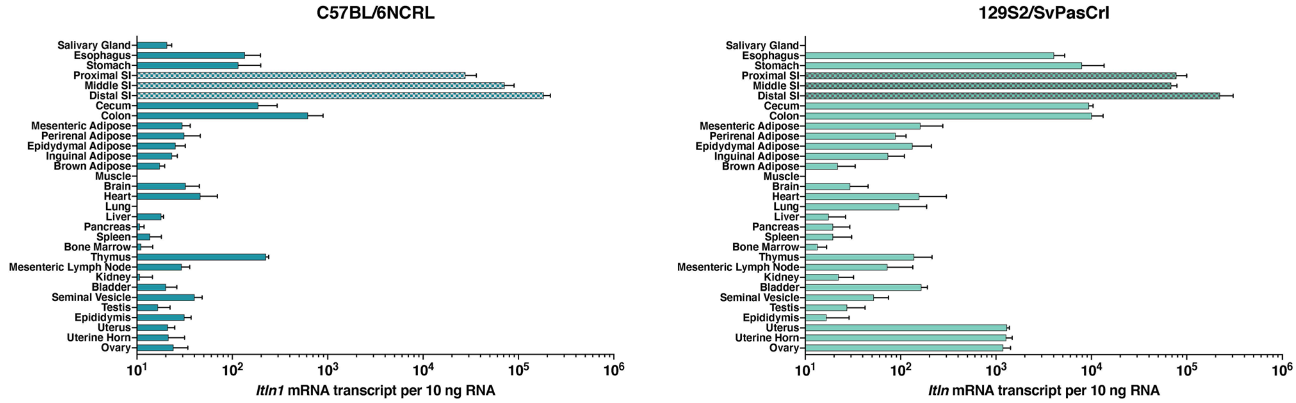

Figure 2. Analysis of intelectin expression in mice. mRNA expression levels measured by RT-qPCR are shown across tissues for **(a)** *Itln1* in C57BL/6NCRL and **(b)** all intelectins in 129S2/SvPasCRL. Absolute quantification of mouse *Itln* mRNA transcript counts from tissue samples (two technical duplicates for each of four mice) determined from standard curves using a sequence specific plasmid and presented as transcripts per 10 ng total RNA. Note that qPCR primers were designed to amplify all six intelectin transcripts from 129S2/SvPasCRL with equal efficiency. Error bars represent standard error of the mean of results from four mice.

could exist. To do so, we made use of previously sequenced mouse strains, including wild-derived mouse strains, to determine CNV across the intelectin region. We analysed 15 laboratory mouse strains that had previously been whole-genome sequenced using Illumina short-read sequencing technology[20,21]. The alignment and variant analysis of these strains used the reference C57BL/6NJ genome, and subsequent genome assemblies of a selection of these strains also only assembled *Itln1*. Therefore, we remapped raw paired-end sequence reads to the previously assembled 603 kb contig spanning the intelectin region[8]. For each sequenced mouse strain, we plotted normalised sequence read depth in 5 kb windows across the assembly. The normalised sequence read depth was not uniform across the region because of variation in the density of repeat-masked sequence, and therefore the number of mapped reads varies across different 5 kb windows. However, the observed variation was highly reproducible across different strains known to have the full complement of intelectins. Compared to 129S7/Sv the sequence from C57BL/6NJ showed a drop in read depth corresponding to the known loss of intelectin genes *Itln2*, *Itln3*, *Itln4*, *Itln5*, and *Itln6*.

The CAST/EiJ strain mouse was found to have a similar size deletion as C57BL/6NJ (Fig. 6), which was confirmed using optical mapping (Fig. 5). Fine mapping using our paralogue ratio tests and paralogue-specific PCR showed that the CAST/EiJ deletion resulted in loss of *Itln2* and *Itln5* but not *Itln6*, delineating it from the deletion event found in C57BL/6NJ (Supplementary table 2). The PANCEVO/EiJ strain, derived from *Mus hortulanus*, also showed a deletion involving the same genes as CAST/EiJ. In the SPRET/EiJ mouse strain, derived from *Mus spretus*, a smaller contiguous deletion was found, which deleted *Itln2*, *Itln3,* and also *Itln5* but not *Itln1*, *Itln6* or *Itln4* (Fig. 6). A similar deletion was found in SKIVE/EiJ (mosaic of *M. m. Musculus* and *M. m. domesticus*) and MOLF/EiJ (*M. m. mollosinus*) strains (Supplementary table 2), suggesting that the deletion occurs across subspecies of *M. musculus*. Several attempts were made to design specific PCR amplifications to precisely determine deletion breakpoints; however, these failed due to the highly repetitive nature of the sequence around the breakpoints.

We extended our analysis from laboratory inbred strains to wild-caught outbred mice. We analysed publicly available sequence data from 29 *Mus musculus domesticus*, *Mus spretus* and *Mus musculus castaneus* mice[22]. A small deletion was discovered in a wild *M. musculus castaneus* mouse from the Himalayas in India (sample H28), and a wild *M. musculus domesticus* mouse from Germany (sample TP51D). This deletion removes both *Itln2* and *Itln3* and is distinct from the deletions observed in the laboratory strains, suggesting that this deletion is polymorphic across wild mouse subspecies populations. This deletion was confirmed by PRT analysis.

Mouse intelectin genes *Itln3* and *Itln5* are polymorphic pseudogenes.

We analysed our sequencing read alignments from the 15 laboratory mouse strains and 29 wild-caught mice for novel single nucleotide variants in the coding regions of the intelectin genes. Due to the sequence similarity between intelectin genes in this family, care was taken to minimize the risk of incorrectly identifying apparent single nucleotide variants that were in fact differences due to mismapping of sequence reads from similar, paralogous sequences. Three factors suggested that this was not a problem in our analyses. Firstly, the ability to readily visualise deletions by sequence read depth analysis suggests that extensive mismapping from non-deleted to deleted regions did not occur. Secondly, alignment of sequence from 129S1/SvlmJ strain against the reference derived from 129S7/Sv gave no variants across the intelectin region. Finally, mapping C57BL/6NJ sequence reads to the contig containing all the intelectin genes resulted in only a single, low quality variant call across the region deleted in C57BL/6NJ. No variants were annotated in the deleted regions of CAST/EiJ or SPRET/EiJ. Taken together, this suggests a very low or non-existent level of sequence mismapping. Nevertheless, all variant calls were visually inspected at the sequence alignment level to check for reads with multiple mismapping sites and for a biologically appropriate ratio of alternative allele counts (i.e., ~ 0.5 for heterozygotes, and ~ 1 for homozygotes).

No variants in the coding regions of the intelectin genes were observed in 129S1/SvlmJ, BALB/cJ or C3H/HeJ, suggesting a recent shared origin of this region across these three strains. For other strains that have a full

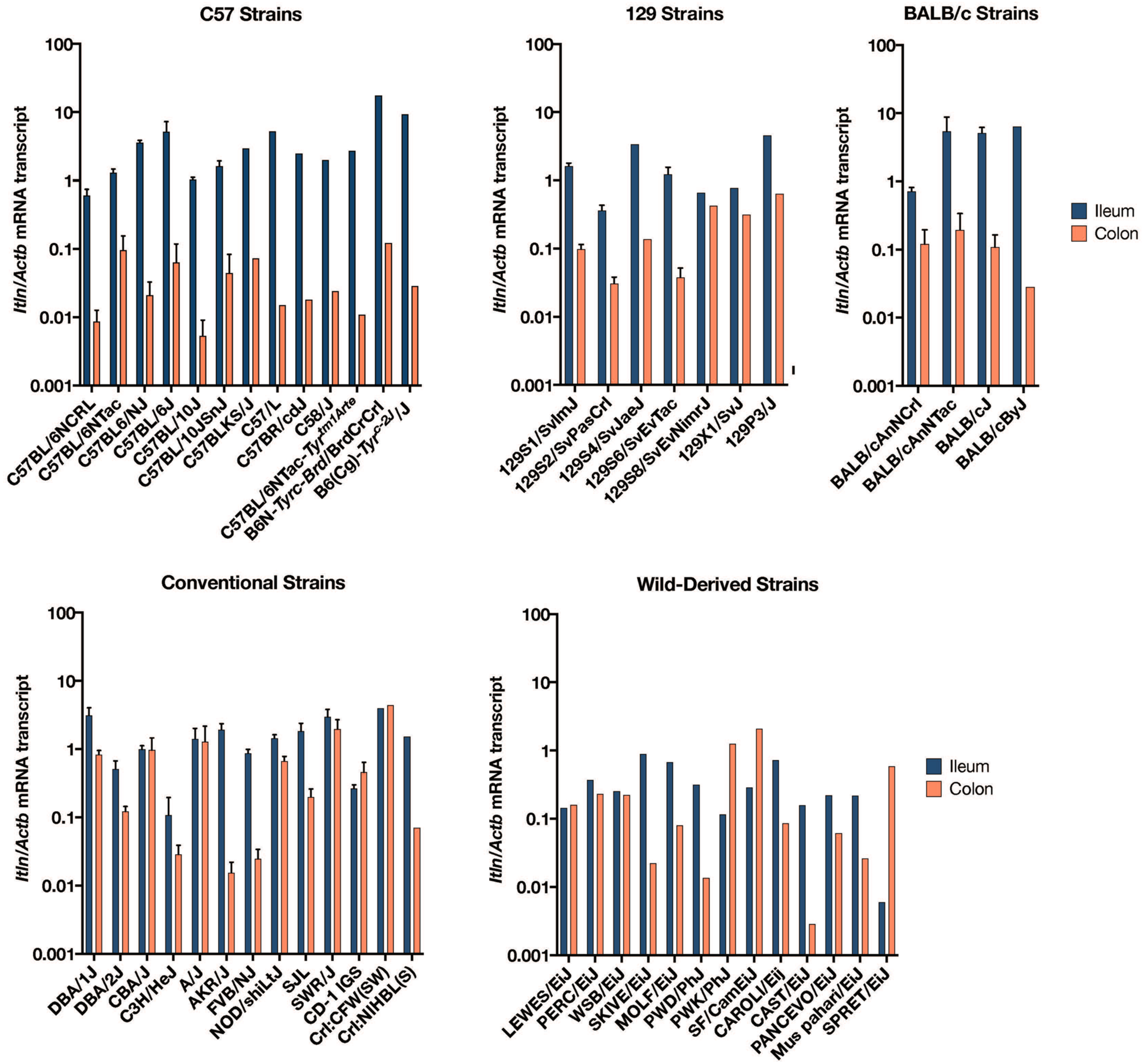

Figure 3. Analysis of intelectin expression in ileum and colon across different mouse strains. Quantification of mouse *Itln* mRNA levels in small intestinal ileum and colon tissue of mouse strains determined by RT-qPCR (two technical duplicates for each of four mice). Note that qPCR primers correspond to targets of identical sequence in *Itln1* in C57BL/6NCRL and all six intelectin mRNA in 129S2/SvPasCRL. Data are normalised to *Actb* transcript levels. Error bars represent standard error of the mean (n = 4 mice). Because of high cost of some strains, specimens from a single mouse were analysed and data presented without error bars.

complement of intelectin genes, between 16 and 31 variants per strain were observed across the intelectin genes. For strains with a deletion, CAST/EiJ had 12 variants and SPRET/EiJ had 40 variants. We focused on single nucleotide variants that were predicted to introduce or removed stop codons from each gene (Table 1), as we would expect these to have a major effect on biological function. A variant in *Itln5* exon 5 changed a cysteine to a stop codon in DBA/2J, CBA/J, A/J, PWK/PhJ and NZO/HlLtJ, and this premature termination codon is likely to suppress translation of the resulting mRNA by nonsense mediated decay[23]. Conversely, in exon 4 of *Itln3*, a stop codon present in the reference sequence is modified to CAG (Gln) in FVB and WSB strains, suggesting that this gene does have potential for expression as a full-length intelectin.

As expected for individuals from outbred populations, wild mice showed much more single nucleotide variation (i.e., 14 to 72 variants per individual). Wild mice populations are polymorphic for both the *Itln5* stop allele and the *Itln3* stop allele seen in certain laboratory strains, suggesting that some laboratory strains inherited these alleles. Similarly, an early truncation allele occurring at glutamine 103 in *Itln5* was found in the *M. m. domesticus* population sampled from Cologne in Germany.

As intelectins are calcium-dependent lectins, we also examined variation affecting either the known calcium- or carbohydrate-binding amino acid residues between and across the intelectin family[2,13]. An alignment of the

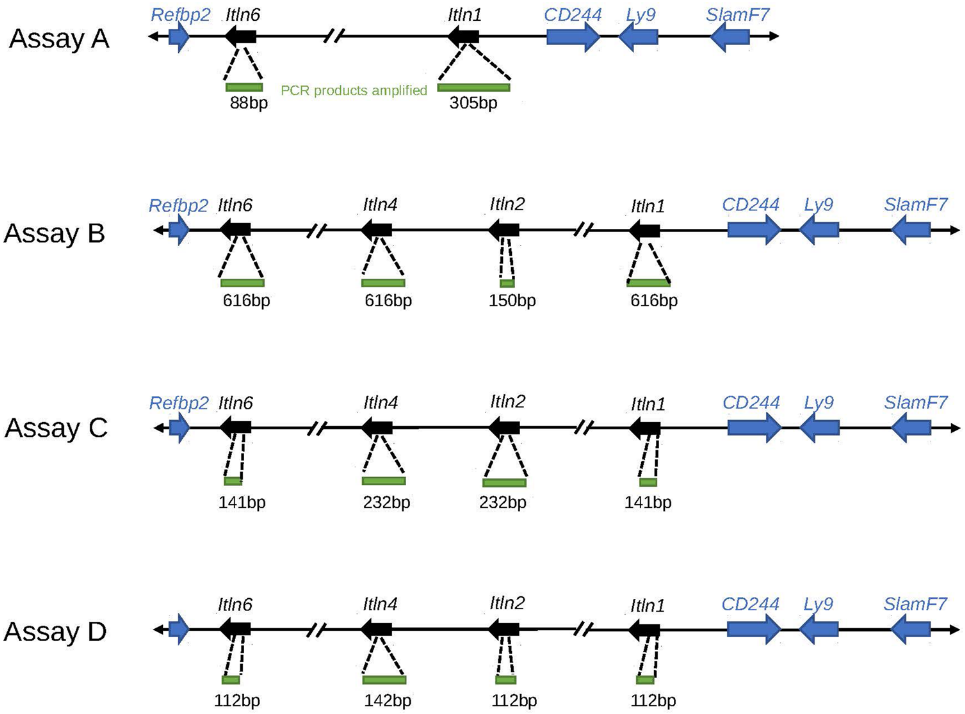

Figure 4. Paralog ratio tests (PRTs) to measure relative copy number of mouse intelectin genes. Four assays **(A–D)** designed to amplify different sized PCR products (green) using a single primer pair in each assay. Predicted amplicon lengths from the contig of the region previously generated from the 129S7 mouse strain.

mouse and human intelectin genes showed that all seven of the predicted carbohydrate-binding amino acids (residues 243, 244, 260, 262, 263, 274, 288 and 297) were identical between human ITLN1 and mouse Itln1, consistent with glycan-binding biochemical studies for these two proteins[2]. However, for other intelectins at least one of these amino acids varies (Fig. 7, highlighted in red). For *Itln6*, the carbohydrate-binding tyrosine at position 297 is polymorphic with an alternative non-polar phenylalanine allele (Table 1), potentially altering carbohydrate-binding. Two amino acid variants in *Itln1* and *Itln4* (i.e., T261S and H263A) may directly (i.e., ligand binding) or indirectly (i.e., configuration of binding pocket) modify glycan binding. In contrast, all ten of the calcium-binding residues are perfectly conserved (Fig. 7, highlighted in blue).

Mouse Itln1 shows evidence of recent adaptive evolution. Given that mouse intelectin genes clearly have a recent evolutionary history of repeated rounds of duplication followed by deletion and point mutations, we wanted to examine evidence for natural selection at the genes at different times following divergence of the lineages leading to humans and mice. To do this, we used the variation data generated for each gene in wild mice and applied the McDonald-Kreitman test to compare ratio of nonsynonymous to synonymous codon variation within mice to this ratio of variation between mice and other species (Table 2). In this case, we took each mouse intelectin coding sequence in turn, and determined the polymorphism within orthologues of that gene in the various mouse strains. We then compared this variation to the divergence observed with mouse paralogues, the single rat orthologue, and a human orthologue (*ITLN1*), with the aim of detecting selection at recent timescales. There is evidence of selection of mouse *Itln1* following the burst of duplication that generated multiple intelectins (Table 2). Because of the likelihood of gene conversion events homogenising sequence between recent duplications, a date for the duplication event cannot be reliably estimated from sequence divergence. However, since the rat has only one intelectin gene we assume that these duplication events occurred after rat-mouse divergence between 9-14MYa[24].

Discussion

Intelectins are a family of calcium-dependent lectins that span vertebrate evolution. Human intelectin-1 is implicated in several disease states, including inflammatory bowel disease, asthma, and obesity; however, its specific functions remain unclear. Mouse models are powerful tools to elucidate the biological function of specific proteins in vivo; however, such studies for intelectins are complicated because of the striking, and as yet incompletely characterised strain-specific variation of the intelectin locus resulting in uncertainty regarding which mouse intelectin gene(s) might represent the corresponding orthologue in humans. Herein, we provide evidence that human *ITLN1* has multiple orthologues in mice, whereas human *ITLN2* does not have a corresponding mouse orthologue. In addition, our data reveal key differences within the mouse intelectin gene family, even between common laboratory strains. Together, our findings provide insight into the evolution of mouse intelectins and

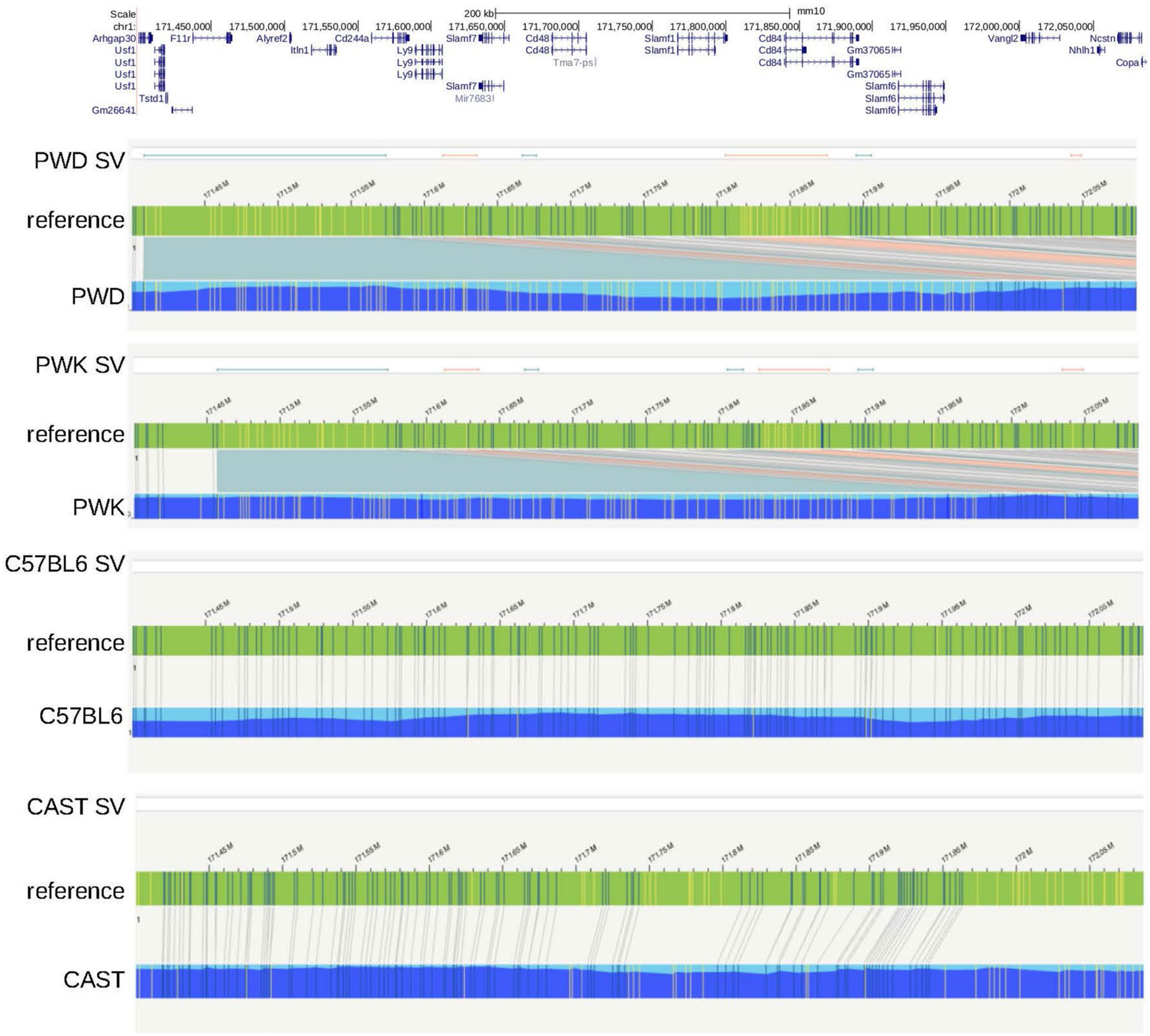

Figure 5. Optical mapping of the intelectin region in laboratory strain mice. Optical mapping results from four strains (PWD, PWK, C57BL/6NCrl, CAST/EiJ) have been aligned against the reference genome (green bar, reference assembly, blue bar new strain assembly with depth of coverage indicated by the dark blue). The yellow and blue vertical lines in these bars represent NtBspQI restriction enzyme cutting sites or DLE-1 binding sites (CTTAAG motif). A blue vertical line indicates sites that are matching between the new assembly and the reference genome, yellow indicates that the site is there in the reference or the new assembly, but are considered non-matching. Regions showing structural variation, with respect to the C57BL6/J reference genome, are annotated as SV, green lines showing putative insertions and red lines showing putative deletions. An extra ~ 400 kb of sequence is in both PWK and PWD strains in the *Itln1* region, consistent with the presence of the full complement of intelectin genes. Gene annotations are at the top and are taken from the UCSC Genome Browser.

highlight the need to carefully consider mouse strain when designing and interpreting experiments whose outcome might be dependent upon intelectin biology and/or report intelectin-specific results.

Phylogenetic tree analysis shows that mice have six intelectin genes resulting from a recent burst of duplications, unlike rats and other murids which have a single intelectin gene. In humans and other primates, an independent duplication event occurring early in (or preceding) primate evolution generated two intelectin genes, *ITLN1* and *ITLN2*. We provide evidence that based on overall protein length and N-terminal sequence of the mature peptide, ITLN2-like proteins are distinguished from ITLN1-like proteins. Our phylogenetic analysis indicates that there is no simple one-to-one relationship between human and mouse intelectin genes. Instead, our data support that all six mouse intelectins are ITLN1-like.

We extensively characterised strain-specific variation of mouse intelectin genes, including analysis of laboratory strains and wild mice. A previous report demonstrated that the 129S7/Sv laboratory mouse strain has six

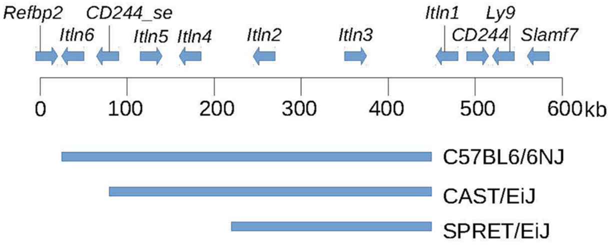

Figure 6. Deletions found by sequence analysis of laboratory strain mice. The region covered by the original 129S7/Sv contig is shown with genes annotated. Extent of deletions found by sequence read depth analysis in laboratory strain mice shown below the size scale.

Location	Amino acid position	Gene	Ref codon	Alt codon	Population	Allele frequency
28777	Y297F	*Itln6*	TAC Y	TTC F	Germany (Heligoland)	0.3
"	"	"	"	"	Germany (Cologne-Bonn)	0.81
"	"	"	"	"	France (Massif Central)	0.68
"	"	"	"	"	Iran (Ahvaz)	0.5
"	"	"	"	"	CAST/EiJ, WSB/EiJ, FVB/NJ, NZO/HlLtJ, NOD/ShiLtJ, NZB/B1NJ, PWK/PhJ, and A/J	(Inbred strains)
117243	Q103*	*Itln5*	CAG Q	TAG *	Germany (Cologne-Bonn)	0.25
118425	C235*	*Itln5*	TGT C	TGA *	France (Massif Central)	0.25
"	"	"	"	"	DBA/2J, CBA/J, A/J, PWK/PhJ, NZO/HlLtJ	(Inbred strains)
174162	T261S	*Itln4*	ACT T	TCT S	NZB/B1NJ	(Inbred strains)
"	"	"	"	"	Germany (Cologne-Bonn)	0.25
"	"	"	"	"	France (Massif Central)	0.125
"	"	"	"	"	Iran (Ahvaz)	0.375
365046	*143Q	*Itln3*	TAG *	CAG Q	Germany (Cologne-Bonn)	0.312
"	"	"	"	"	Iran (Ahvaz)	0.5
"	"	"	"	"	FVB, WSB	(Inbred strains)
463452	T261S	*Itln1*	ACT T	TCT S	SPRET/EiJ	(Inbred strains)
"	"	"	"	"	Germany (Cologne-Bonn)	0.18
"	"	"	"	"	France (Massif Central)	0.18
"	"	"	"	"	Iran (Ahvaz)	0.125
463445	H263A	*Itln1*	CAT H	GCT A	Germany (Cologne-Bonn)	0.18
"	"	"	"	"	France (Massif Central)	0.18
"	"	"	"	"	Iran (Ahvaz)	0.125
"	"	"	"	"	FVB, WSB	(Inbred strains)

Table 1. Putative functional amino acid variants identified in intelectin genes (all lab mice strains shown were homozygous for the alternative variant).

intelectin genes, whereas the C57BL/6J strain has only one due to a 420 kb deletion[5]. Using a newly developed collection of paralogue ratio tests we determined that 15 common laboratory mouse strains have the full complement of six intelectin genes. In contrast, all tested sub-strains of the C57 line have a single family member, *Itln1*. Using PCR primers designed to amplify across the deletion breakpoint in C57BL/6J, sequence data from the PCR products confirmed that the breakpoint was identical in these sub-strains, including C57 parent lines: C57L/J, C57BR/cdJ, and C58/J. Optical mapping supports these observations. These results imply that the deletion event occurred prior to human/researcher-driven mouse inbreeding. A similar analysis indicated that a deletion in CAST/EiJ and PANCEVO/EiJ mice removes *Itln2* and *Itln5,* and a deletion in SPRET/EiJ, SKIVE/EiJ, and MOLF/EiJ strains removes *Itln2*, *Itln3* and *Itln5,* indicating that these deletions are not identical to the C57 line deletion.

We detected no single nucleotide variants in the coding regions of the intelectin genes the genome sequences of 129S1/SvlmJ, BALB/cJ or C3H/HeJ mice, suggesting a recent shared origin of this region across these three strains. However, for the other laboratory strains, including those that have the full complement of six intelectin genes, we observed between 16 and 31 coding-region variants per strain, similar to inbred CAST/EiJ (12 variants) and SPRET/EiJ (40 variants), which possess locus deletions. As expected, we observed a higher frequency of single nucleotide variants in wild-caught mice (between 14 to 72 variants per individual). Of the single nucleotide variants predicted to have profound consequences, a premature stop codon (C235*) in exon 5 of *Itln5* was observed in five lab strains (DBA/2 J, CBA/J, A/J, PWK/PhJ, and NZO/HlLtJ) and a wild-caught mouse, and a

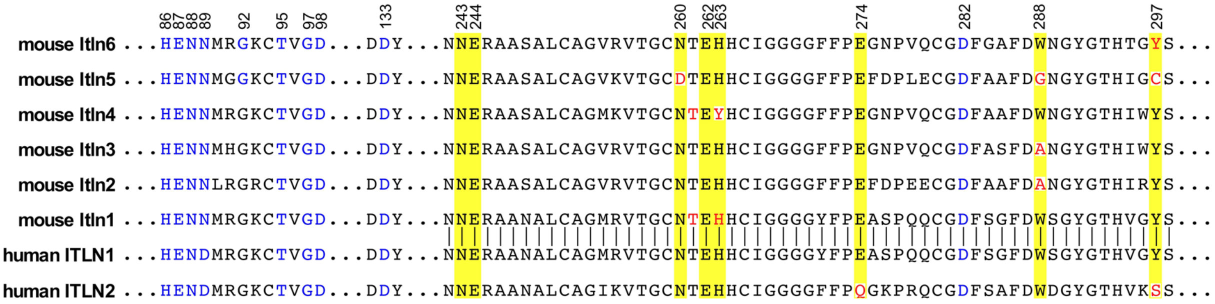
86 87 88 89 92 95 97 98 133 243 244 260 262 263 274 282 288 297
mouse Itln6 ...HENNMRGKCTVGD...DDY...NNERAASALCAGVRVTGCNTEHHCIGGGGFFPEGNPVQCGDFGAFDWNGYGTHTGYS...
mouse Itln5 ...HENNMGGKCTVGD...DDY...NNERAASALCAGVKVTGCDTEHHCIGGGGFFPEFDPLECGDFAAFDGNGYGTHIGCS...
mouse Itln4 ...HENNMRGKCTVGD...DDY...NNERAASALCAGMKVTGCNTEYHCIGGGGFFPEGNPVQCGDFAAFDWNGYGTHIWYS...
mouse Itln3 ...HENNMHGKCTVGD...DDY...NNERAASALCAGVRVTGCNTEHHCIGGGGFFPEGNPVQCGDFASFDANGYGTHIWYS...
mouse Itln2 ...HENNLRGRCTVGD...DDY...NNERAASALCAGVRVTGCNTEHHCIGGGGFFPEFDPEECGDFAAFDANGYGTHIRYS...
mouse Itln1 ...HENNMRGKCTVGD...DDY...NNERAANALCAGMRVTGCNTEHHCIGGGGYFPEASPQQCGDFSGFDWSGYGTHVGYS...
human ITLN1 ...HENDMRGKCTVGD...DDY...NNERAANALCAGMRVTGCNTEHHCIGGGGYFPEASPQQCGDFSGFDWSGYGTHVGYS...
human ITLN2 ...HENDMRGKCTVGD...DDY...NNERAANALCAGIKVTGCNTEHHCIGGGGFFPQGKPRQCGDFSAFDWDGYGTHVKSS...

Figure 7. Comparison of the calcium- and glycan-binding regions of mouse and human intelectins. The aligned partial amino acid sequences of human and mice intelectins are shown (amino acid residue numbering from human ITLN1). Amino acids previously shown[2] to mediate calcium coordination are shown in blue and those mediating glycan binding highlighted with yellow. Residues identical to those previously shown to bind carbohydrate in human ITLN1 are shown in black within the yellow background and the polymorphic amino acids are shown in red.

	Itln1	*Itln2*	*Itln5*	*Itln6*
Human *ITLN1*	0.598	0.014	0.159	0.537
Rat *Itln1*	0.064	0.200	0.981	0.490
Mouse *Itln1*	–	0.194	0.893	0.688
Mouse *Itln2*	**0.001**	–	0.901	0.091
Mouse *Itln5*	**0.007**	0.154	–	0.096
Mouse *Itln6*	0.025	0.409	0.861	–

Table 2. McDonald-Kreitman analysis of mouse *Itln* genes. P values are shown indicating the comparison between the ratio of nonsynonymous to synonymous codon variations within the four *Itln* genes (columns) against divergence across species and with paralogues (rows). For example, the top left p value of 0.598 represents the comparison of variation within mouse *Itln1* compared to divergence between mouse *Itln1* and human *ITLN1*. Statistically significant values with a false discovery rate of < 10% are shown in bold.

glutamine codon replaced the premature stop codon in exon 4 of *Itln3* (*143Q) of the reference genome, in two lab strains (FVB and WSB) and two wild-caught mice.

Complementary evidence supports the concept that *Itln1* is the functional homologue of human *ITLN1*[5]. We highlight here that both genes encode a preproprotein of identical length (313 aa), and that the residues responsible for calcium coordination (10 residues) and those mediating carbohydrate binding (8 residues) are identical in the two orthologous proteins. Prior biochemical studies confirm that glycan-recognition selectivity of Itln1 was analogous to its human counterpart; however, subtle differences in binding selectivity, kinetics, and affinity were reported[2,5]. Our data also show that mouse *Itln1* is expressed in the small intestine confirming initial reports[4].

Our findings suggest that adaptive evolution has occurred in mouse *Itln1* during the last 9–14 million years, since repeated rounds of duplication resulted in expansion of the intelectin locus. The amino acid changes in Itln1 adjacent to the carbohydrate-binding residues may have resulted in subtle alteration of affinity or spectrum of carbohydrate-binding during mouse evolution. Moreover, tissue specificity outlined here and/or potential inducibility may define different functions of individual mouse intelectins[25]. Other differences between species are that human ITLN1 exists as a disulfide-linked trimer, whereas mouse Itln1 lacks cysteines required for intermolecular disulfide-linked trimer formation[5], and that within the intestine mouse *Itln1* is expressed in Paneth cells, whereas human *ITLN1* is expressed in goblet cells. Nevertheless, a strong argument can be made for mouse *Itln1* to have essentially similar functions in the gut as human *ITLN1*.

The relationships and functions of the other mouse intelectins are less clear. As noted, our data support that all six mouse intelectins are *ITLN1*-like, with no mouse orthologue of human *ITLN2*. Mouse *Itln2*, despite its numerical designation, is not the orthologue of human *ITLN2*. Of the six encoded intelectins, only *Itln1* and *Itln6* were found to be constitutively expressed at high levels, where *Itln1* is present in the small intestine and *Itln6* is present in the colon. It is interesting that like *Itln1*, *Itln6* shares amino acid identity at the eight positions mediating glycan binding, although a conservative polymorphism at Y297F was noted in some strains. We speculate that perhaps differences in gene regulation and/or protein secretion in development and/or under environmental stress will explain the distinguishing anatomic patterns of expression.

From the data presented here, other mouse intelectin genes have variant residues at one or more key amino acids that govern carbohydrate binding, have polymorphic early truncating codons and/or have polymorphic large deletions. Combined with the lack of evidence for adaptive evolution at these other intelectin genes, this suggests a lack of critical function, and that these genes are, or are in the process of becoming, pseudogenes. Therefore, we may be witnessing a snapshot of an evolutionary process, likely common in other immunity gene families, called birth-and-death evolution[26]. Future work requires focus on identifying any novel functions of these mouse intelectin genes, including variation in topographical location, which does hint at the possibility of unique roles. If any gene does encode a fully functional protein, it is likely that polymorphic variation within wild mouse populations and between laboratory strains will have functional consequences.

This study also emphasises the importance of examining natural variation for gene regions that are absent in the reference genome of the organism under study[27]. While now being addressed in humans using long-read sequencing and efforts by consortia like the telomere-to-telomere consortium, such analyses in model or domestic animals lag behind. Moreover, our study highlights the careful considerations required by researchers investigating innate immunity genes in mice, where strain and sub-strain variations can complicate interpretations of gene function[24–30].

Methods

Mouse husbandry and tissue analysis.

The Institutional Animal Care and Use Committee at the University of California, Davis, approved all procedures involving live animals and methods of euthanasia; experiments were performed following AVMA guidelines and with strict adherence to IACUC-approved protocols. Briefly, animals were deeply anesthetised with a cocktail of ketamine and xylazine (100/mg/kg and 10 mg/kg, respectively) prior to euthanasia. Tissue samples were dissected immediately after mice were euthanised and submerged in RNAlater (Ambion Inc, Austin, TX). The RNAlater specimen tubes were incubated with gentle rocking overnight at 4 ºC, and then stored long-term at -20 ºC. This study was carried out in compliance with the ARRIVE guidelines.

DNA samples and extraction.

Genomic DNA was isolated and purified using the QIAamp DNA minikit (Qiagen, Germantown, MD) according to the manufacturer's protocol. Isolated DNA was quantified by ultraviolet absorbance spectroscopy (260 nm) using a NanoDrop spectrophotometer (Thermo Scientific/NanoDrop Products, Wilmington, DE).

Phylogenetic analysis.

Intelectin coding sequences for the phylogenetic tree were retrieved from the list of orthologues and paralogues of human *ITLN1* curated by Ensembl (https://www.ensembl.org, release 98). Mouse *Itln* sequences were identified from the full length 129 contig accession number HM370554, available from Genbank (https://www.ncbi.nlm.nih.gov/nuccore/HM370554).

Evolutionary analyses were conducted in MEGA X[31]. For the protein tree, amino acid sequences were inferred from coding DNA sequences and aligned using ClustalW (v2.1), and the phylogenetic tree inferred by maximum likelihood and a JTT matrix-based model, Atlantic salmon (*Salmo salar*) *Itln1* was used as an outgroup. The tree with the highest log likelihood is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.4683)). The rate variation model allowed for some sites to be evolutionarily invariable, but no sites were found to fit in that category of the model. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Following bootstrap analysis (500 iterations), the percentage of trees in which the associated taxa clustered together is shown next to the branches.

RNA isolation and expression analysis.

The general procedures for RNA isolation and synthesis of cDNA were previously described by our group[32,33]. Briefly, the RNAlater solution was decanted and the tissue was homogenised in guanidine thiocyanate buffer. Total RNA was isolated using cesium chloride gradient ultracentrifugation, quantified using ultraviolet absorption spectrometry at 260 nm, and reverse transcribed to cDNA using an oligo-(dT)12–18 primer. The single-stranded cDNA product was purified using the Qiagen PCR purification kit (Qiagen, Valencia, CA), and diluted to 10 ng/µl based on the input concentration of total RNA. Real-time PCR was performed using as templates the cDNA from experimental tissues and gene-specific plasmids as external standards for quantification. The primer pairs for qPCR were: *Itln1* forward 5'- ACC GCACCTTCACTGGCTTC-3', *Itln1* reverse 5'- CCAACACTTTCCTTCTCCGTATTTC-3', and *Actb* forward 5'-GGCTGTATTCCCCTCCATCG-3', reverse 5'- CCAGTTGGTAACAATGCCATGT-3'. Absolute quantification of specific mRNA from tissue was determined by extrapolation of the detection threshold (crossing point) to the crossing point for gene-specific external plasmid standard analysed within each run. Reproducibility assessments of this approach were previously reported[32]. For analysis across multiple strains, data were analysed using the delta CT method normalising gene expression to *Actb* in each sample. A negative control reaction that omitted template cDNA was included with each set of reactions to check for possible cross-contamination.

For sequence analysis of RT-PCR products, intestinal cDNA (distal small intestine and colon) from 129S2/SvPasCrl mice was used as a template in a PCR reaction using oligonucleotide primers whose sequences correspond to regions of identical sequence in mRNA of all six paralogues of intelectin (m129ItlnCom-4 s 5'-GCC TCAGCAGAGAAAGGTTCC-3' and m129ItlnCom-287a 5'GAAGGTCTGGTAGATGACACCATTC-3'). Primers were designed using MacVector Software (MacVector, Apex, NC) and synthesised by Invitrogen Life Technologies (Invitrogen, Carlsbad, CA). The PCR reactions were initiated by denaturation of the DNA template at 95 °C for 10 min followed by 45 cycles consisting of 95 °C for 15 s, a − 1 °C per two-cycle 'touchdown' annealing temperature for 5 s (i.e., 65 °C to 58 °C), and 10 s at 72 °C. The PCR product was purified by passage through a PCR clean-up column following the manufacturer's protocol (Qiagen). The purified sample was then sequenced using an Illumina platform (Genewiz, South Plainfield, NJ). Sequence data was filtered to remove reads that failed base-calling quality checks, and the reads with identical sequence were assigned to a particular intelectin gene by comparison to a reference sequence.

DNA sequence analysis. All mouse genome coordinates use the GRCm38 assembly (Genbank accession number GCA_000001635.2). The coding sequence for human *ITLN1*, human *ITLN2* and Rat *Itln1* were retrieved from Genbank (https://www.ncbi.nlm.nih.gov/nuccore) with accession numbers AB036706.1, AY065973.1 and XM_017598901 respectively. The coding sequence for each mouse *Itln* gene was retrieved from the BAC contig of the 129S7 intelectin locus from GenBank with accession number (HM370554) (https://www.ncbi.nlm.nih.gov/nuccore/HM370554).

Whole genome sequencing (WGS) data for 29 wild-caught mice were accessed at the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) under accession number PRJEB9450.These 29 wild-caught mice had been sequenced by Illumina Hiseq 2000 at the Max Planck Institute[22]. The 15 laboratory mouse strains were sequenced by the Wellcome Sanger Institute and accessed at the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/)[20] (Supplementary table 3).

The quality of the raw fastq files was assessed using FastQC v 0.11.5, followed by adapter removal using Cutadapt v 01.11[34]. Following mapping of the processed fastq files to the repeat-masked BAC contig (accession number HM370554) using BWA-MEM v 0.7.15[35], the resulting SAM file was converted to a BAM file and sorted and indexed using SAMTools v 1.3.2[36]. Multiple bam files from the same mouse sample were merged using SAMtools, and labelled as a single new unified read group by Picard v2.1[37]. The resulting bam file was subjected to local realignment using GATK 3.6 and 3.8. First, the RealignerTargetCreator command in GATK was used to identify misalignments and these were saved in 'interval' format. Next, realigning these reads to the BAC contig was carried out locally using the 'IndelRealigner' command using the information that is given in the 'interval'. PCR duplicates were identified using Picard v 2.1 and removed using SAMTools, resulting in the final single bam file for each mouse.

Paralogue ratio test. The paralogue ratio test (PRT) is a form of quantitative PCR that uses shared PCR primers to minimise differences in amplification kinetics between cases and controls[38,39]. We designed four paralogue ratio tests (PRTs, assays A-D) to examine relative copy numbers of intelectin genes in mice by designing a primer pair specific to a subset of intelectin genes, ensuring at least 5 mismatches between the primer and non-amplified intelectins. PCR primers were 5'-TATTCCTGTCTCAGCTCCTAG-3', 5'-GTCACAGGTAAAKCCAGAAGG-3' for assay A, 5'-TGTAAAYCCCTCCTTGACTCC-3', 5'-GATAAATGRCCCAGTMCTTGCC-3' for assay B, 5'-CACACGTACACCTTCCTG-3', 5'-GAGTAGTCTGCYKTATTTCAG-3' for assay C and 5'-GTAATGYTGACTTYGGCCTTC-3', 5'-GTTGAGCWTGGCAGATTGT-3' for assay D with IUPAC codes K Y, M and R indicating that the primers were synthesised with a mix of the two indicated nucleotides at that position.

PCR amplification was in a final volume of 10 µl, which included 5–10 ng genomic DNA, 0.5U *Taq* DNA polymerase and a final concentration of 1 mM primer in 1 × KAPA Buffer A (a Tris-ammonium sulphate buffer with a final concentration of 1.5 mM $MgCl_2$). Thermal cycling was one cycle of initial denaturation at 94 °C for 2 min, followed by 35 cycles of denaturation at 94 °C for 30 s, then PRT1 and PRT2 were annealed at 60 °C and PRT3 and PRT4 were annealed at 62 °C. All assays were annealed for 30 s; this was then followed by an extension step for 30 s at 72 °C. Lastly, one extra extension step was carried out at 72 °C for 5 min. PCR amplification products from different intelectin genes were detected based on amplicon size using standard 2% agarose gel electrophoresis, ethidium bromide staining and visualisation under UV light.

Calling variation from sequence alignment files. Across the whole region (603274 bp) of the *Itln* locus, the number of reads mapping to non-overlapping 5 kb windows was calculated using SAMTools (v1.3.2)[36]. In each single window, reads were counted, normalised to average read count and plotted to visualise gain or loss. Single nucleotide variation for the exonic regions for each individual mouse from the bam file was called using FreeBayes (v1.1)[40], and converted to vcf format using VCFtools (v0.1.14)[37] on a minimum quality score of 30 and a minimum depth of 15. All vcf files are available at https://doi.org/10.25392/leicester.data.13679035.v1.

Optical mapping. We generated optical maps across the whole-genome of four different mice, from two mouse subspecies. C57BL/6J (B6) and C57BL6Crl (B6N) of *Mus musculus domesticus,* CAST/EiJ of *M. musculus castaneus* and PWD/Ph (PWD) and PWK/Ph (PWK) of *M. m. musculus* origin. First megabase-scale high molecular weight DNA was extracted according to the Saphyr Bionano Prep Animal Tissue DNA Isolation Soft Tissue Protocol (Document Number: 30077; Document Revision: B). Briefly, cell nuclei were isolated from splenic tissue and embedded in agarose plugs. DNA in plugs was purified with Proteinase K and RNAse, then high molecular weight (HMW) genomic DNA was extracted from the agarose plugs using agarase and purified by drop dialysis. HMW DNA was resuspended overnight before quantification with the Qubit BR dsDNA assay, then kept at 4 °C until labelling.

We performed the Bionano Direct Labelling and Staining (DLS) protocol (Document Number: 30024 Revision: I) on 750 ng of DNA from each sample using direct labelling enzyme (DLE-1) to label all its recognition sites (CTTAAG). After an initial clean-up step, the labelled HMW DNA was pre-stained, homogenised, and quantified with the Qubit HS dsDNA assay, before using an appropriate amount of backbone stain YOYO-1. The molecules were then imaged using the Bionano Saphyr System (Bionano Genomics, San Diego). The resulting de-novo optical maps were then generated and mapped against an in-silico optical map of the mouse genome reference sequence.

Received: 31 March 2021; Accepted: 13 July 2021
Published online: 30 July 2021

References

1. Watanabe, T., Watanabe-Kominato, K., Takahashi, Y., Kojima, M. & Watanabe, R. Adipose tissue-derived omentin-1 function and regulation. *Compr. Physiol.* **7**(3), 765–781 (2011).
2. Wesener, D. A., Dugan, A. & Kiessling, L. L. Recognition of microbial glycans by soluble human lectins. *Curr. Opin. Struct. Biol.* **44**, 168–178 (2017).
3. Chen L, Li J, Yang G. A comparative review of intelectins. *Scand. J. Immunol.* e12882 (2020).
4. Komiya, T., Tanigawa, Y. & Hirohashi, S. Cloning of the novel gene intelectin, which is expressed in intestinal paneth cells in mice. *Biochem. Biophys. Res. Commun.* **251**(3), 759–762 (1998).
5. Tsuji, S. *et al.* Human intelectin is a novel soluble lectin that recognizes galactofuranose in carbohydrate chains of bacterial cell wall. *J. Biol. Chem.* **276**(26), 23456–23463 (2001).
6. Pemberton, A. D. *et al.* Innate BALB/c enteric epithelial responses to *Trichinella spiralis*: Inducible expression of a novel goblet cell lectin, intelectin-2, and its natural deletion in C57BL/10 mice. *J. Immunol.* **173**(3), 1894–1901 (2004).
7. Wrackmeyer, U., Hansen, G. H., Seya, T. & Danielsen, E. M. Intelectin: A novel lipid raft-associated protein in the enterocyte brush border. *Biochemistry* **45**(30), 9188–9197 (2006).
8. Lu, Z. H. *et al.* Strain-specific copy number variation in the intelectin locus on the 129 mouse chromosome 1. *BMC Genomics* **12**(1), 1–11 (2011).
9. Yan, J. *et al.* Comparative genomic and phylogenetic analyses of the intelectin gene family: Implications for their origin and evolution. *Dev. Comp. Immunol.* **41**(2), 189–199 (2013).
10. Lönnerdal, B., Jiang, R. & Du, X. Bovine lactoferrin can be taken up by the human intestinal lactoferrin receptor and exert bioactivities. *J. Pediatr. Gastroenterol. Nutr.* **53**(6), 606–614 (2011).
11. Akiyama, Y. *et al.* A lactoferrin-receptor, intelectin 1, affects uptake, sub-cellular localization and release of immunochemically detectable lactoferrin by intestinal epithelial Caco-2 cells. *J. Biochem.* **154**(5), 437–448 (2013).
12. Dierick, M., Vanrompay, D., Devriendt, B., & Cox, E. Minireview: Lactoferrin, a versatile natural antimicrobial glycoprotein which modulates host innate immunity. *Biochem. Cell Biol.* **99**(1), 61–65 (2020).
13. Wangkanont, K., Wesener, D. A., Vidani, J. A., Kiessling, L. L. & Forest, K. T. Structures of Xenopus embryonic epidermal lectin reveal a conserved mechanism of microbial glycan recognition. *J. Biol. Chem.* **291**(11), 5596–5610 (2016).
14. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**(8), 955–962 (2008).
15. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**(12), 1118–1125 (2010).
16. Jostins, L., Ripke, S., Weersma, R. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**(7422), 119–124. https://doi.org/10.1038/nature11582 (2012).
17. Jimmy, Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**(9) 979–986. https://doi.org/10.1038/ng.3359 (2015).
18. Katrina, M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**(2) 256–261. https://doi.org/10.1038/ng.3760 (2017).
19. Eric, B. *et al.* Human intelectin-1 (ITLN1) genetic variation and intestinal expression. *Sci. Rep.* **11**(1). https://doi.org/10.1038/s41598-021-92198-9 (2021).
20. Adams, D. J., Doran, A. G., Lilue, J. & Keane, T. M. The Mouse Genomes Project: A repository of inbred laboratory mouse strain genomes. *Mamm. Genome* **26**(9–10), 403–412 (2015).
21. Lilue, J. *et al.* Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* **50**(11), 1574–1583 (2018).
22. Harr, B. *et al.* Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. *Sci. Data* **3**(1), 1–14 (2016).
23. Kurosaki, T., Popp, M. W. & Maquat, L. E. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. *Nat. Rev. Mol. Cell Biol.* **20**(7), 406–420 (2019).
24. Steppan, S. J., Adkins, R. M. & Anderson, J. Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. *Syst. Biol.* **53**(4), 533–553 (2004).
25. Voehringer, D. *et al.* *Nippostrongylus brasiliensis*: Identification of intelectin-1 and-2 as Stat6-dependent genes expressed in lung and intestine during infection. *Exp. Parasitol.* **116**(4), 458–466 (2007).
26. Nei, M. & Rooney, A. P. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152 (2005).
27. Lilue, J., Shivalikanjli, A., Adams, D.J., & Keane, T.M. Mouse protein coding diversity: What's left to discover? *PLoS Genet.* **15**(11), e1008446 (2019).
28. Festing, M. F., Simpson, E. M., Davisson, M. T. & Mobraaten, L. E. Revised nomenclature for strain 129 mice. *Mamm. Genome* **10**(8), 836 (1999).
29. Mekada, K. *et al.* Genetic differences among C57BL/6 substrains. *Exp. Anim.* **58**(2), 141–149 (2009).
30. Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nat. Genet.* **43**(7), 648 (2011).
31. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**(6), 1547–1549 (2018).
32. Wehkamp, J. *et al.* Paneth cell antimicrobial peptides: Topographical distribution and quantification in human gastrointestinal tissues. *FEBS Lett.* **580**(22), 5344–5350 (2006).
33. Castillo, P. A. *et al.* An experimental approach to rigorously assess paneth cell α-defensin (Defa) mRNA expression in C57BL/6 mice. *Sci. Rep.* **9**(1), 1–14 (2019).
34. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**(1), 10–12 (2011).
35. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint 13033997 (2013).
36. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16), 2078–2079 (2009).
37. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**(15), 2156–2158 (2011).
38. Armour, J. A. *et al.* Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res.* **35**(3), e19–e19 (2007).
39. Hollox, E.J. Analysis of copy number variation using the paralogue ratio test (PRT). in *Genotyping*. 127–146 (Springer, 2017).
40. Garrison, E., & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv preprint 12073907 (2012).

Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. We acknowledge research grant support from the National Institutes of Health [U01AI125926 (Mucosal Immunology Study Team) and R37AI32738].

Author contributions

Conceptualisation—F.A., E.B.N., E.J.H. and C.L.B. Investigation—all authors. Resources—E.J.H., C.L.B., L.O.-H. and B.L. Supervision—E.J.H., C.L.B., L.O.-H. and B.L. Visualisation—F.A., E.B.N., L.O.-H. and E.J.H. Writing—original draft—E.J.H., C.L.B., E.B.N. and F.A. Writing—review and editing—all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-94679-3.

Correspondence and requests for materials should be addressed to C.L.B. or E.J.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.