Title:

Differential allelic representation (DAR) identifies candidate eQTLs and improves transcriptome analysis

Short title:
Differential allelic representation and eQTL bias in transcriptomics

Lachlan Baer[1], Karissa Barthelson[1,2], John Postlethwait[3], David Adelson[1], Stephen Pederson[4,5,6†], Michael Lardelli[1†*]
[1] School of Biological Sciences, University of Adelaide, Adelaide, SA, Australia
[2] Childhood Dementia Research Group, College of Medicine and Public Health, Flinders Health and Medical Research Institute, Flinders University, Bedford Park, SA, Australia
[3] Institute of Neuroscience, University of Oregon, Eugene, OR, USA ORCID: 0000-0002-5476-2137
[4] Black Ochre Data Labs, Indigenous Genomics, Telethon Kinds Institute, Adelaide, SA, Australia
[5] Dame Roma Mitchell Cancer Research Laboratories, Adelaide Medical School, University of Adelaide, Adelaide, SA, Australia
[6] John Curtin School of Medical Research, Australian National University, Canberra, ACT, Australia

[†]Joint senior authors
*Author for correspondence (michael.lardelli@adelaide.edu.au)

ABSTRACT

RNA-sequencing analysis excels in its ability to infer transcriptomic differences between discrete genetic states. This advantage is often leveraged to explore the consequences of single mutations causing disease, where prior knowledge of expected outcomes may not be established. Successful interpretation of such RNA-seq studies reveals the direct impacts of the mutation, as well as the homeostatic responses of the biological system. Recent studies have highlighted that, when homozygous mutations are studied in non-isogenic backgrounds, genes from the same chromosome as the mutation tend to be over-represented among differentially expressed (DE) genes. One hypothesis suggests that DE genes chromosomally linked to the mutation may not be true biological responses to the disruption of the mutation but, instead, result from differences in the representation of expression quantitative trait loci (eQTLs) that differ between the sample groups being compared. This can be problematic since including spurious DE genes in a functional enrichment study may result in incorrect inferences of mutation effect. Here we show that *chromosomally co-located differentially expressed genes* (CC-DEGs) can also be observed in analyses of dominant mutations in heterozygotes. We define a method and a metric to quantify, in RNA-seq data, localised differential allelic representation (DAR) between groups of samples subject to DE analysis. We show how the DAR metric can predict regions prone to eQTL-driven differential expression, and how it can improve functional enrichment analyses through gene exclusion or weighting of gene-level rankings. Advantageously, this improved ability to identify probable eQTLs also reveals examples of CC-DEGs that *are* likely to be functionally related to a mutant phenotype. This situation was predicted by R.A. Fisher in 1930 as due to selection for advantageous linkage disequilibrium after chromosomal rearrangements. By comparing the genomes of zebrafish (*Danio rerio*) and medaka (*Oryzias latipes*), a teleost with a conserved ancestral karyotype, we find possible examples of chromosomal aggregation of CC-DEGs during evolution of the zebrafish lineage. The ability to identify and exclude eQTL artefacts from true transcriptomic responses to mutation provides exciting opportunities for developing new approaches for analysing RNA-seq data. Our DAR metric provides a solid foundation for addressing the eQTL issue in new and existing datasets because it relies solely on RNA-seq data, which will thus improve our understanding of the mechanisms of disease-causing mutations.

INTRODUCTION

Morphologies of chromosomes change over evolutionary timescales due to the processes of translocation, inversion, duplication, and deletion. These processes change the order of genes along a chromosome, and so change the interactions between those genes. Changes in chromosome morphology create new regulatory relationships between genes and disrupt previous ones because genes can share regulatory elements and a regulatory element for one gene may be in an intron of a different nearby gene. The expression of genes is also influenced by their location within chromatin regulatory domains existing within and between chromosomes (reviewed by (1)), which are also disrupted after chromosome rearrangements.

To survive and reproduce, living cells/organisms must be resilient in the face of continuous environmental and genetic change. Such resiliency has driven the evolution of robust systems of homeostasis. One aspect of homeostasis is the ability to vary the expression of genes in different circumstances to maintain appropriate cellular functions. Transcriptome analysis monitors variations in gene expression at the RNA level. When transcriptome analysis is used to examine a biological system before and after a change (environmental or genetic), the differences in transcript levels reflect both the direct impacts of the change and homeostatic responses to it.

The alleles of genes positioned in cis on the same chromosome may show genetic linkage leading to non-random co-assortment into gametes (2). However, despite linkage, the shuffling of alleles between homologous chromosomes through recombination over successive generations means that the probability of any two particular alleles of two genes being chromosomally co-located in *cis* in a randomly selected individual should be the product of each allele's frequency in the population. Occasionally, a particular pair of alleles of two linked, functionally-related genes can act in concert to increase an individual's likelihood of survival and reproduction. Co-inheritance of these two alleles bestows a fitness advantage and the frequency of this pair in the population increases. The two alleles are said to be in "linkage disequilibrium". Positive selection for the pair of alleles can eventually drive them to fixation in the population – they become the only extant alleles for those two genes.

Sir Ronald Aylmer Fisher FRS is recognized for his very significant contributions to developing the science of statistics and, in particular, for applying statistical methods to the study of genetics. Fisher's work, '*The genetical theory of natural selection'* published in 1930 (3) is regarded as central to the "modern synthesis" reconciling Darwin's theory of evolution with Mendelian genetics. In discussing the phenomenon of linkage disequilibrium between pairs of alleles across a population he asserted that,

> "*…the presence of pairs of factors in the same chromosome, the selective advantage of each [pair] of which reverses that of the other, will always tend to diminish recombination, and therefore to increase the intensity of linkage in the chromosomes of that species*."

An expected outcome of this tendency, acting in an evolutionary background of substantial chromosome rearrangement, is that rearrangements establishing linkage between functionally-related genes facilitate selectively advantageous linkage disequilibrium. Therefore, we predict that, over evolutionary timescales:

1) Functionally-related genes may group together to show linkage within chromosomes.

2) The smaller the selective advantage of functionally-related gene pairing, the tighter the genetic linkage between them must be to exert positive selective pressure on chromosome structure over evolutionary time.

3) A gene with multiple, distinct functions (each subject to independent homeostatic mechanisms) will drive accumulation of sets of genes distinct for each function near that gene on the same chromosome.

Considerable evidence exists for the clustering of functionally-related genes on chromosomes in both prokaryotes and eukaryotes (4,5). Such clustering should be revealed as CC-DEGs in transcriptome data when mutations change the activity of those chromosomally co-located genes and initiate homeostatic responses. However, the reality of functional relationships between CC-DEGs is challenged by the possibility that their observed differential expression results from unequal distribution between experimental comparison groups of alleles with discrete levels of expression (due to differences in allelic transcription rate and/or transcript stability), i.e. "expression quantitative trait loci", eQTLs, as illustrated in Figure 1. Indeed, obvious enrichments of DE genes on the same chromosome as a mutation-of-interest have previously been observed in analyses involving comparisons of homozygous mutant with homozygous wild-type individuals (6,7). We recently observed such CC-DEGs when comparing transcriptomes from pools of embryos either wild-type or homozygous for a loss-of-function mutation in the zebrafish gene *fmr1* (orthologous to the human *FMR1* gene mutated in Fragile X Syndrome) (8). However, that analysis also revealed enrichment in

homozygous mutant embryos for a particular gene set previously seen as affected in *Fmr1* knock-out mice (8,9), supporting the possibility that some of the CC-DEGs do participate in *fmr1* functions, and highlighting challenges in this aspect of transcriptome analysis.

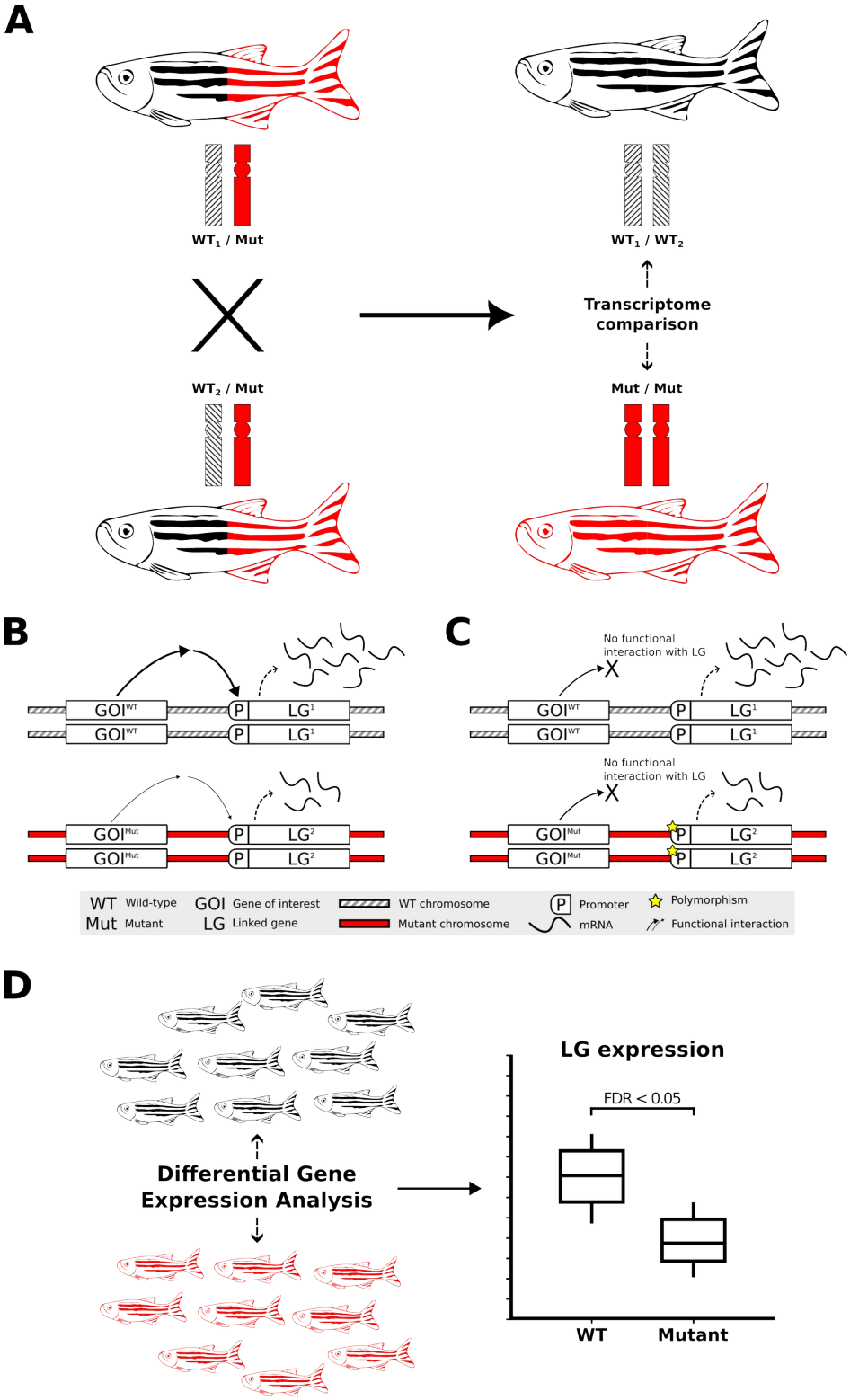


Figure 1. Transcriptome analysis of homozygous mutants compared to their wild-type siblings and the impact of non-isogenic genetic backgrounds on gene expression. **A)** Experimental selection of progeny homozygous for a mutant allele of a gene-of-interest (GOI, mutation-bearing chromosome indicated in red) necessarily involves increased homozygosity for alleles of genes linked to that mutation (i.e. on the same chromosome). The rates of transcription or transcript degradation for these alleles may differ significantly from their corresponding alleles on the homologous wild-type chromosomes (shaded differentially to illustrate that these wild-type chromosomes are not isogenic). **B)** Differential expression of alleles of a linked bystander gene, LG, between wild-type and mutant chromosomes due to a functional interaction between

the GOI and the LG. **C)** eQTL-driven differential expression of LG between wild-type and homozygous mutant chromosomes in the absence of a functional interaction between the GOI and the LG. The expression of LG differs independently of the GOI genotype when LG's different alleles are eQTLs. **D)** Breeding to produce genotype groups (e.g., homozygous GOI mutants for comparison to wild-type) selects for differential representation between those groups for the alleles of neighbouring LGs. When those alleles are eQTLs, they can show differential expression between the groups that is not a phenotypic effect of the mutation but can, mistakenly, be inferred as such.

White et al. (10) recently published an analysis of gene expression differences between the individual, inbred descendants of an ancestral cross between two discrete isogenic zebrafish strains. This allowed comparison of gene expression from discrete chromosomal segments isogenic for one or other of the ancestral genotypes. The authors demonstrated evidence for the existence and importance of eQTLs in DE gene analysis, and suggested some ways to identify such eQTLs in transcriptome analyses.

Chromosomally co-located allele specific expression has been shown to affect RNA-seq analyses in studies involving recessive mutations (10), but has not yet been investigated for dominant mutations. An early-onset inherited (familial) form of Alzheimer's disease (EOfAD) is caused by dominant alleles of a small number of genes. The majority of EOfAD mutations are missense mutations (or in-frame insertions/deletions) of the human gene *PRESENILIN 1* (*PSEN1*) (11). In contrast, some frameshift mutations in this gene cause a very distinct, dominant, functional effect - an inflammatory skin disease, familial Acne Inversa (fAI) without Alzheimer's disease (12–14). We have examined the young adult brain transcriptomes from a number of zebrafish models of these mutations (12,15–22). As expected from the very different disease phenotypes of human *PSEN1* frame-preserving and frameshift mutations, our previously published analyses of these zebrafish mutation model transcriptomes have revealed very distinct functional effects on energy metabolism and cell signalling pathways (12). However, among the genes detected as DE between the brains of the wild-type and the dominant heterozygous mutant zebrafish, we were surprised to observe that a greater than expected number were chromosomally colocalised with the mutation-of-interest (i.e. many are CC-DEGs). Here we investigate this colocalization phenomenon further by examining between-group differential allelic representation in brain transcriptome datasets from zebrafish. We also extend our investigation to similar mouse transcriptome datasets. We find strong support for the importance of eQTLs, but also for Fisher's postulate on colocalization of functionally related genes. We show that using differential allelic representation information can improve the detection of significant gene sets.

RESULTS

Chromosomal co-localisation of DE genes in heterozygous mutant analyses

Our previous observation of CC-DEGs in our analysis of homozygous *fmr1* mutant zebrafish embryos (8) led us to re-examine our numerous adult zebrafish brain transcriptome datasets where we had previously analysed the effects of dominant mutations in heterozygotes. Surprisingly, we saw CC-DEGs in a number of our analyses as illustrated by Manhattan plots of DE gene expression (Figure 2A and Supplementary Figure 1A-B). To confirm that CC-DEGs are not a phenomenon specific to zebrafish, we searched in public databases of transcriptome data for other datasets comparing wild-type and mutant (but non-transgenic) individuals. While suitable datasets were few, we found that an analysis of different *APOE* alleles (human nomenclature to denote the use of humanised alleles) in mice showed CC-DEGs for 3-month-old cerebral cortices homozygous for the Alzheimer's disease-protective ε2 allele (*APOE2*) compared to cortices homozygous for the neutral-risk *ε3* allele (*APOE3*, (23), Figure 2B). Interestingly, CC-DEGs were less obvious in the analysis of cortices homozygous for the Alzheimer's disease-risk *ε4* allele (*APOE4*, Supplementary Figure 1D), supporting that mutations/alleles that are more disruptive to cell biology cause the appearance of DE genes distributed throughout the genome.

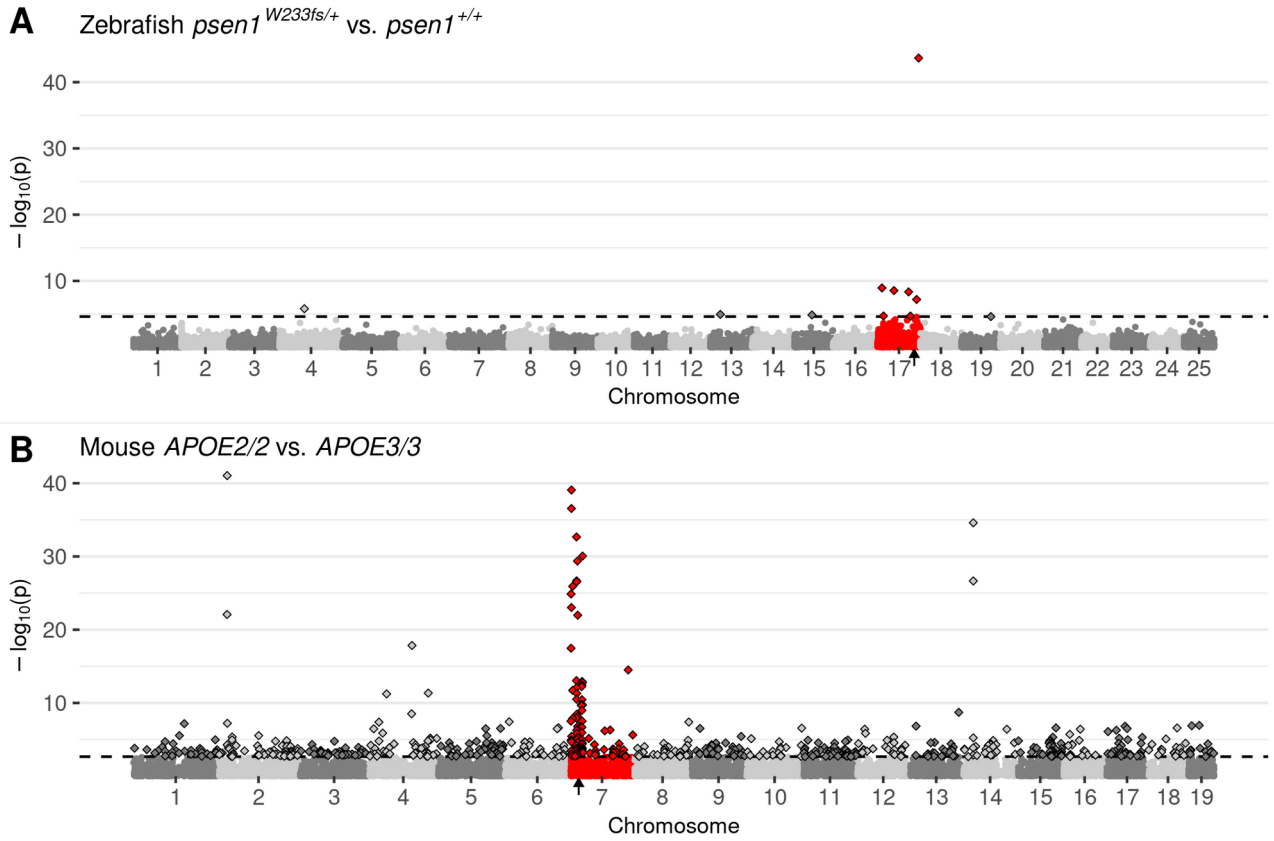


Figure 2. Manhattan plots highlighting CC-DEGs from differential expression testing in: **A)** Brain transcriptomes from a comparison of zebrafish *psen1*$^{W233fs/+}$ vs. *psen1*$^{+/+}$ siblings at 6 months of age. *psen1*^{W233fs} is a dominant fAI-like frameshift allele. **B)** 3-month-old cerebral cortex transcriptomes from a comparison of homozygous male mice bearing humanised *APOE2* or *APOE3* alleles. Non-random accumulations of DEGs are supported by Bonferroni-adjusted Fisher's exact test *p*-values for enrichment of DE genes on the mutant chromosome, A) *psen1*$^{W233fs/+}$ vs. *psen1*$^{+/+}$: p = 7.41e-7, B) *APOE2/2* vs. *APOE3/3*: p = 2.03e-10. Genes are plotted along the x-axis based on their genomic positions along the chromosomes distinguished by alternating shades of grey. Genes on the chromosomes containing the mutations are highlighted in red. The approximate locations of the mutated genes are indicated with small black arrows on the x-axis. The raw *p*-values are plotted along the y-axis at the -$\log_{10}$ scale such that the most significant genes exist at the top of the plot. The cut-off for gene differential expression (FDR-adjusted *p*-value < 0.05) is indicated by a dashed horizontal line. Genes classified as differentially expressed under this criterion are represented as diamonds with a black outline.

White et al.'s analysis (10) of DE genes in the inbred descendants of two wild-type isogenic zebrafish strains strongly supports the conclusion that many CC-DEGs are eQTLs and showed how homozygosity for regions of the genome can enhance observation of allelic expression differences. Transcriptome comparisons

between individuals or groups of individuals should not reveal eQTLs when comparisons are made within single isogenic (e.g. inbred) backgrounds or when sequence differences between comparison groups (here described as differential allelic representation, DAR) are sufficiently limited. Therefore, we sought to test whether DAR between genotype groups in transcriptome datasets could be used to determine the likelihood that any particular gene is differentially expressed due to eQTL effects. To facilitate this analysis, we assembled a computational workflow to quantify DAR between samples (see below).

Assembly of a pipeline for analysis of differential allelic representation (DAR) in transcriptome datasets

Our pipeline for calculating a differential allele representation (DAR) metric between sample groups is summarised in Figure 3 (see Methods for extended details). Briefly, we determined single nucleotide variants based on the GATK Best Practice Workflow for RNA-seq short variant discovery (24). Monoallelic calls (where only the reference allele is reported at a single nucleotide locus across all samples) were excluded due to the computational implications for downstream processing. Bi- and multiallelic calls (one or more alternative alleles reported at a single nucleotide locus across all samples) were then analysed using the R programming language (25). We present the calculation of the DAR metric as a modification of the Euclidean distance formula,

$$DAR_{1,2} = \sqrt{\frac{\sum_{i \in \{A,C,G,T\}} \left(p_{i_1} - p_{i_2}\right)^2}{2}} \quad (1)$$

where $DAR_{1,2}$ represents the DAR value between sample groups 1 and 2 at a single nucleotide locus. p_{i_1} and p_{i_2} correspond to the proportion of the total reported allele counts (A, C, G or T) in sample groups 1 and 2 respectively at a single nucleotide locus. Dividing the Euclidean distance of allele proportions by √2 results in an easy-to-interpret DAR metric where 0 represents identical allele proportions and 1 represents complete difference (non-identity).

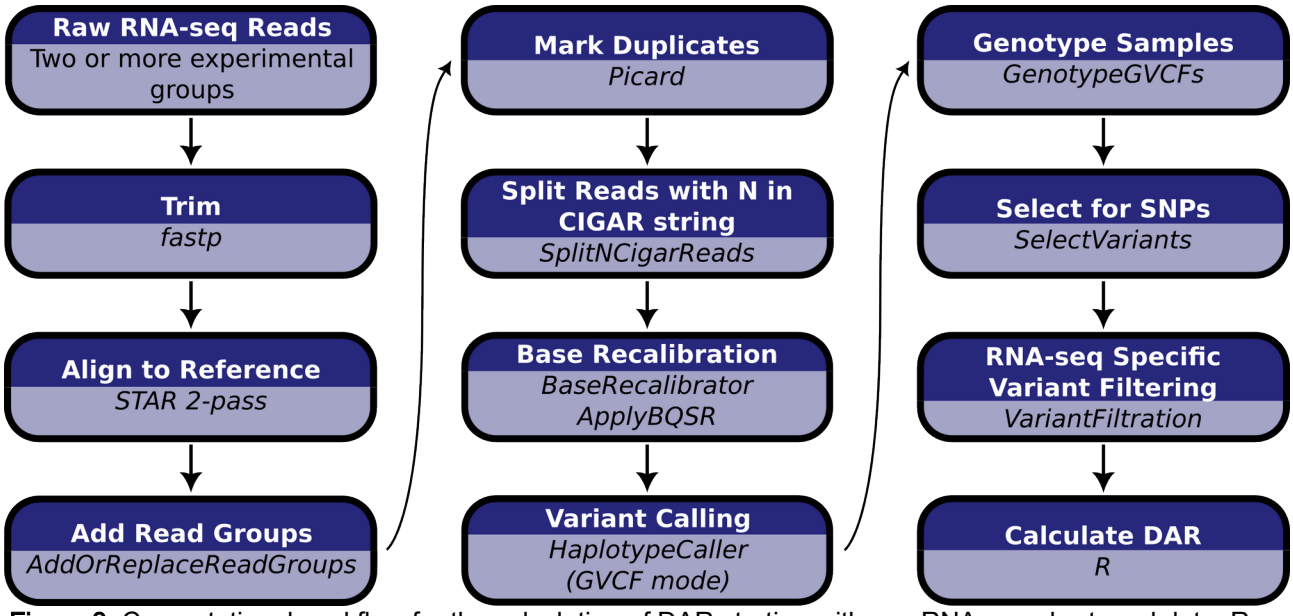


Figure 3. Computational workflow for the calculation of DAR starting with raw RNA-seq short read data. Raw RNA-seq reads must consist of at least two experimental groupings to allow the calculation of DAR between them.

DAR analysis of the humanised *APOE* alleles in a mouse cerebral cortex transcriptome dataset

After establishing the pipeline to quantify DAR between samples, we quantified variation in this metric across entire genomes with a particular focus on chromosomes containing mutations of interest. We chose the mouse *APOE* cortex transcriptome dataset for an initial DAR analysis as it displayed the strongest signal of CC-DEGs on the mutant chromosome (Figure 2B). For visualisation purposes, and to alleviate noise, we smoothed the DAR metric using an elastic sliding window approach to calculate the mean DAR value at each single nucleotide variant locus along with the nearest five loci on either side (n = 11). See Supplementary Figure 2 for a summary of the elastic window sizes. Interestingly, the cortex transcriptome comparison of mice homozygous for *APOE2* with those homozygous for *APOE3* revealed a region containing peaks of very

high DAR localised around the *APOE* gene on chromosome 7 (Figure 4). An inspection of the cumulative distribution of DAR across each chromosome supports the notion that the mutant chromosome contains the most DAR of all the chromosomes (Supplementary Figure 3). In a completely isogenic strain, this result would not be expected. However, the mice in this study were engineered through targeted replacement of the murine *Apoe* locus with human *APOE* alleles by homologous recombination in mouse 129 strain embryonic stem cells followed by repeated backcrossing into the C57BL/6J strain (26). The strain 129 alleles of genes located near (in tight linkage with) the *APOE* locus are less likely to be replaced by strain C57BL/6J alleles through recombination than alleles of syntenic genes distant from *APOE* or alleles of genes on other chromosomes.

Having identified a region of high DAR potentially prone to eQTL-dependent effects, we looked for CC-DEGs within our differential expression analysis. We observed a clear clustering of CC-DEGs around the *APOE* gene within the identified regions of high DAR, strongly supporting the existence of eQTLs as drivers of differential expression (Figure 4). The remaining DE genes on chromosome 7 found within regions of low DAR are more likely to represent biological responses to the mutation because they are DE even without close linkage. However, genes may become DE as a secondary effect of eQTLs located elsewhere in the genome (27), so the existence of a DE gene in a region of low DAR does not confirm a direct, primary functional relationship with any mutation. We also noticed that the *APOE* region showed relatively low DAR compared to the surrounding region, which was unexpected as each sample group has a different *APOE* allele. On closer inspection, it was evident that variants in this region were removed during quality control filtering due to being classified as potential single nucleotide polymorphism (SNP) cluster artefacts (> 3 SNPs in a 35bp window). This most likely occurs because the *APOE* alleles in these mice are humanised and therefore exhibit substantial variation in the *APOE* region as reads were aligned to the mouse genome (GRCm39, Ensembl release 104).

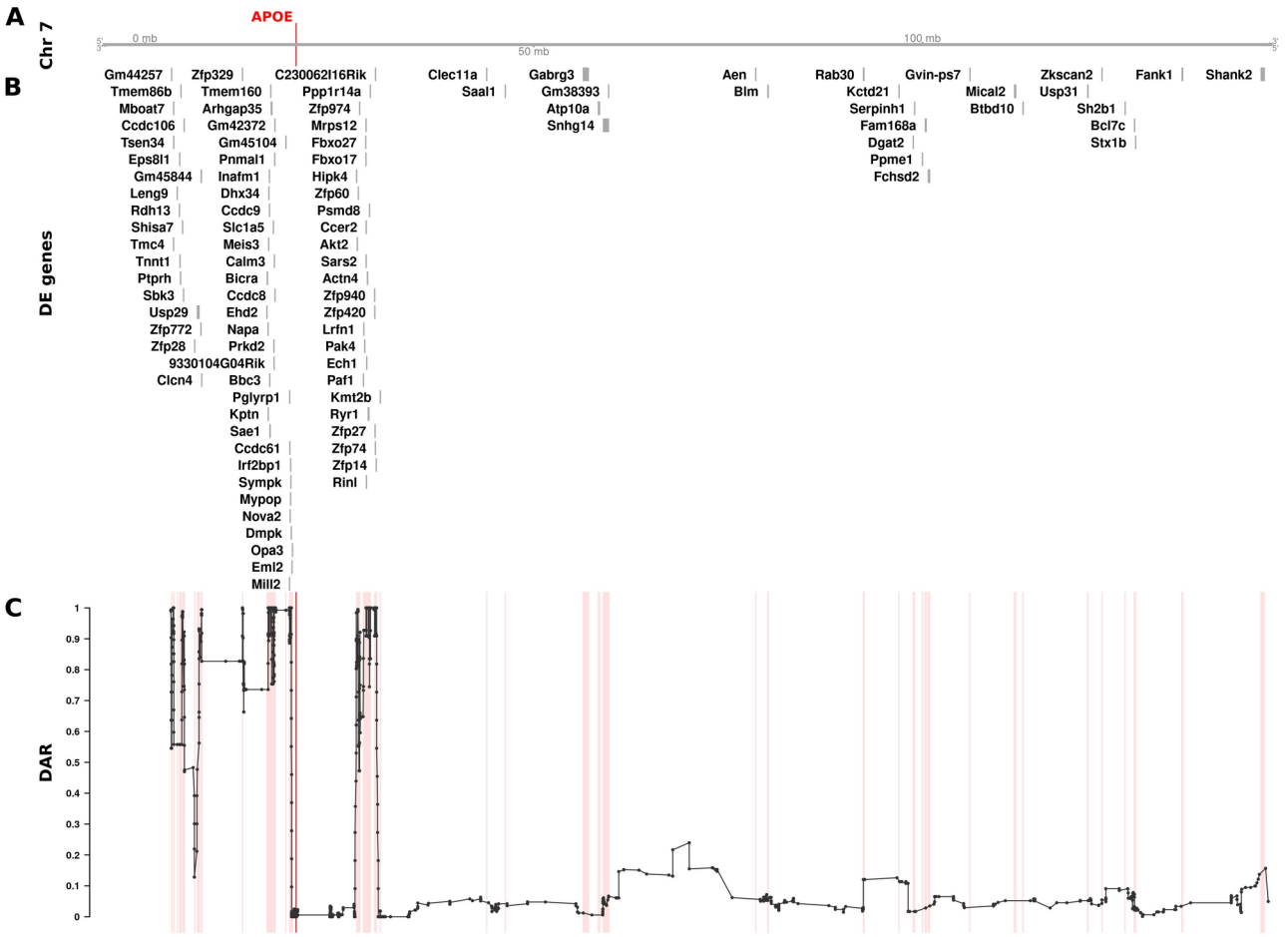


Figure 4. The relationship between DE genes and DAR along the entirety of mouse Chromosome 7 between male *APOE2/2* and *APOE3/3* mouse cerebral cortices (at 3 months of age). The plot contains four sets of information represented by separate tracks horizontally. Track A represents the axis of Chromosome 7. The position of the *APOE* gene is marked and labelled in bold red. Track B displays differentially expressed genes according to their positions along the chromosome. 99 of 1126 total genes (8.79%) on Chromosome 7 that were expressed in the dataset were classified as DE (FDR < 0.05). Track C shows the trend in DAR as a connected scatterplot with each point in black representing the DAR value at a single nucleotide variant

position (elastic sliding window size = 11 variants). Positions of the DE genes shown in track B are indicated by light red lines.

Identified zebrafish *psen1* CC-DEGs are unlikely to be eQTLs

DAR analysis of the mouse *APOE* dataset showed convincing evidence implicating eQTLs as the primary drivers of differential expression for a large proportion of genes. Unexpectedly, our analysis of DAR in the comparisons of 6-month-old brain transcriptomes in zebrafish *psen1*$^{T428del/+}$ (EOfAD-like) vs. *psen1*$^{+/+}$ siblings and in *psen1*$^{W233fsl/+}$ (fAI-like) vs. *psen1*$^{+/+}$ siblings did not support the hypothesis that those DE genes clustered near *psen1* on zebrafish Chromosome 17 are eQTLs. Nevertheless, the identified DE genes are indeed clustered with *psen1* on Chromosome 17 (supported by Bonferroni-adjusted Fisher's exact test *p*-values for enrichment of DE genes on Chromosome 17 despite consistently low DAR along this chromosome; *psen1*$^{T428del/+}$ vs. *psen1*$^{+/+}$: $p = 8.63e\text{-}3$, *psen1*$^{W233fsl/+}$ vs. *psen1*$^{+/+}$: $p = 7.41e\text{-}7$). As shown in Figure 5, the DAR metric along the length of Chromosome 17 in the *psen1*$^{T428del/+}$ vs. *psen1*$^{+/+}$ and the *psen1*$^{W233fsl/+}$ vs. *psen1*$^{+/+}$ comparisons was uniformly below 0.2, as expected from the fact that these mutations were generated in one, very inbred, strain of zebrafish. Indeed, the Tübingen (Tu) strain zebrafish in the facility used to produce these mutant fish were bred in relatively small populations in isolation for over 15 years. It is also notable that the genes that were DE due to the EOfAD-like *psen1*^{T428del} mutation were not the same as those DE due to the fAI-like *psen1*^{W233fs} mutation, consistent with these two very different types of mutation in the same gene (reading frame-preserving vs. frameshift respectively) having discrete molecular effects (as revealed and discussed in our previously published analysis (12)). (Note that the mechanisms underlying the differential effects of frame-preserving and frameshift mutations on *PSEN1* function are currently unclear (28)). The genes that were DE due to the EOfAD-like mutation are clustered very close to *psen1,* while those DE due to the fAI-like mutation are greater in number and more widely distributed on Chromosome 17 (Figure 5). This finding is consistent with the second prediction from Fisher's assertion (gene pairings with a smaller selective advantage require tighter genetic linkage, see Introduction) because the relatively late-acting (after reproductive maturity), dementia-inhibiting function of *psen1* that is disrupted by the EOfAD-like mutation should only exert mild selective pressure on the organism's survival and reproduction. In contrast, the fAI-like mutation affects the role of Psen1 protein in Notch signalling and in other cellular signalling pathways that are important in embryo development and immune system function, and so likely exerts stronger selective effects. It is also relevant to note that meiotic recombination is suppressed in male zebrafish relative to females so that recombination distances along of the majority of chromosomes, including Chromosome 17, are less than 50 centiMorgans in males (29). This means that, in male zebrafish, alleles of any two genes anywhere on Chromosome 17 are genetically linked.

Due to the low numbers of DE genes in this dataset, we looked for additional evidence that the clustering of DE genes on Chromosome 17 had not occurred by chance. Therefore, we investigated whether DE clustering was still observed after applying a less stringent DE significance threshold. We performed an additional analysis using an FDR-adjusted *p*-value < 0.15 (i.e. 3α for α = 0.05) allowing for detection of additional DE genes, along with a likely increase in false positives. This change increased the total number of DE genes from 10 to 20 across the entire genome in the EOfAD-relevant *psen1*$^{T428del/+}$ vs. *psen1*$^{+/+}$ comparison, and from 11 to 30 in the fAI-relevant *psen1*$^{W233fsl/+}$ vs. *psen1*$^{+/+}$ comparison. Both comparisons remained significant for enrichment of DE genes on Chromosome 17 with Bonferroni-adjusted Fisher's exact test *p*-values of 1.88e-2 and 2.73e-12 respectively.

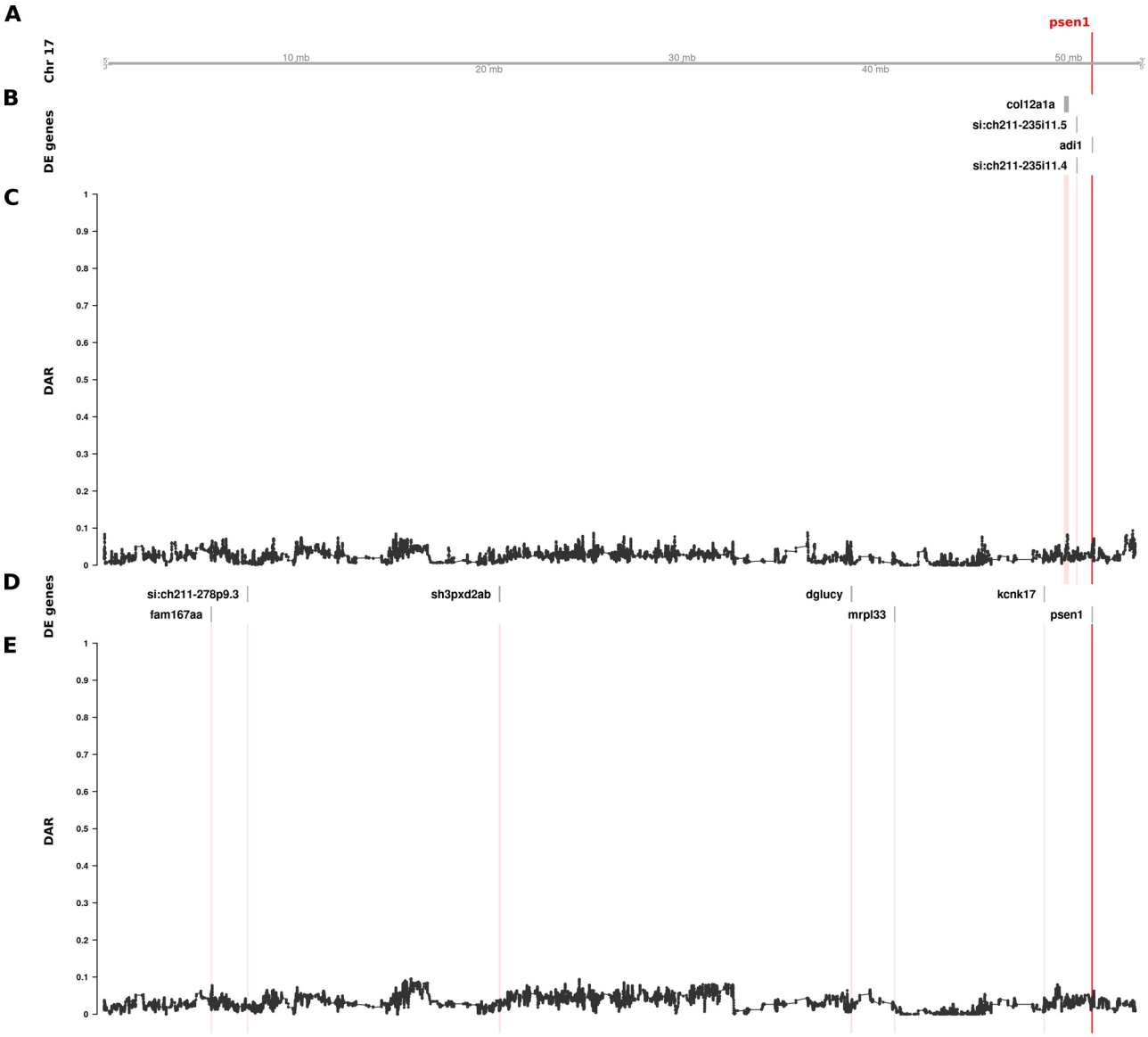


Figure 5. The relationship between DE genes and DAR along the entirety of zebrafish Chromosome 17 in brain transcriptome comparisons of 6-month-old wild type fish against sibling fish heterozygous for either of two different *psen1* mutations. Track A represents the axis of Chromosome 17. The position of *psen1* is marked and labelled in bold red. Tracks B and C show the results from the EOfAD-relevant *psen1*$^{T428del/+}$ vs. *psen1*$^{+/+}$ comparison, while tracks D and E show the fAI-relevant *psen1*$^{W233fsI/+}$ vs. *psen1*$^{+/+}$ comparison. Tracks B and D display DE genes from their respective comparisons, while tracks C and E show the trend in DAR as a connected scatterplot with each point representing the DAR value at a single nucleotide variant position (elastic sliding window, n = 11 variants). The positions of DE genes are highlighted on the DAR tracks in a light red colour.

Aggregation of *psen1* CC-DEG over evolutionary time

Our analysis of zebrafish *psen1* mutant brain transcriptomes supports Fisher's contention that selection for linkage disequilibrium may influence chromosome structure over evolutionary timescales. Therefore, we looked for evidence of structural rearrangements favouring clustering of DE genes during teleost (bony fish) speciation. Medaka (*Oryzias latipes*) is a small freshwater teleost sharing a common ancestor with zebrafish approximately 290 million years ago (30). Its widespread use as a model organism has encouraged detailed characterisation of its genome. The medaka genome has not undergone any major chromosomal rearrangements in the 323 ± 9.1 million years since divergence from its common ancestor with zebrafish (31). Therefore, it provides a suitable reference outgroup to assess how the zebrafish genome has changed over time.

We accessed The Synteny Database (32) to assess whether CC-DEG revealed in our various analyses exist on separate chromosomes in the medaka genome. Although our efforts were limited by low numbers of DE

genes, we did indeed find evidence supporting that some *psen1* CC-DEG have become chromosomally co-located during zebrafish evolution. Of the seven DE genes chromosomally co-located with *psen1* (on zebrafish Chromosome 17) in fAI-like *psen1*^{W233fs} mutants, two (*fam167aa*, *sh3pxd2ab*) were predicted to have orthologues on separate medaka primary chromosomes, while another two (*psen1*, *dglucy)* were predicted to be located on alternate scaffolds (Supplementary Figure 4). The remaining three DE genes (*mrpl33*, *kcnk17, si:ch211-278p9.3*) were either not present in the Ensembl version 71 database or not predicted to have medaka orthologues. As Ensembl version 71 was released in 2013 we wanted to confirm our findings with the most recent Ensembl database (version 108, released 2022). By searching the gene names of the zebrafish genes DE on Chromosome 17, we confirmed that *fam167aa* and *sh3pxd2ab* are located on medaka Chromosomes 24 and 1 respectively as predicted by The Synteny Database. However, four DE genes (*psen1*, *dglucy*, *mrpl33*, *kcnk17)* where orthologues could not be found on the medaka primary chromosomes of Ensembl version 71 have since been assigned to medaka Chromosome 22 and represent syntenies since the last common ancestor of medaka and zebrafish. While these four genes on medaka Chromosome 22 have been linked together with *psen1* for at least 323 ± 9.1 million years in zebrafish, *fam167aa* and *sh3pxd2ab* may be examples of genes that have been captured more recently during zebrafish evolution due to functional interactions with *psen1*.

Using a DAR metric to exclude probable eQTLs improves gene set enrichment analyses

eQTL artifacts due to linkage represent noise that could reduce the sensitivity of functional pathway analysis of mutations. Therefore, we investigated whether we could reduce such noise (exclude eQTLs) by excluding genes located in those parts of the genome differing between mutant and wild-type individuals. Using the previously calculated DAR metric across the genome, we set gene exclusion thresholds at DAR values ranging from 0.1 to 1.0 at 0.1 intervals prior to functional analysis. We excluded from analysis those genes located in regions with a DAR value greater than the threshold. As the maximum possible DAR metric value is 1, a threshold at this value does not remove any genes and, therefore, represents the results of analyses where no eQTLs are excluded. We utilised two well-established and complementary pathway analysis methods at each of the ten thresholds to observe the effects of DAR metric-based filtering; Rotation Gene Set Testing (ROAST) implemented using the *fry* function in Bioconductor package *limma* (33,34), and Gene Set Enrichment Analysis (GSEA) implemented using Bioconductor package *fgsea* (35,36). We chose the 186 *a priori* gene sets defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) (37) to be tested in each analysis.

To examine the effects of different DAR metric thresholds, we applied these to a previously unpublished zebrafish brain transcriptome dataset generated by our laboratory that includes transcriptome data on the effects of homozygosity for a mucopolysaccharidosis type-IIIB (MPS-IIIB)-like mutation in the gene *naglu*, *naglu*^{A603Efs}, at 7 days post fertilisation (dpf, see also **Data description and generation**). This dataset displays a highly variable pattern of DAR along the mutant chromosome and clustering of DE genes near the mutation site (Figure 6). Mutations in the human *NAGLU* gene causing MPS-IIIB have a clearly understood pathological mechanism; they reduce a lysosomal enzyme activity required in the degradation of heparan sulfate (38). In contrast to EOfAD, where the pathological mechanism(s) is disputed, interpretation of the results of functional analyses of mutations causing MPS-IIIB is less ambiguous. In fact, the KEGG collection of gene sets includes a gene set/pathway directly related to heparan sulfate degradation, "*Glycosaminoglycan Degradation*". We performed functional analysis of *naglu*$^{A603Efs/A603Efs}$ vs. *naglu*$^{+/+}$ siblings from this dataset at each of the aforementioned thresholds and the results are summarised in Figure 7.

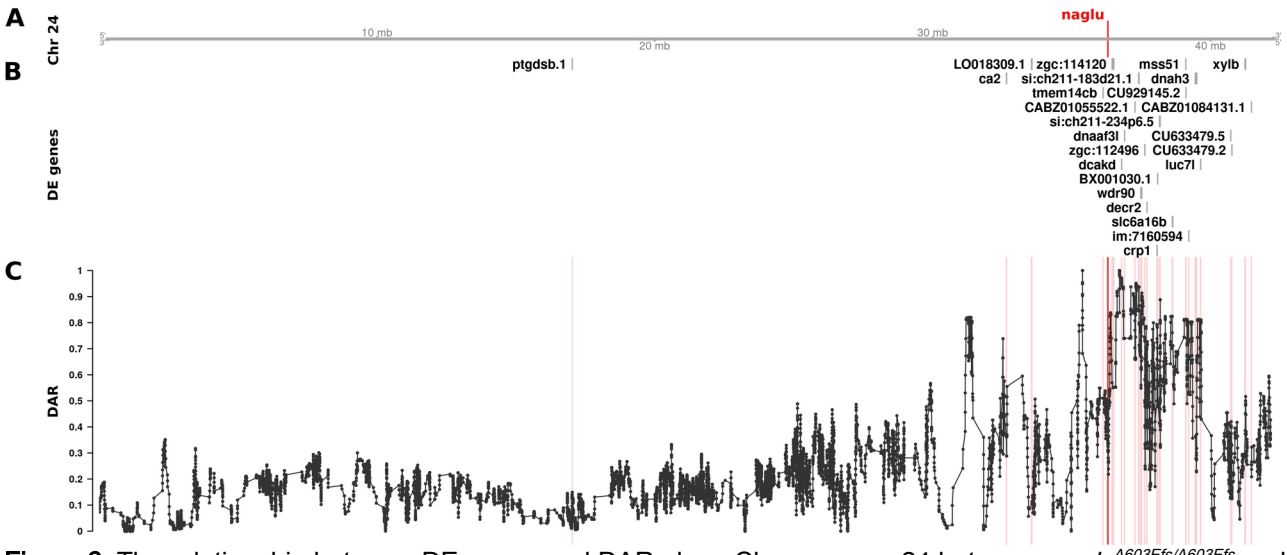


Figure 6. The relationship between DE genes and DAR along Chromosome 24 between *naglu*$^{A603Efs/A603Efs}$ and *naglu*$^{+/+}$ 7 dpf larval zebrafish. Track A represents the axis of Chromosome 24. The position of *naglu* is marked and labelled in bold red. Track B displays DE genes according to their positions along the chromosome. Track C shows the trend in DAR as a connected scatterplot with each point representing the DAR value at a single nucleotide variant position (elastic sliding window, n = 11 variants). Positions of the DE genes shown in track B are indicated by light red lines.

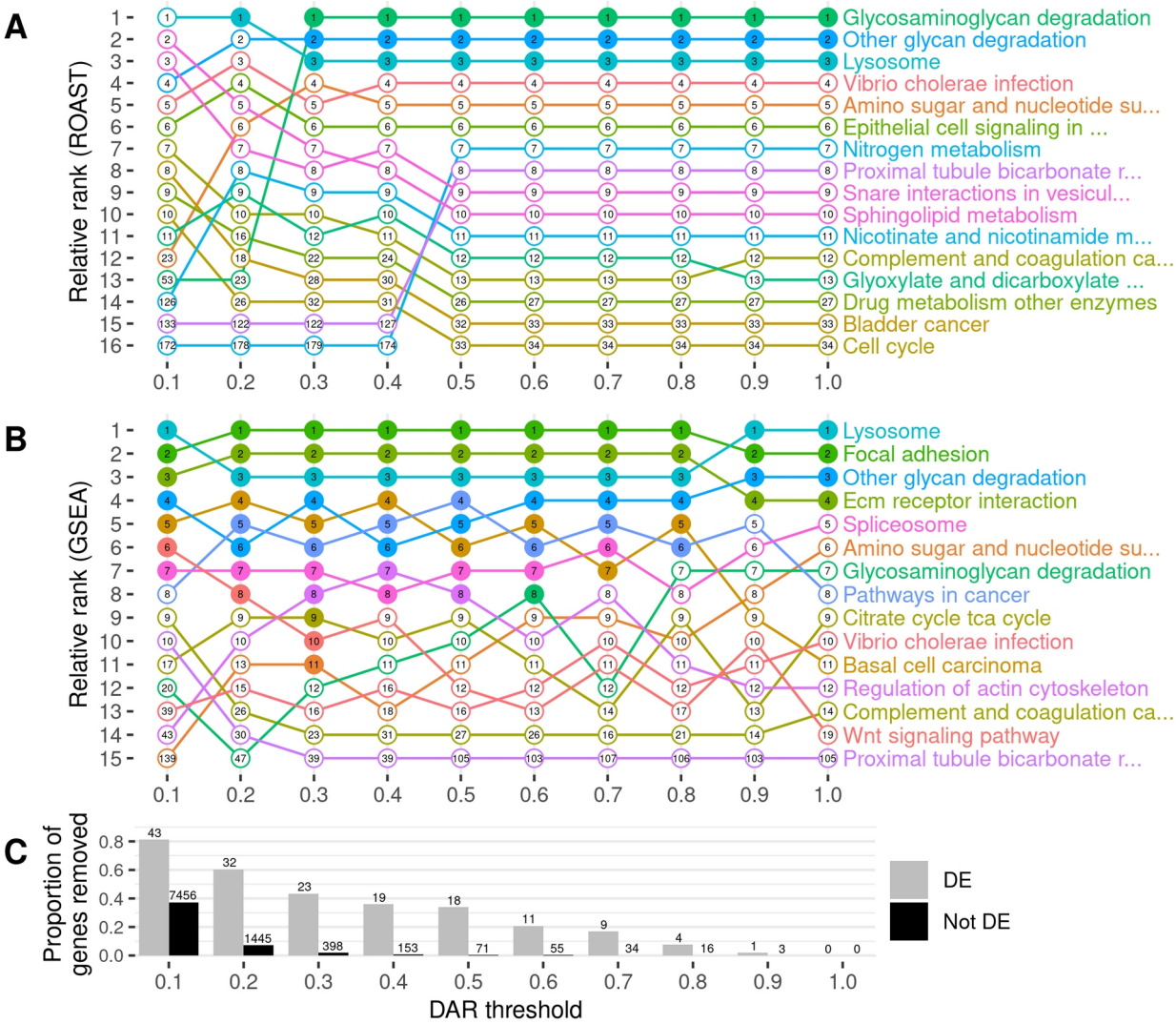


Figure 7. The impact of gene exclusion by DAR thresholding on the outcomes of functional enrichment analysis using KEGG gene sets. Panel **A** displays the outcomes of ROAST, while panel **B** displays the outcomes of GSEA. Gene sets are displayed only if they were found in the top ten most significant gene sets for at least one DAR threshold. The relative ranking between the displayed gene sets is represented along the y-axis for each threshold indicated on the x-axis. Filled dots indicate that the gene set was classified as significantly enriched (FDR-adjusted *p*-value < 0.05) at the respective threshold. The numbers inside the dots show the overall ranking of the gene set among all 186 KEGG gene sets tested. Panel **C** displays the proportion of DE (grey) and non-DE (black) genes that were removed at each threshold. The number of genes this equates to is displayed above each bar of the chart.

Setting DAR thresholds caused only minor impact on the ranking and significance of gene sets when using ROAST. The lower thresholds of 0.1 and 0.2 showed clear perturbations in rankings but this is unsurprising due to the large proportion of genes excluded. Results at low thresholds would be expected to be unreliable due to the removal of genes representing real gene expression differences caused by the mutation. This is evident from the loss of statistical significance of all three pathways that had previously shown significance at thresholds greater than 0.2. The consistent significance of the pathways (gene sets) *Glycosaminoglycan Degradation*, *Other Glycan Degradation*, and *Lysosome* at most thresholds gives confidence in the results of ROAST because these three pathways have previously been implicated in MPS-IIIB (39).

In contrast to the effects on ROAST, the imposition of DAR thresholds before GSEA offers a number of interesting observations. Substantial alterations in rankings are seen crossing each threshold for most pathways. This trend remains true even when very few genes are removed from the analysis, for example by DAR thresholds at 0.8 and 0.9 (20 and 4 genes excluded in total respectively). A general trend showing an increasing number of significant gene sets is observed as the DAR metric threshold is lowered. Eleven pathways are classified as significant at a threshold of 0.3 in comparison to four pathways when no filtering was performed (i.e. at a DAR threshold = 1). Most notably, the pathway *Glycosaminoglycan Degradation* that is relevant in MPS-IIIB achieves statistical significance uniquely at a DAR threshold of 0.6 where a total of 66 genes (11 DE, 55 not DE) are excluded from the analysis. Interestingly, and despite being an expected true

positive, this pathway does not achieve significance at any other threshold. Closer inspection of the *p*-values of pathways that achieved significance at one threshold or more (Supplementary Figure 5), found that most pathways return smaller *p*-values as the DAR metric threshold is lowered, until too many genes are excluded. The *Lysosome* and *Regulation of Actin Cytoskeleton* pathways show this trend even as the number of leading-edge genes (those driving the enrichment) become fewer.

Due to our finding that GSEA produced a greater number of significant gene sets at lower DAR thresholds, we examined more closely the genes remaining after imposition of a DAR threshold = 0.3. Interestingly, Fisher's exact test revealed that significant enrichment for clustering of DE genes on Chromosome 24 persisted (Bonferroni-adjusted *p*-value = 6.16e-3) despite the exclusion of 20 of the 25 DE genes initially found on this chromosome. Two of the five remaining genes on Chromosome 24 at DAR=0.3 can be found in The Synteny Database. These genes exist on medaka chromosomes 17 and 20 indicating their aggregation onto Chromosome 24 during the past ~323 million years, see Supplementary Figure 6.

Using DAR to weight the GSEA ranked list

GSEA requires a ranked list of genes to calculate gene set enrichment scores. In typical GSEA, a ranking statistic is calculated from the results of differential expression analysis, which is used to define the order of the ranked list. Often, the ranking statistic is calculated by multiplying the sign of each gene's $\log_2$ fold-change value by $-\log_{10}$ of its respective *p*-value, resulting in the most significant up-regulated genes being positioned at the top of the list, and the most significant down-regulated genes at the bottom. Genes at the top or bottom of the list provide a greater contribution to their associated pathways' enrichment scores than those in the middle of the list. To deemphasise the contribution to differential expression of genes more likely to be eQTLs, we weighted each gene's ranking statistic prior to GSEA by multiplying it by the complement of the gene's DAR value (1 – DAR). See Supplementary Figure 7 for the impact of weighting on gene rankings in the *naglu*$^{A603Efs/A603Efs}$ vs. *naglu*$^{+/+}$ 7 dpf sibling larval transcriptome dataset. This resulted in genes existing in high DAR regions being pushed towards the middle of the list, where they would have less impact on the GSEA algorithm. Weighting gene ranks using the DAR metric in the *naglu*$^{A603Efs/A603Efs}$ vs. *naglu*$^{+/+}$ dataset caused one pathway that was not initially significant, *Glycosaminoglycan Degradation,* to achieve significance (FDR-adjusted *p*-value = 1.56e-2). Given the expectation of *Glycosaminoglycan Degradation* being a true positive, this result supports DAR-weighting as a viable strategy when using GSEA. The other five initially significant pathways, *Lysosome*, *Focal Adhesion*, *Other Glycan Degradation*, *ECM Receptor Interaction*, and *Spliceosome*, all retained their significance when using the weighted approach (Figure 8).

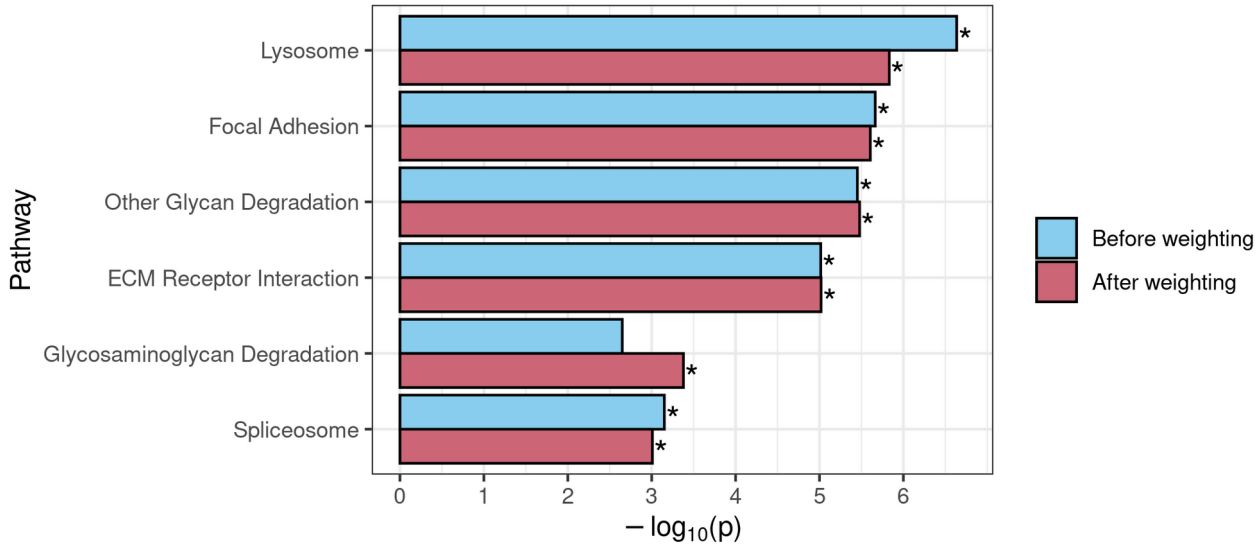


Figure 8. Comparison of GSEA results for KEGG gene sets that achieved significance before and/or after use of DAR to weight the gene-level ranking statistic in the *naglu*$^{A603Efs/A603Efs}$ vs. *naglu*$^{+/+}$ 7 dpf sibling larval transcriptome dataset. An asterisk indicates that a pathway was determined to be significantly enriched (FDR-adjusted *p*-value < 0.05).

We also examined the effects of DAR-weighting in our separate brain transcriptome dataset from 6-month-old zebrafish that compares a heterozygous mutation in the *sorl1* gene (*sorl1*$^{R122Pfs/+}$, an EOfAD-like model) to wild-type siblings (21). This dataset displayed CC-DEG (Supplementary Figure 1) and DAR levels of around 0.4 were noted as a concern (Supplementary Figures 8, 9). A comparison of GSEA results before and after weighting also showed compelling results for pathways we expect to be perturbed by the mutation

(Supplementary Figure 10). Gene sets associated with neurodegeneration, i.e. *Parkinsons disease*, *Huntingtons disease* and *Alzheimers disease*, all showed increased significance using the DAR-weighting method. Additionally, *Oxidative phosphorylation*, the most significantly enriched gene set and one we see consistently enriched in all our zebrafish models of EOfAD, also increased in significance. Three of the pathways (*Viral myocarditis*, *Cytokine-Cytokine Receptor Interaction* and *Galactose metabolism*) that were initially classified as significant, but were on the verge of non-significance, were shown to lose significance after DAR-weighting.

DISCUSSION

High throughput RNA-sequencing is currently the most advanced technology for gene expression analysis. However, experimental artifacts are widespread and can lead to misinterpretation of results when not appropriately addressed. There is considerable evidence for transcript length and guanine-cytosine (GC) content as complexities of RNA-seq contributing to bias in differential expression and functional gene set analyses (40,41). Techniques such as *cqn* (42), *EDASeq* (43) and *GOSeq* (44) have been developed to assist in the correction of these issues. As we uncover new challenges in the interpretation of RNA-seq data, we require the continuous development of procedures to identify and alleviate such situations.

White et al. (10) recently proposed that CC-DEGs might be mistakenly interpreted as differential expression caused by a mutation of interest when, in reality, they are an allele-specific expression (eQTL)-dependent effect. The authors highlighted that eQTLs can be revealed in contexts of differential allelic representation between sample groups subject to comparison; a situation commonly encountered when using non-isogenic strains of model organisms. White et al. used inbred descendants of a cross between two discrete isogenic wild-type zebrafish strains and identified chromosomal sub-regions of homozygosity corresponding to one or other of the original strains. When gene expression data from these individuals was grouped according to such regional homozygosity, and then compared according to its derivation from the original strains, evidence was revealed for an eQTL basis of differential expression. While the analysis was appropriately designed to address the eQTL issue, the regions of chromosomal homozygosity analysed were relatively small and lacked the genetic context that would commonly be encountered in RNA-seq experiments investigating the effects of mutations. Generally, homozygosity for a mutant or wild-type sequence is the criterion for sorting individual samples into comparison groups, resulting in samples homozygous for a large portion of a mutation-bearing chromosome near the site of the mutation. The majority of the 82 homozygosity regions analysed by White et al. were limited to 3Mbp or less, which accounts for ~6% of the average length of zebrafish chromosomes (GRCz11, Ensembl version 101). In the context of a zebrafish transcriptome comparison such as described in Figure 1, recombination will often alter the size of the shared homozygous region to a length shorter than an entire chromosome. However, the region of homozygosity around the mutation locus may remain larger than 3 Mbp, particularly as recombination is suppressed during meiosis in male zebrafish (29).

Our discovery of clustering of, apparently, non-eQTL DE genes around heterozygous *psen1* mutations is consistent with Fisher's assertion that functional relationships between genes creates selective pressure driving the chromosomal co-location of advantageous allelic pairs during evolution to allow linkage disequilibrium. However, the differential representation of eQTLs in groups under comparison is also a parsimonious explanation for CC-DEG. In an attempt to resolve these two explanations, we developed a means to measure localised between-group differential allelic representation in RNA-seq datasets. Our DAR metric allows for refined identification of genomic regions that may be prone to eQTL expression bias in transcriptome studies. Such bias was clearly highlighted in our analysis of the mouse *APOE* dataset (Figure 4) where we found extreme clustering of genes around the *APOE* locus, apparently differentially expressed due to differential representation of eQTLs in the comparison groups. By the same measure, the CC-DEGs associated with heterozygous, dominant *psen1* mutations are unlikely to be eQTLs. Intriguingly, the CC-DEGs differ for the functionally very distinct *psen1*^{T428del} EOfAD-like and *psen1*^{W233fs} fAI-like mutations, and the degree of clustering of the CC-DEG is consistent with the selective disadvantage that each mutation would cause. In particular, the fAI-like mutation, which we have shown to affect significantly Notch signalling (12) (a mechanism of central importance to cell differentiation and embryo development, reviewed in (45)) shows differentially expressed genes scattered across the entire length of Chromosome 17, in contrast to the fewer, tightly clustered DE genes of the EOfAD-like mutation that, in humans, causes an adult-onset disease. In male zebrafish, the entirety of chromosome 17 shows a recombination distance of less than 35 cM (~118 cM in females) (29) meaning that alleles of any pair of genes on this chromosome can exist in a state of linkage disequilibrium and so might be under positive selection for chromosomal co-location if co-involved in Notch signalling-dependent processes.

Our limited investigation of chromosome evolution in zebrafish revealed some evidence consistent with the idea that selection for linkage disequilibrium may be influential during genome evolution. A number of CC-DEGs due to mutation of *psen1* were found to reside on separate chromosomes in an ancestral chromosome arrangement. However, a far more expansive and detailed investigation of this issue would be necessary for confidence that generation of linkage disequilibrium is an important positive selective pressure on chromosomal rearrangements. If such selective pressure is important, then why have CC-DEGs not been observed more widely in transcriptome studies? It is notable that the only situations where our laboratory has observed distinct CC-DEG patterns are in well-controlled comparisons of single gene mutations in 2 dpf embryos (8), 7 dpf larvae, or in young (~6-month-old zebrafish & ~3-month-old mouse) whole brains/cortices. If genes are connected in complex, homeostatic networks, then disturbance of only one network node would cause great expression changes in only the most closely co-functional nodes, much as pulling on one point in a spider's web causes greatest displacement for the most closely associated threads. In contrast, influences that disturb multiple nodes simultaneously (such as environmental change) would, like wind over a spider's web, give much more widespread network expression change. The magnitude of a mutation's phenotypic effect may also be a factor. In our analysis of human *APOE* alleles in humanised mouse cortices, the *APOE4* allele may differ in its activity from *APOE3* more significantly than the *APOE2* allele differs from *APOE3*, contributing to the greater number and broader chromosomal distribution of genes that are DE in the *APOE4/4 vs. APOE3/3* comparison than the *APOE2/2* vs. *APOE3/3* comparison.

Identification of possible eQTLs in transcriptome comparisons is important if their exclusion from analyses can improve the identification of affected gene sets. Our investigation into utilising the DAR metric to exclude possible eQTLs in the *naglu*$^{A603Efs/A603Efs}$ vs. *naglu*$^{+/+}$ 7 dpf sibling larval transcriptome dataset revealed a number of interesting phenomena. Both ROAST and GSEA methods returned the expected *Glycosaminoglycan Degradation* gene set as significantly enriched at various stages of the thresholding procedure. ROAST, however, appeared far more robust to the removal of genes in comparison to GSEA. ROAST consistently reported *Glycosaminoglycan Degradation* as significantly enriched at most diversity thresholds, together with two other gene sets known to be implicated with MPS disease (*Other Glycan Degradation*, and *Lysosome*). These gene sets also remained consistent in their respective rankings as defined by their *p*-values. GSEA, on the other hand, displayed considerably more variability in its outcomes. Aside from the four gene sets that remained significant at all thresholds using GSEA (*Lysosome*, *Focal Adhesion*, *Other Glycan Degradation*, *ECM Receptor Interaction*), we noticed that excluding genes had a clear impact on the relative rankings of gene sets. We also observed a general increase in the number of significant gene sets as more genes from high DAR regions were excluded, suggesting that the removal of eQTL-biased genes may improve the sensitivity of GSEA. *Glycosaminoglycan Degradation* only achieved significance at a diversity metric threshold of 0.6, possibly meaning that a threshold near this value is most suitable for removing bias from the dataset. However, choosing an optimal DAR threshold in this manner is arbitrary and therefore we considered alternative solutions, i.e. weighting-based approaches.

The disparity in the effects of thresholding between ROAST and GSEA is considerable but not surprising. Indeed, GSEA is well-known to be prone to bias, such as the false positive calls attributed to sample-specific gene length bias by Mandelboum et al. (40). GSEA requires a gene-level ranked list as determined by differential expression analysis to calculate downstream gene set scores. It is also a competitive test, meaning that the magnitude of enrichment of each gene set affects the significance assessment of other gene sets. Therefore, biased genes influence the ranking of genes showing mutation-driven differential expression, and gene sets containing the biased genes affect the significance of gene sets without them. In comparison, ROAST is a non-competitive test that works directly from the gene expression values, without requiring intermediate calculations of gene-level scores. Therefore, in the context of ROAST, biased genes will only impact the outcomes for a particular gene set if they are a member of that gene set. This provides a likely explanation for ROAST appearing more robust to experimental biases than GSEA.

In terms of good bioinformatics practise, excluding information is generally undesirable in contrast to applying statistical methods such as gene-level weightings to address differences in confidence. The DAR-based weighting approach presented compelling findings because *Glycosaminoglycan Degradation*, a pathway we expect to see enriched in studies of MPS-IIIB, only achieved significance after the ranked list was weighted with DAR, along with all other gene sets, which showed consistent significance across all thresholds. While the gene sets we expected to see enriched in the *sorl1*$^{R122Pfs/+}$ vs. *sorl1*$^{+/+}$ comparison were already statistically significant before weighting, the increase in significance after weighting further reinforces the validity of this method. This finding provides confidence that using DAR to weight ranking statistics and adjust gene rankings can be a suitable approach to improving GSEA in situations where eQTLs are present. While this approach is not applicable for ROAST, as it does not require intermediate gene-level rankings, ROAST appears far more robust to eQTL artefacts and may simply be a preferable choice in many situations.

MATERIALS AND METHODS

Availability of code

All code to reproduce the described analyses can be found below.

- *APOE* mouse dataset: https://github.com/baerlachlan/211001_APOE_Mm
- *psen1* mutant zebrafish dataset: https://github.com/baerlachlan/210408_psen1_fADfAI
- *naglu* larval zebrafish dataset: https://github.com/baerlachlan/211130_Q96K97del_A603fs
- *sorl1* dataset: https://github.com/baerlachlan/210216_sorl1_snv

Data description and generation

Two of the RNA-seq datasets involved in this analysis have already been described in previously published analyses. For information about animal ethics, genome editing, breeding strategy and RNA-seq data generation for each dataset please refer to their respective publications. The *psen1* mutant zebrafish dataset is described in (12) and is available in the GEO database (GSE164466). The *APOE* mouse dataset is described in (26) and is available at the AD Knowledge Portal (accession number syn20808171, https://adknowledgeportal.synapse.org/). The $naglu^{A603Efs/A603Efs}$ vs. $naglu^{+/+}$ 7 dpf sibling larval RNA-seq dataset is a currently unpublished dataset from our group. However, it is available from the GEO database (GSE217196). It is comprised of RNA-seq libraries derived from whole zebrafish 7 dpf sibling larvae either heterozygous for the EOfAD-like mutation $psen^{Q96_K97del}$, homozygous for the MPSIIIB-like mutation $naglu^{A603fs}$, or wild-type.

Differential expression analysis

Paired-end raw RNA-seq reads in FASTQ format were assessed for quality with *FastQC* (46) and *ngsReports* (47). Adapters, poly-G regions and bases with a PHRED score below 20 were trimmed from raw RNA-seq data using *fastp* (v0.23.2) (48). Any resulting reads were removed if they were shorter than 35bp after trimming. Trimmed reads were aligned to the Ensembl genome (GRCz11 release 101 for zebrafish, GRCm39 release 104 for mouse) using *STAR* (v2.7.7a) (49) in two-pass mode to achieve better alignment around novel splice junctions. Aligned reads were counted at the gene-level using *featureCounts* from the *Subread* package (v2.0.1) (50) if they were unique and mapped to strictly exonic regions. The analysis of count data was then performed in R (25). Genes were retained for downstream analysis if they had a minimum of 1 count per million in at least the number of samples equalling the smallest genotype group. Differences in library sizes was accounted for by normalisation using the trimmed mean of M-values method (51). Conditional Quantile Normalisation (*CQN*, (42)) was performed prior to differential expression analysis if guanine-cytosine content or length bias was detected in the dataset. Differential expression analysis was performed by fitting a negative binomial generalized log-linear model to gene-level counts and likelihood ratio testing using *edgeR* (52). The design matrix was specified with an intercept of 0 and coefficients corresponding to genotypes. Contrasts were defined and the null hypothesis was tested that the specified contrasts of the coefficients are equal to zero. Genes were considered to be differentially expressed in each comparison if the false discovery rate (FDR)-adjusted *p*-value was less than 0.05.

Variant calling and the DAR metric

Variants were determined based on the Genome Analysis Toolkit (GATK) best practices workflow for RNA-seq short variant discover (SNPs + Indels) (24). Unique Molecular Identifiers (UMIs) were incorporated into FASTQ headers with *fastp* (v0.20.1) where this information was available to improve deduplication methods. Adapters were trimmed from raw reads using *fastp* (v0.20.1) and resulting trimmed reads were discarded if they were shorter than 35bp. Trimmed reads were aligned to the Ensembl genome (GRCz11 release 101 for zebrafish, GRCm39 release 104 for mouse) using *STAR* (v2.7.7a) in two-pass mode. Read groups were assigned to the aligned reads based on sample and sequencing lane with the GATK's (v4.2) *AddOrReplaceReadGroups* to allow accurate correction of technical batch effects. Reads were grouped based on their UMI and mapping coordinates by tagging BAM files with *umitools group* (v1.1.1) (53). Reads were deduplicated with Picard *MarkDuplicates* (GATK v4.2) and those that spanned splicing regions were split using *SplitNCigar* reads (GATK v4.2). A set of known variants was obtained from the Ensembl database for the corresponding release chosen for the Ensembl genome during alignment. Base Quality Score Recalibration (BQSR) was performed with *BaseRecalibrator* and *ApplyBQSR* (GATK v4.2) to detect systematic errors in sequencing and adjust base quality scores accordingly. Variants were called in GVCF mode with *HaplotypeCaller* and joint-genotyped with *GenotypeGVCFs* (GATK v4.2). The resulting calls were filtered based on the GATK's recommended specific hard filters (phred-scaled *p*-value using Fisher's exact test for strand bias (FS) > 60, variant confidence/quality by depth (QD) > 2, root mean square mapping

quality (MQ) < 40, strand odds ratio (SOR) > 4). Lastly, variants were selected for only single nucleotide polymorphisms (SNPs) using the GATK (v4.2) *SelectVariants* tool.

Calculation of the DAR metric was performed in R. Genotype calls were assessed at each SNP locus that contained a non-reference allele in at least one sample. Samples were grouped according to mutant genotype and SNPs were excluded from the analysis if the number of samples with missing or filtered genotype calls was greater than 50% within each group. Allele-level counts for each SNP were normalised as a proportion of the total number of reported alleles for each group (2n alleles reported for n genotypes called). The Euclidean distance formula was then implemented using the base R *dist* function on the normalised allele counts between the sample groups contrasted during differential expression analysis. The calculated distance was divided by its maximum possible value of $\sqrt{2}$ to achieve the DAR metric ranging between 0 and 1.

Functional gene set analysis

Because not all genes contained SNP information to assign a DAR value, we smoothed the DAR metric using an elastic window approach of n=11 SNPs, allowing for a broader representation of DAR. Genes were assigned a DAR value based on the average of any regions the gene was located in. Prior to functional enrichment analysis, gene exclusion thresholds were established at DAR values ranging from 0.1 to 1.0 at intervals of 0.1. Here, any genes that had an assigned DAR value above the threshold were excluded from the analysis. Two separate methods were chosen to perform functional gene set analysis: ROAST (33), implemented with the *fry* function from the *limma* package (34), and GSEA (35), implemented with the *fgseaMultilevel* function from the *fgsea* package (36), Both methods were performed at each of the 10 established DAR thresholds.

The Synteny Database

Analysis of conserved synteny between zebrafish and medaka was performed using the dot plot feature of the Synteny Database (32). The *Danio rerio* Ensembl version 71 genome was selected as the source genome and the *Oryzias latipes* Ensembl version 71 genome was selected as the outgroup genome. Note that Ensembl version 71 was selected as it was the most recent available release on the Synteny Database. Plots were generated by inputting the zebrafish Ensembl gene identifiers, and selecting the corresponding chromosome to display along the x-axis.

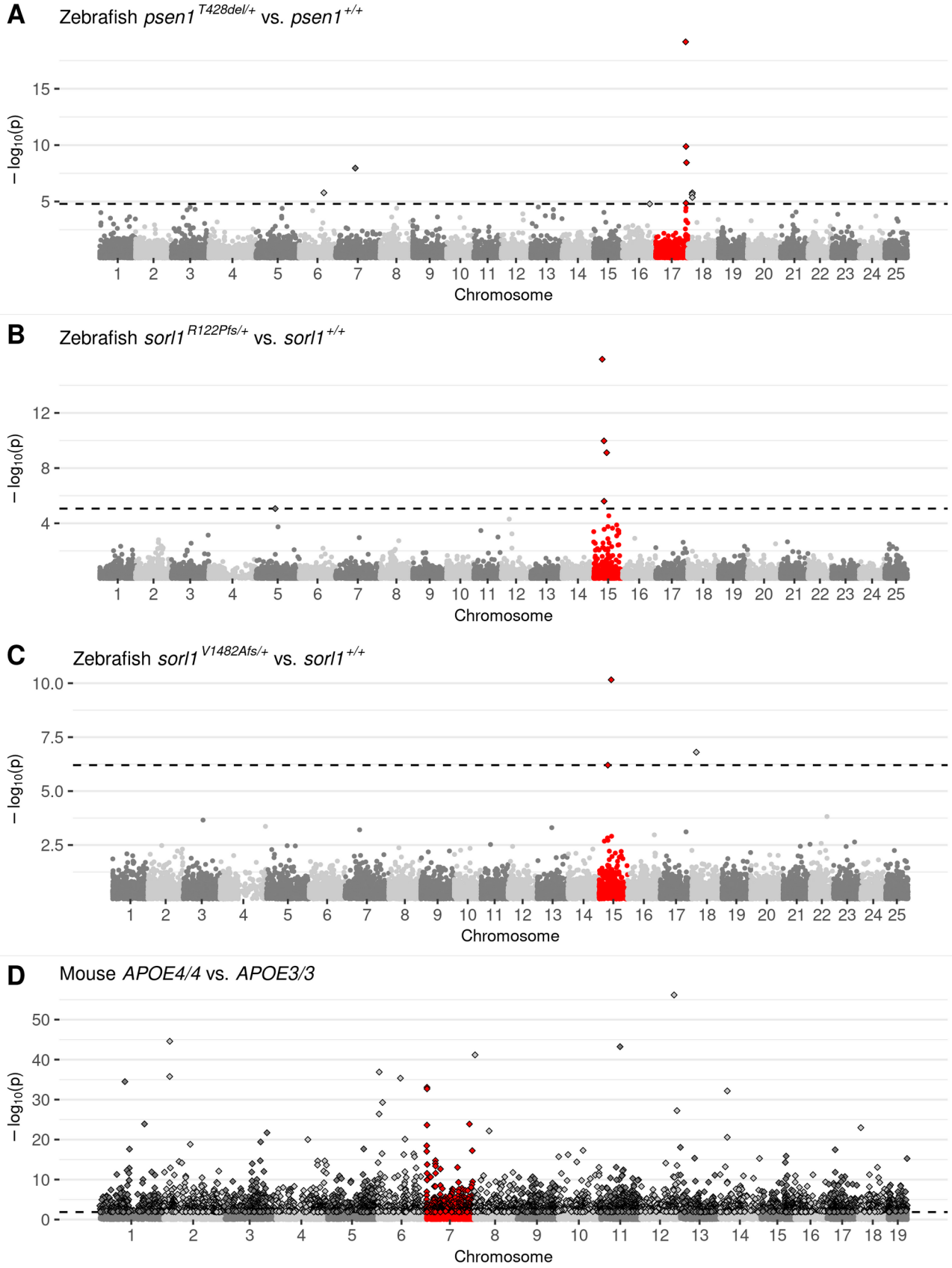
ACKNOWLEDGEMENTS

The authors wish to thank Julian Catchen for assistance with the use of the Synteny Database. The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Support for these studies was provided by the NIH RF1 AG051504 and P01 AG030128. We thank Drs. Patrick Sullivan and Nobuyo Maeda for generating human APOE targeted replacement mice and providing access through Taconic. The zebrafish transcriptome data was produced with assistance from the South Australian Genomics Centre. This work was also supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide.

FUNDING

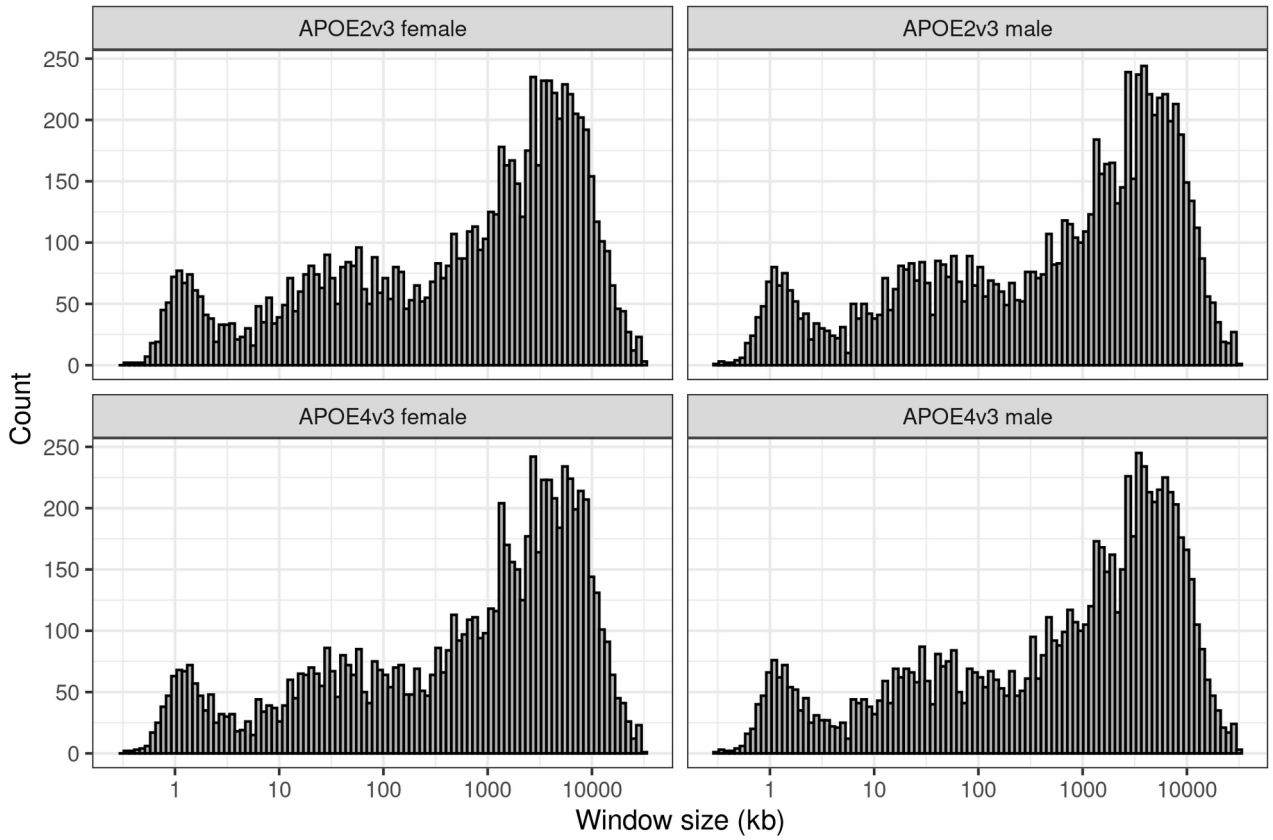
Production of some of the transcriptome datasets analysed in this manuscript was supported by funds from the Carthew Family Charity Trust. LB is supported by the Australian Government Research Training Program Scholarship. KB is supported by a Race Against Dementia – Dementia Australia Research Foundation Post-doctoral Fellowship and funds from Flinders University. DA, ML, and JP are academic employees of their respective universities. SP is a postdoctoral fellow supported by the Telethon Kids Institute.

SUPPLEMENTARY INFORMATION

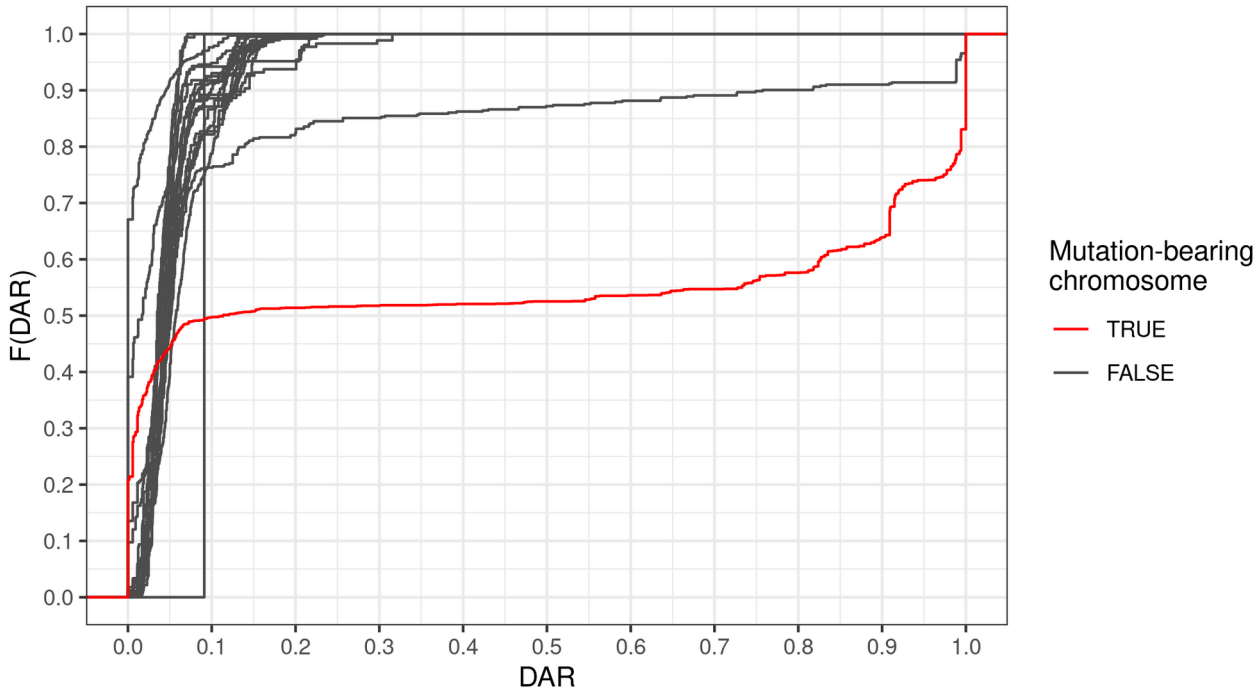


Supplementary Figure 1. Manhattan plots highlighting CC-DEG from differential expression testing in heterozygous vs. wild-type zebrafish siblings (6-month-old brains) and between homozygous mouse mutants (3-month-old male cortices). **A)** zebrafish EOfAD-like *psen1*$^{T428del/+}$ vs. *psen1*$^{+/+}$ **B)** zebrafish EOfAD-like *sorl1* $^{R122Pfs/+}$ vs. *sorl1*$^{+/+}$ **C)** zebrafish EOfAD-like *sorl1*$^{V1482Afs/+}$ vs. *sorl1*$^{+/+}$ **D)** mouse *APOE4/4* vs. *APOE3/3*. Non-random accumulation of DEGs was tested using a Bonferroni-adjusted Fisher's exact test *p*-value for

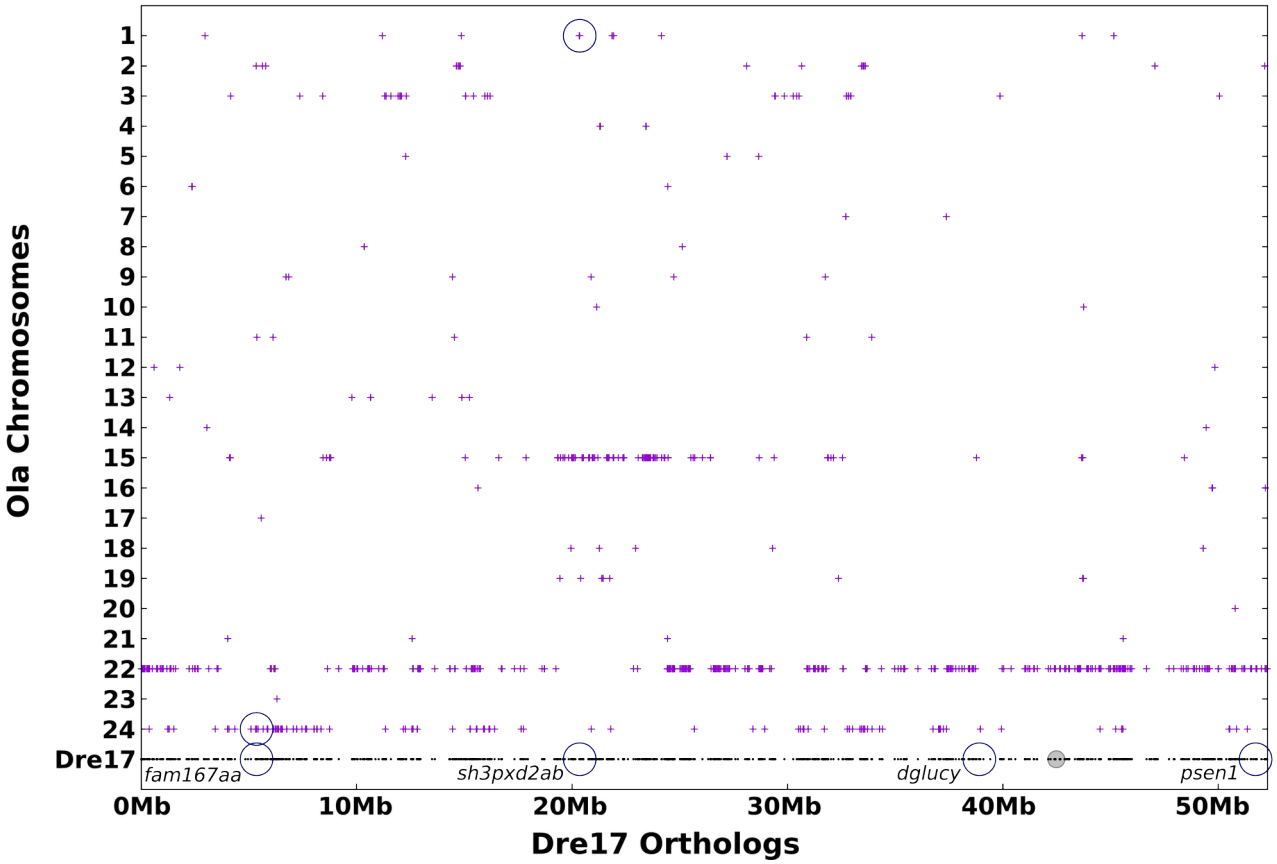
enrichment of DE genes on the mutant chromosome, A) *psen1*$^{T428del/+}$ vs. *psen1*$^{+/+}$: p = 8.63e-3, B) *sorl1*$^{R122Pfs/+}$ vs. *sorl1*$^{+/+}$: p = 1.92e-4, C) *sorl1*$^{V1482Afs/+}$ vs. *sorl1*$^{+/+}$: p = 9.24e-2, D) *APOE4/4* vs. *APOE3/3*: p = 1.00. Genes are plotted along the x-axis based on their chromosomal position in alternating shades of grey for visual distinction between chromosomes. Genes on the chromosome containing the mutation are highlighted in red. The raw p-values are plotted along the y-axis at the $-\log_{10}$ scale such that the most significant genes exist at the top of the plot. The cut-off for gene differential expression (FDR-adjusted p-value < 0.05) is indicated by a dashed horizontal line. Genes classified as differentially expressed under this criterion are represented as diamonds with a black outline.



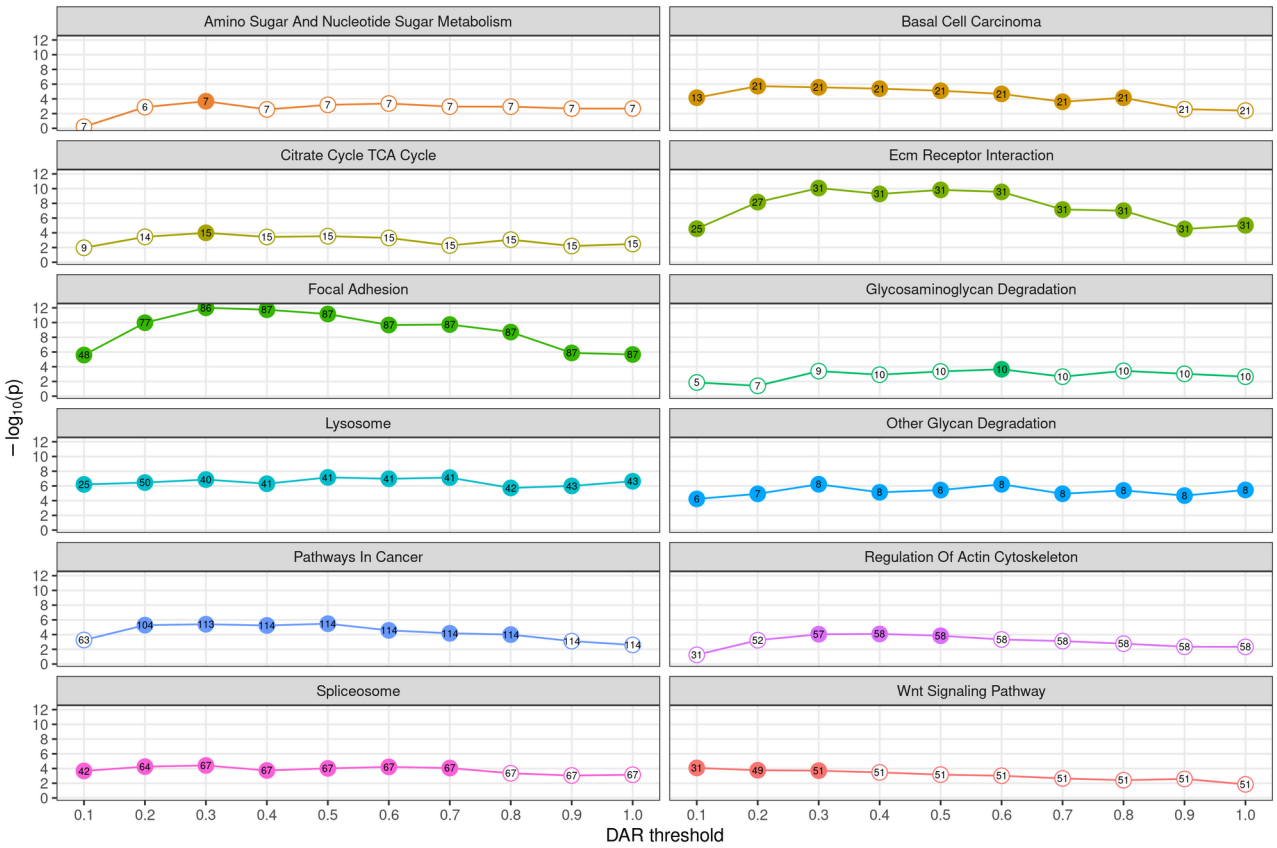
Supplementary Figure 2. Elastic window sizes of the mouse APOE dataset summarised in 100 bins of equal size. Window sizes are determined based on the genomic distance between 11 subsequent SNPs.



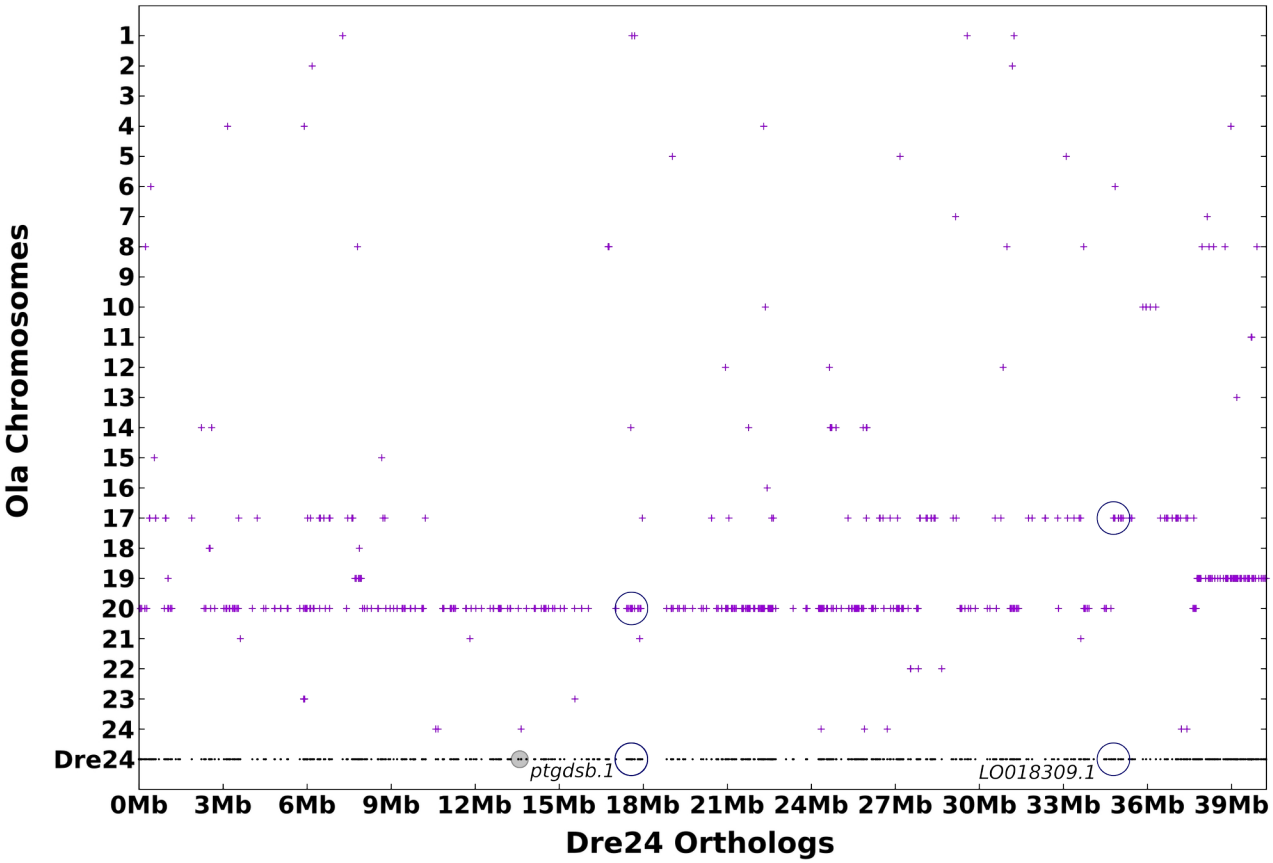
Supplementary Figure 3. Cumulative distribution of DAR by chromosome between male *APOE2/2* and *APOE3/3* mouse cortices (3 months). The mutant chromosome exhibits the most regions of high DAR.



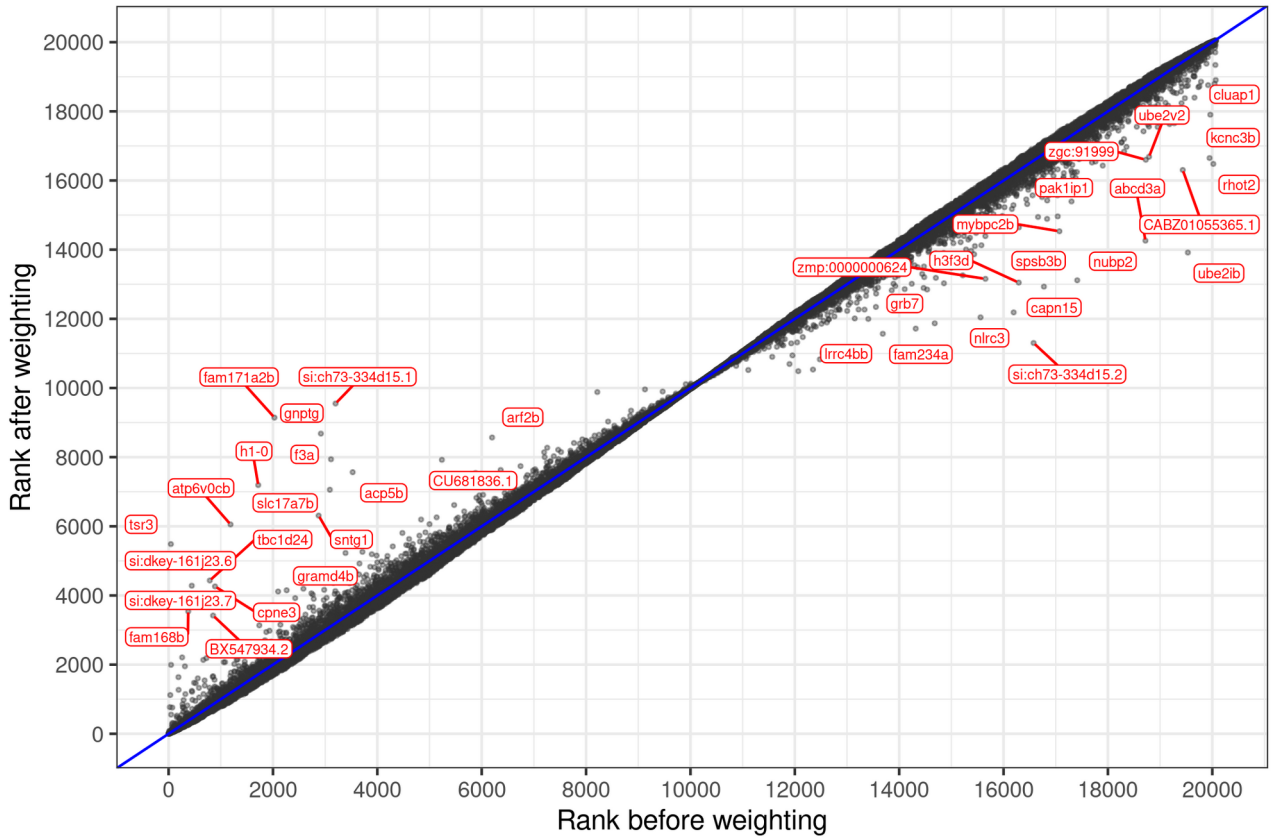
Supplementary Figure 4. Dotplot displaying the chromosomal location of medaka (Ola) genes predicted by the Synteny Database (32) to be orthologues of zebrafish (Dre) Chromosome 17 DE genes from the fAI-relevant, *psen1*$^{W233fs/+}$ vs. *psen1*$^{+/+}$ comparison. Genes are plotted along the x-axis based on their position on Chromosome 17 in zebrafish, while the y-axis indicates the chromosome they are located on in medaka. The four DE genes predicted to have orthologues are circled in blue. The two genes without a secondary blue circle plotted along the y-axis (*dglucy* and *psen1*), were predicted by the Synteny Database to exist on alternate scaffolds of medaka Ensembl version 71, which are not plotted.



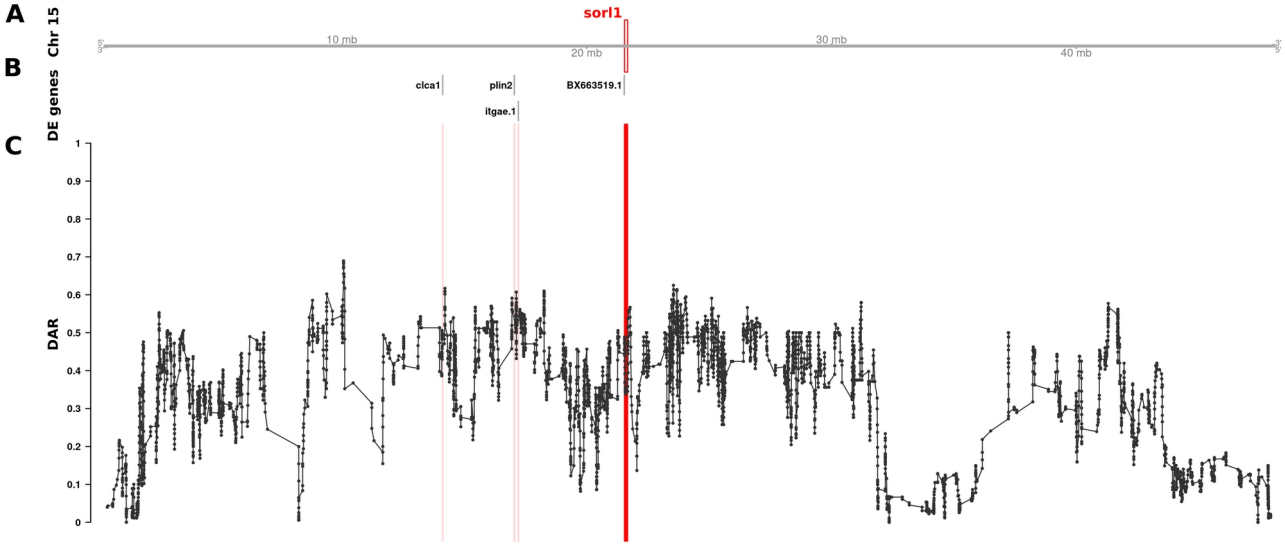
Supplementary Figure 5. The effects of gene exclusion by DAR threshold on KEGG gene set *p*-values from GSEA testing. The gene sets displayed are those that achieved significance for at least one DAR threshold value. *p*-values are plotted on a -$\log_{10}$ scale along the y-axis such that the most significant results exist at the top of each graph. Each position along the x-axis represents a different DAR gene exclusion threshold. Dots on the graph filled with colour indicate that the gene set was classified as significantly enriched (FDR-adjusted *p*-value < 0.05). The number inside a dot corresponds to the number of leading-edge genes that contributed to the respective gene set's enrichment score.



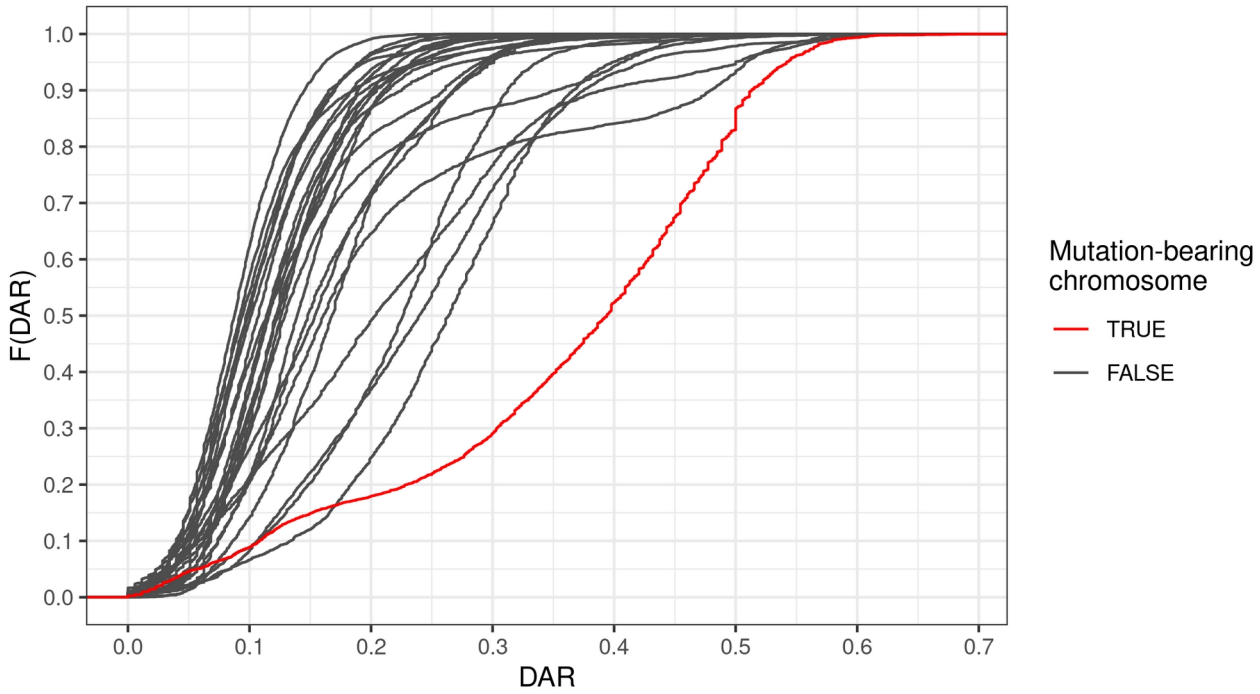
Supplementary Figure 6. Dotplot displaying the chromosomal location of medaka (Ola) genes predicted by the Synteny Database (32) to be orthologues of zebrafish (Dre) Chromosome 24 DE genes from the *naglu*$^{A603Efs/A603Efs}$ vs. *naglu*$^{+/+}$ 7 dpf sibling larval RNA-seq dataset comparison at a DAR threshold of 0.3. Genes are plotted along the x-axis based on their position on Chromosome 24 in zebrafish, while the y-axis indicates the chromosome they are located on in medaka. The two DE genes predicted to have orthologues are circled in blue.



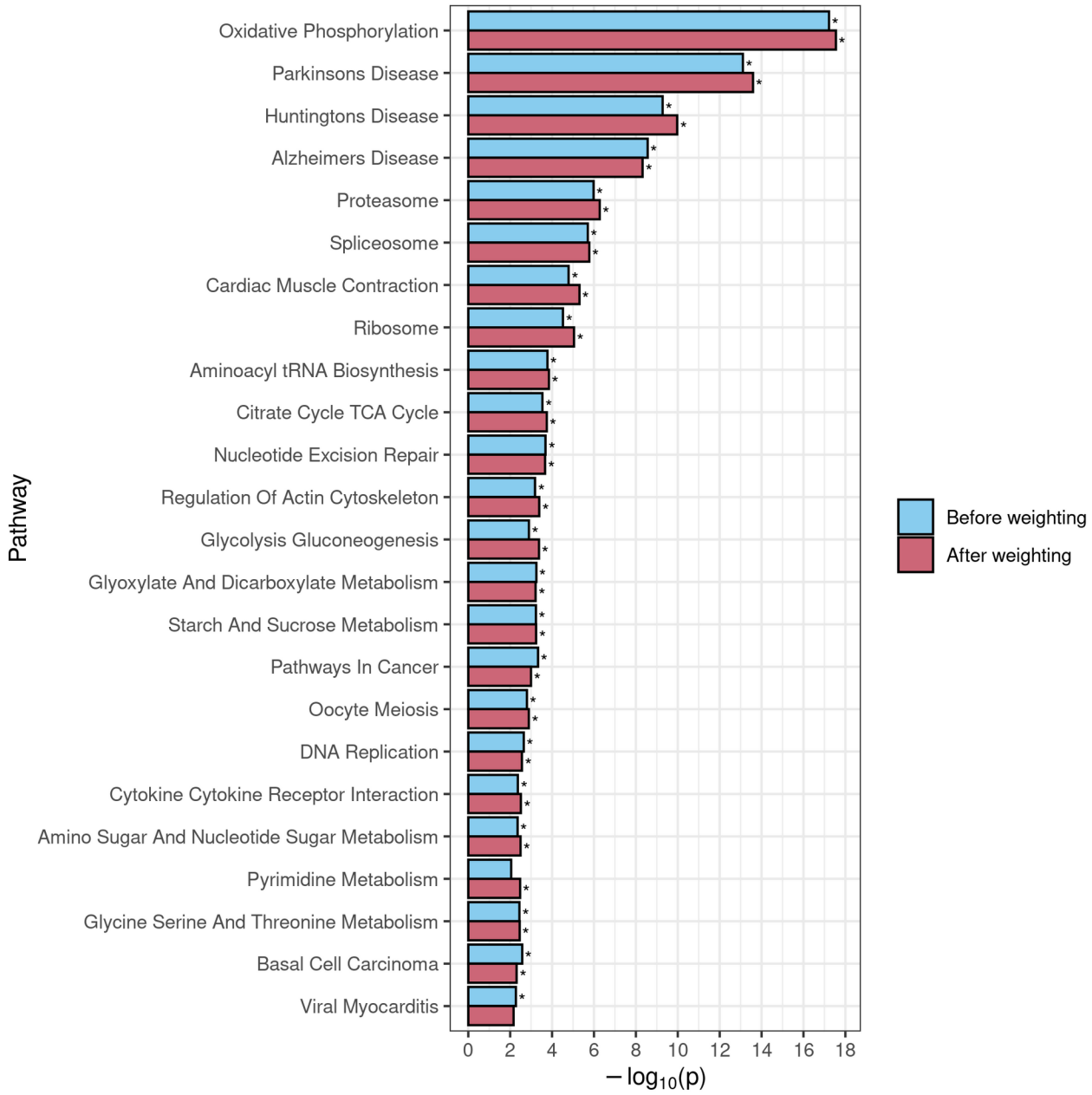
Supplementary Figure 7. The effects of using DAR to weight gene-level rankings in the MPS-IIIB-relevant transcriptome comparison of *naglu*$^{A603Efs/A603Efs}$ vs. *naglu*$^{+/+}$ 7 dpf sibling larvae. Genes that are plotted close to the blue diagonal line are least impacted by the weighting method. Genes that were most substantially affected (rank change > 2000) are labelled in red.



Supplementary Figure 8. The relationship between DAR and DE genes along Chromosome 15 in the comparison of *sorl1*$^{R122Pfs/+}$ vs. *sorl1*$^{+/+}$ sibling zebrafish brains. The plot contains four sets of information represented by separate tracks. Track A represents the axis of Chromosome 15. The position of the *sorl1* gene is marked and labelled in bold red. Track B displays differentially expressed genes according to their positions along the chromosome. Track C shows the trend in DAR as a connected scatterplot with each point representing the DAR value at a single nucleotide variant position (elastic sliding window, n = 11 variants). Positions of the DE genes shown in track B are indicated by light red lines.



Supplementary Figure 9. Cumulative distribution of DAR by chromosome between *sorl1*$^{R122Pfs/+}$ and *sorl1*$^{+/+}$ sibling zebrafish. The mutant chromosome exhibits the most regions of high DAR.



Supplementary Figure 10. Comparison of GSEA results for KEGG gene sets that achieved significance before and/or after using DAR to weight the gene-level ranking statistic in the *sorl1*$^{R122Pfs/+}$ vs. *sorl1*$^{+/+}$ siblings dataset. An asterisk is used to denote that the pathway was determined to be significantly enriched.

REFERENCES

1. Oudelaar AM, Higgs DR. The relationship between genome structure and function. Nat Rev Genet. 2021 Mar;22(3):154–68.
2. Morgan TH. Random Segregation Versus Coupling in Mendelian Inheritance. Science. 1911 Sep 22;34(873):384–384.
3. Fisher RA. The genetical theory of natural selection [Internet]. Oxford: Clarendon Press; 1930 [cited 2022 Oct 25]. Available from: https://www.biodiversitylibrary.org/bibliography/27468
4. Salih H, Adelson DL. QTL global meta-analysis: are trait determining genes clustered? BMC Genomics. 2009;10(1):184.
5. Tiirikka T, Siermala M, Vihinen M. Clustering of gene ontology terms in genomes. Gene. 2014 Oct;550(2):155–64.
6. Dooley CM, Wali N, Sealy IM, White RJ, Stemple DL, Collins JE, et al. The gene regulatory basis of genetic compensation during neural crest induction. Mullins MC, editor. PLOS Genet. 2019 Jun 14;15(6):e1008213.
7. Nuvolone M, Hermann M, Sorce S, Russo G, Tiberi C, Schwarz P, et al. Strictly co-isogenic C57BL/6J-Prnp-/- mice: A rigorous resource for prion science. J Exp Med. 2016 Mar 7;213(3):313–27.
8. Barthelson K, Baer L, Dong Y, Hand M, Pujic Z, Newman M, et al. Zebrafish Chromosome 14 Gene Differential Expression in the fmr1hu2787 Model of Fragile X Syndrome. Front Genet. 2021 May 31;12:625466.
9. Kong SW, Sahin M, Collins CD, Wertz MH, Campbell MG, Leech JD, et al. Divergent dysregulation of gene expression in murine models of fragile X syndrome and tuberous sclerosis. Mol Autism. 2014 Dec;5(1):16.
10. White RJ, Mackay E, Wilson SW, Busch-Nentwich EM. Allele-specific gene expression can underlie altered transcript abundance in zebrafish mutants. eLife. 2022 Feb 17;11:e72825.
11. Ryman DC, Acosta-Baena N, Aisen PS, Bird T, Danek A, Fox NC, et al. Symptom onset in autosomal dominant Alzheimer disease: a systematic review and meta-analysis. Neurology. 2014 Jul 15;83(3):253–60.
12. Barthelson K, Dong Y, Newman M, Lardelli M. PRESENILIN 1 Mutations Causing Early-Onset Familial Alzheimer's Disease or Familial Acne Inversa Differ in Their Effects on Genes Facilitating Energy Metabolism and Signal Transduction. J Alzheimers Dis. 2021 Jun 29;82(1):327–47.
13. Wang B, Yang W, Wen W, Sun J, Su B, Liu B, et al. Gamma-secretase gene mutations in familial acne inversa. Science. 2010 Nov 19;330(6007):1065.
14. Duchatelet S, Miskinyte S, Delage M, Ungeheuer MN, Lam T, Benhadou F, et al. Low Prevalence of GSC Gene Mutations in a Large Cohort of Predominantly Caucasian Patients with Hidradenitis Suppurativa. J Invest Dermatol. 2020 Oct;140(10):2085-2088.e14.
15. Hin N, Newman M, Pederson S, Lardelli M. Iron Responsive Element-Mediated Responses to Iron Dyshomeostasis in Alzheimer's Disease. J Alzheimers Dis. 2021 Dec 7;84(4):1597–630.
16. Barthelson K, Pederson SM, Newman M, Jiang H, Lardelli M. In-Frame and Frameshift Mutations in Zebrafish Presenilin 2 Affect Different Cellular Functions in Young Adult Brains. J Alzheimers Dis Rep. 2021 May 4;5(1):395–404.
17. Barthelson K, Newman M, Lardelli M. Brain transcriptomes of zebrafish and mouse Alzheimer's disease knock-in models imply early disrupted energy metabolism. Dis Model Mech. 2022 Jan 1;15(1):dmm049187.
18. Barthelson K, Pederson SM, Newman M, Lardelli M. Brain Transcriptome Analysis of a Protein-Truncating Mutation in Sortilin-Related Receptor 1 Associated With Early-Onset Familial Alzheimer's Disease Indicates Early Effects on Mitochondrial and Ribosome Function. J Alzheimers Dis. 2021 Feb 2;79(3):1105–19.
19. Jiang H, Pederson SM, Newman M, Dong Y, Barthelson K, Lardelli M. Transcriptome analysis indicates dominant effects on ribosome and mitochondrial function of a premature termination codon mutation in the zebrafish gene psen2. Key B, editor. PLOS ONE. 2020 Jul 13;15(7):e0232559.
20. Hin N, Newman M, Kaslin J, Douek AM, Lumsden A, Nik SHM, et al. Accelerated brain aging towards transcriptional inversion in a zebrafish model of the K115fs mutation of human PSEN2. Paisan-Ruiz C, editor. PLOS ONE. 2020 Jan 24;15(1):e0227258.

21. Barthelson K, Pederson SM, Newman M, Lardelli M. Brain transcriptome analysis reveals subtle effects on mitochondrial function and iron homeostasis of mutations in the SORL1 gene implicated in early onset familial Alzheimer's disease. Mol Brain. 2020 Dec;13(1):142.
22. Newman M, Hin N, Pederson S, Lardelli M. Brain transcriptome analysis of a familial Alzheimer's disease-like mutation in the zebrafish presenilin 1 gene implies effects on energy production. Mol Brain. 2019 Dec;12(1):43.
23. Zhao N, Ren Y, Yamazaki Y, Qiao W, Li F, Felton LM, et al. Alzheimer's Risk Factors Age, APOE Genotype, and Sex Drive Distinct Molecular Pathways. Neuron. 2020 Jun;106(5):727-742.e6.
24. Auwera G van der, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra. First edition. Sebastopol, CA: O'Reilly Media; 2020. 467 p.
25. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/
26. Sullivan PM, Mezdour H, Aratani Y, Knouff C, Najib J, Reddick RL, et al. Targeted Replacement of the Mouse Apolipoprotein E Gene with the Common Human APOE3 Allele Enhances Diet-induced Hypercholesterolemia and Atherosclerosis. J Biol Chem. 1997 Jul;272(29):17972–80.
27. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013 Oct;45(10):1238–43.
28. Jayne T, Newman M, Verdile G, Sutherland G, Münch G, Musgrave I, et al. Evidence For and Against a Pathogenic Role of Reduced γ-Secretase Activity in Familial Alzheimer's Disease. Murphy MP, editor. J Alzheimers Dis. 2016 May 23;52(3):781–99.
29. Singer A, Perlman H, Yan Y, Walker C, Corley-Smith G, Brandhorst B, et al. Sex-Specific Recombination Rates in Zebrafish (*Danio rerio*). Genetics. 2002 Feb 1;160(2):649–57.
30. Steinke D, Salzburger W, Meyer A. Novel Relationships Among Ten Fish Model Species Revealed Based on a Phylogenomic Analysis Using ESTs. J Mol Evol. 2006 Jun;62(6):772–84.
31. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, et al. The medaka draft genome and insights into vertebrate genome evolution. Nature. 2007 Jun;447(7145):714–9.
32. Catchen JM, Conery JS, Postlethwait JH. Automated identification of conserved synteny after whole-genome duplication. Genome Res. 2009 Aug;19(8):1497–505.
33. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010 Sep 1;26(17):2176–82.
34. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47–e47.
35. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005 Oct 25;102(43):15545–50.
36. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis [Internet]. Bioinformatics; 2016 Jun [cited 2022 Oct 25]. Available from: http://biorxiv.org/lookup/doi/10.1101/060012
37. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27–30.
38. Schmidtchen A, Greenberg D, Zhao HG, Li HH, Huang Y, Tieu P, et al. NAGLU Mutations Underlying Sanfilippo Syndrome Type B. Am J Hum Genet. 1998 Jan;62(1):64–9.
39. Parenti G, Andria G, Ballabio A. Lysosomal Storage Diseases: From Pathophysiology to Therapy. Annu Rev Med. 2015 Jan 14;66(1):471–86.
40. Mandelboum S, Manber Z, Elroy-Stein O, Elkon R. Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias. Huang S, editor. PLOS Biol. 2019 Nov 12;17(11):e3000481.
41. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010 Apr;464(7289):768–72.
42. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostat Oxf Engl. 2012 Apr;13(2):204–16.
43. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinformatics. 2011 Dec 17;12:480.
44. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14.

45. D'Assoro A, Leon-Ferre R, Braune EB, Lendahl U. Roles of Notch Signaling in the Tumor Microenvironment. Int J Mol Sci. 2022 Jun 2;23(11):6241.
46. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Internet]. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
47. Ward CM, To TH, Pederson SM. ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files. Hancock J, editor. Bioinformatics. 2020 Apr 15;36(8):2587–8.
48. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884–90.
49. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21.
50. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013 May 1;41(10):e108–e108.
51. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.
52. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.
53. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017 Mar;27(3):491–9.