Genetics: Early Online, published on November 26, 2013 as 10.1534/genetics.113.157800

Effect of Genetic Variation in a Drosophila Model of Diabetes-Associated Misfolded Human Proinsulin

Bin Z. He[*,1,2], Michael Z. Ludwig[*], Desiree A. Dickerson[*], Levi Barse[*], Bharath Arun[*], Bjarni J. Vilhjálmsson[‡‡], Soo-Young Park[†], Natalia A. Tamarina[†], Scott B. Selleck[§], Patricia J. Wittkopp[§§], Graeme I. Bell[†,‡], Martin Kreitman[*,2]

[*] Department of Ecology and Evolution, The University of Chicago, Chicago, IL 60637

[†] Department of Medicine, The University of Chicago, Chicago, IL 60637

[‡] Department of Human Genetics, The University of Chicago, Chicago, IL 60637

[‡‡] Department of Epidemiology, and Department of Biostatistics, Harvard School of Public Health, Harvard University, Boston, MA 02115

[§] Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, 16802

[§§] Department of Ecology and Evolutionary Biology, and Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI 48109

[1] Current Address: FAS Center for Systems Biology, Harvard University, 52 Oxford Street, Cambridge, MA 02138

Supporting information is available online at http://www.genetics.org/content/

Running title: Natural Variation in Fly Modifies Mutant-insulin Phenotype

Key words: mutant insulin, Drosophila, genome-wide association study, heparan sulfate proteoglycan, sulfateless

[2] Corresponding authors:

Martin Kreitman

Mailing address: Department of Ecology and Evolution, The University of Chicago, 1101 E 57th Street, Chicago, IL 60637-1573.

Phone: +1 773 702 1222. Fax: +1 773 702 9740.

Email: martinkreitman@gmail.com.

Bin Z. He

Mailing address: FAS Center for Systems Biology, Harvard University, 52 Oxford Street, Cambridge, MA 02138

Phone: +1 312 550 8421. Fax: +1 617 496 5425

Email: binhe@fas.harvard.edu

ABSTRACT

The identification and validation of gene-gene interactions is a major challenge in human studies. Here, we explore an approach for studying epistasis in humans using a *Drosophila melanogaster* model of neonatal diabetes mellitus. Expression of the mutant preproinsulin ($hINS^{C96Y}$) in the eye imaginal disc mimics the human disease: it activates conserved stress response pathways and leads to cell death (reduction in eye area). Dominant-acting variants in wild-derived inbred lines from the *Drosophila Genetics Reference Panel* produce a continuous, highly heritable distribution of eye degeneration phenotypes in a $hINS^{C96Y}$ background. A genome-wide association study (GWAS) in 154 sequenced lines identified a sharp peak on chromosome 3L, which mapped to a 400bp linkage block within an intron of the gene *sulfateless* (*sfl*). RNAi knock-down of *sfl* enhanced the eye degeneration phenotype in a mutant-hINS-dependent manner. RNAi against two additional genes in the heparan sulfate (HS) biosynthetic pathway (*ttv* and *botv*), in which *sfl* acts, also modified the eye phenotype in a $hINS^{C96Y}$-dependent manner, strongly suggesting a novel link between HS-modified proteins and cellular responses to misfolded proteins. Finally, we evaluated allele-specific expression difference between the two major *sfl*-intronic haplotypes in heterozygtes. The results showed significant heterogeneity in marker-associated gene expression, thereby leaving the causal mutation(s) and its mechanism unidentified. In conclusion, the ability to create a model of human genetic disease, map a QTL by GWAS to a specific gene, validate its contribution to disease with available genetic resources, and the potential to experimentally link the variant to a molecular mechanism, demonstrate the many

advantages *Drosophila* holds in determining the genetic underpinnings of human disease.

INTRODUCTION

Limitations imposed by human subject research can be overcome by investigating models of human disease in experimental organisms. *Drosophila* can provide genetic insights relevant to human biology and disease, owing to the conservation of fundamental cellular and developmental processes. We constructed a fly model of protein misfolding disease, by creating a transgene of a diabetes-causing, human mutant preproinsulin ($hINS^{C96Y}$) that could be expressed in the eye imaginal discs and other tissues (Park et al., 2013). This misfolded proinsulin protein causes the loss of insulin-secreting pancreatic beta cells and diabetes in humans and mice (Støy et al., 2007). When misexpressed in the *Drosophila* eye imaginal disc, it disrupts eye development, resulting in a reduced eye area in adult flies (Park et al., 2013).

In the accompanying paper (Park et al., 2013), we crossed the transgenic line bearing the mutant preproinsulin and an eye-specific Gal4 driver ($GMR>>hINS^{C96Y}$) with a subset of the lines from the Drosophila Genetics Reference Panel (DGRP). The F1 lines displayed a wide, nearly continuous, range of heritable eye degeneration phenotypes, suggesting a polygenic basis for this genetic background variation (Park et al., 2013). To investigate the genetic basis of this background variation, here we performed a genome-wide association study in a larger set of 154 DGRP lines.

Drosophila's many favorable attributes for mapping quantitative trait loci (QTL) — a high density of common variants, relatively little population subdivision, a decay of linkage disequilibrium (LD) over a scale of only 100's of bp, controlled crosses allowing repeat measurements, and excellent resources for confirmatory genetics — allowed us to identify a variant in the heparan sulfate (HS) biosynthesis pathway gene, *sulfateless*

(*sfl*), contributing to the eye degeneration phenotype, and then confirm a genetic interaction between mutant hINS and *sfl* by RNAi knockdown analysis. Two other genes in the HS biosynthetic pathway, *tout-velo* (*ttv*) and *brother of tout-velo* (*botv*), displayed a similar interaction upon genetic analysis, implicating HS-modified proteins, or proteoglycans (HSPG), in the response to misfolded proteins.

We then tested the hypothesis that the intronic *sfl* variants act by decreasing gene expression by measuring the relative expression level of each allele in 15 heterozygotes containing both alleles. The results are mixed, with seven crosses showing a difference that is consistent with the hypothesis; however, overall there is only modest correlation between the genotype and the expression level, which leaves the causal mutation(s) and its mechanism yet to be identified.

Although our model of neonatal diabetes in the fly — transgenic expression of a mutant disease-causing human insulin allele — is Mendelian, the severity of the disease trait is exquisitely sensitive to genetic background, and behaves as a complex trait. We discuss the prospects for modeling complex human disease in the fly with this general approach.

MATERIALS AND METHODS

Drosophila stocks and crosses

The {GMR-Gal4, UAS-hINS^{C96Y}} line was generated by crossing the GMR-Gal4 line (Stock #1104, Bloomington Stock Center) with the UAS-hINS^{C96Y} line (Park et al., 2013), and obtaining the recombinant 2nd chromosome, which was balanced over CyO. DGRP

lines were obtained from the Bloomington stock center. RNAi lines against *sfl* (GD5070), *ttv* (GD4871), *botv* (GD37186) were from the Vienna *Drosophila* RNAi center. Mutant lines for *ttv* (*ttv*^{681}) and *botv* (*botv*^{510}) were described previously (Ren et al., 2009).

Eye area measurement

All crosses were reared at 25□C. Total eye area was measured as described in (Park et al., 2013). At least 10 images (independent flies) passing the quality check were collected for each cross. Raw data is available in **Table S1**.

Principal Component Analysis

The whole-genome SNP dataset for the 154 DGRP lines used for GWAS (see **Table S2** for the list of line numbers) was downloaded from the DGRP website (http://dgrp.gnets.ncsu.edu/, freeze 1). To characterize population structure, 900K SNPs (after LD pruning using PLINK v1.07, with parameter --indep-pairwise 50 5 0.5) were used to identify the top 15 principal components (PCs) (SmartPCA software in EIGENSOFT v3.0, no outlier exclusion). We then estimated the correlation between the hINS^{C96Y} phenotype (line mean) and projection length in the direction of the top five principle components in each DGRP line to test whether population structure is a confounding source of association in GWAS.

Genome wide association using linear regression

The mean eye area of 154 DGRP lines crossed to the hINS^{C96Y} line was regressed on each SNP with a minor allele frequency (MAF) > 5% (PLINK 1.07, quantitative trait

mode). 1,616,121 autosomal and 256,948 SNPs on the X chromosome were tested. The F1 males inherited their X chromosome from the common transgene-containing strain. The identity by descent of this X chromosome allowed us to test whether the X-linked SNPs in the DGRP sample conformed to a null distribution assuming no association (although linkage is likely to cause deviation from this expectation). This was tested in quantile-quantile (Q-Q) plot analysis.

Association by mixed linear model to control for genetic relatedness

A Python implementation of EMMAX (Kang et al., 2010; Segura et al., 2012) was used to estimate the Genetic Related Matrix (GRM) using inverse variance weighted SNPs. The GRM is plotted using the pheatmap package in R to visualize any cryptic relatedness (Kolde, 2011). When performing mixed linear model regression, we used the GRM estimated from just the X-chromosome SNPs, for which the mixed model yields a narrow sense heritability of 0.83 (SNPs with MAF>0.05). By doing so, we increase our power to detect associations at loci on the other chromosomes, because those are not included in the GRM (Listgarten et al., 2012). The ~250K SNPs on the X chromosome are sufficient for inferring the population structure in the sample and thereby controlling population stratification. This is evident by the uniform p-value distribution in the Q-Q plots (**Fig. S4**). To assess the genome-wide significance threshold while accounting for both the relatedness structure in the data as well as the non-independence between SNPs due to LD, we performed a permutation procedure (details in **Text S1**).

Conditional analysis using sfl intronic SNPs as covariates

To identify possible secondary associations in *sfl* or elsewhere in the genome independent of the intronic QTL variants in *sfl*, we fit a linear model with the most significant variant, a 18 bp /4 bp insertion/deletion polymorphism, as a covariate. This analysis was performed either within the *sfl* locus or genome-wide. The p-values were corrected for multiple testing using Bonferroni's method.

Estimate proportion of variance explained by common SNPs

We first used GCTA (v1.0) to estimate the genetic relatedness matrix with all SNPs with minor allele frequency greater than 5% (--maf 0.05). We then used the restricted maximum likelihood method implemented in GCTA to estimate the quantity V_G / V_P (--reml), i.e. the narrow sense heritability.

Expression of sfl and CG32396

Expression profiles in adult tissues were assessed using data from FlyAtlas (Chintapalli et al., 2007) and modENCODE (Roy et al., 2010). To assay expression in the eye imaginal discs, we isolated total RNA from 10 pairs of discs from 3rd instar larvae. Individual larva was sexed and dissected in 1X phosphate buffer saline (PBS); the eye portions of the eye-antennal disc were collected and the isolated discs immediately dissolved in 300 µl Trizol (Invitrogen). Total RNA was extracted according to the manufacturer's instructions. cDNA libraries were constructed using (dT)20 primers after DNase I treatment (Invitrogen). Real time quantitative PCR was performed with primer pairs targeting either *sfl* or CG32396, with expression of the gene *rp49* as an

179 endogenous reference (SYBR-Green assay). Primers used for qRT-PCR are listed in
180 **Table S3**.
181
182 ***RNAi and validation studies***
183 All RNAi lines were originally from the Vienna Drosophila RNAi Center as P-element
184 insertion lines on a co-isogenic w1118 background. Each RNAi line was first tested to
185 determine whether it alone had an effect on eye development by crossing it to GMR-
186 Gal4 and comparing the eye area of the F1 males (or females) to the control cross
187 between w1118 and GMR-Gal4. In all crosses, GMR-Gal4 was used as the maternal
188 parent. To test its effect on the $hINS^{C96Y}$-induced eye degeneration phenotype, the
189 RNAi line was crossed to the GMR>>$hINS^{C96Y}$ line (used as maternal parent), so that
190 both $hINS^{C96Y}$ and the RNAi constructs are driven by GMR-Gal4. The resulting
191 phenotype was compared to the cross between $hINS^{C96Y}$ females and w1118 males. At
192 least 10 individual flies were measured per cross and a t-test was used to determine
193 significance at 0.05 level with multiple testing correction. For mutant lines, GMR-Gal4
194 was replaced with w1118 in the first test and used as a control. The same scheme was
195 used for the second test. It is worth noting that because the mutants were tested in
196 heterozygous states, only dominant interaction with $hINS^{C96Y}$ will be revealed.
197
198 ***sfl expression studies***
199 Six lines carrying the 18 bp indel allele and eight carrying the 4 bp allele were chosen
200 and paired to form 15 crosses (**Figure S1A**). Three sets of ten late 3rd instar (wandering
201 stage) larvae were collected from each cross and dissected in 1X PBS to isolate eye

imaginal discs. RNA isolation and cDNA library preparation are the same as described above. Genomic DNA was extracted from adult flies from the same cross. Because the 18 bp/4 bp polymorphism is in the intron of *sfl*, a SNP in the cDNA was identified that could be used to distinguish the two alleles in each cross (**Figure S1B**). Four such SNPs were chosen and pyro-sequencing assays were designed (primers listed in **Table S3**). Pyro-sequencing was performed as previously described (Wittkopp, 2011). Briefly, each of the three cDNA and one gDNA sample per cross was analyzed by pyrosequencing in four replicate PCR amplifications to determine relative expression. The ratio in genomic DNA analysis was used to account for amplification bias. The resulting 12 ratios were first log2 transformed and analyzed using ANOVA according to the model $y_{ij} = \boldsymbol{\alpha + L_i + \varepsilon_{ij}}$, where α is the estimate of the relative expression ratio, which is expected to be significantly different from zero when the two alleles are differentially expressed; L_i is a random effect term for the biological replicates (i = 1,2,3). For 13 of the 15 crosses the p-value > 0.1; for these crosses the data were fit to a reduced ANOVA model $y_i = \boldsymbol{\alpha + \varepsilon_i}$, from which the estimate and the 95% confidence interval for the ratio of expression (α) were calculated. In the two cases where the random effect term was nominally significant ($P < 0.1$), a linear mixed-effect model was fit using the lme package in R to obtain an estimate and 95% confidence interval for the same ratio.

RESULTS

Effect of natural variation on hINS^{C96Y}-induced eye phenotype

We crossed the transgenic fly line (w; P{GMR-Gal4}, P{UAS-hINS^{C96Y} }/CyO) as the maternal parent to 178 inbred lines from DGRP (only 154 were used in the subsequent GWAS analyses due to genome sequence availability). These lines represent a spectrum of natural variation, except for recessive lethal variants, which were eliminated in the formation of the DGRP. Among several eye phenotypes observed — rough eye, reduced total area, distortion of the oval shape and black lesion spots — we chose total eye area as the phenotype to carry out a GWAS. We quantified eye area in ten male progeny from each hINS^{C96Y} x DGRP cross. We observed a continuously varying distribution of this phenotype, ranging from 13% to 86% of wild type fly eye area (**Figure 1**). ANOVA indicated that 58.6% of the variance is between genotypes (approximately equal to the broad sense heritability (Falconer, 1981, p115), indicating a large genetic component. Males were chosen for measurement and analysis because they showed a more severe phenotype than females (Park et al., 2013). However, we also measured F1 females for a subset of 38 lines and found a strong correlation between the two sexes from the same cross (r=0.8, **Figure S2**).

The observed variation in eye degeneration is consistent with the hypothesis that it reflects differences in cellular response to the expression of hINS^{C96Y}. The severity of the eye degeneration phenotype is not correlated with body size of the same individual, or the mean eye size of the same line; neither is it correlated with GAL4 protein levels in eye imaginal discs (Park et al., 2013). The GWAS described below showed no evidence for association between eye area and SNPs in or surrounding the *glass* (*gl*) locus, the trans-activator of GMR-Gal4, a result consistent with Gal4 protein measurements and the fact that the eye degeneration phenotype is insensitive to GMR-Gal4 gene dose

247 when hINS^{C96Y} is present in single copy (Figure 3 in Park et al., 2013). Finally, when we
248 expressed hINS^{C96Y} in the notum (rather than the eye) and measured the loss of
249 macrochaetae in F1 crosses to 38 DGRP lines for which we also collected eye
250 degeneration data, we observed no correlation between the two traits, indicating that
251 the degeneration phenotypes are not caused by line-specific differences in mutant
252 insulin expression (Park et al., 2013).
253
254 ***Genome-wide association analysis***
255 To identify candidate genetic loci and variants underlying the phenotypic variation, we
256 carried out GWAS on the F1 males from the crosses of hINS^{C96Y} and 154 DGRP lines.
257 We used mean eye area as a quantitative trait to perform single marker regression for
258 1.6 million autosomal SNPs, restricted to bi-allelic sites for which the minor allele
259 frequency is at least 5%. The result revealed a strong peak on chromosome 3L and
260 minor ones on other major chromosome arms (**Figure 2C**). The most significant SNP
261 underlying the chromosome 3L peak has a raw p-value of 2.4×10^{-8} (t-test); the
262 Bonferroni corrected P = 0.04.
263 Population stratification is a potential confounder for GWAS – it can inflate the test
264 statistic for non-associated variants if the population structure correlates with the
265 phenotype. We assessed its impact in our study in three ways. First, we evaluated the
266 Quantile-Quantile (Q-Q) plots for autosomal and X-linked variants. Neither showed a
267 systematic shift towards low p-values compared to the null expectation, which would be
268 expected if population structure induces false association signals (**Figure 2A,B**).
269 Second, we used Principle Component Analysis (PCA) to calculate the top eigenvectors

explaining the most genetic variation in the sample. Plotting the phenotype of each cross against the coordinate of each of the top five eigenvectors revealed no correlation between the two (Materials and Methods, **Figure S3**). Third, because all F1 males inherited their X-chromosome from the GMR>>hINS^{C96Y} tester line, we expect no association between the phenotype and X-linked SNPs. Indeed, we only found an excess of low p-values in autosomal variants, but not in X-linked ones (**Figure 2A,B**). The above analyses suggested that population stratification does not correlate with the trait and does not influence the results of the association study.

Cryptic relatedness, i.e. unknown genetic relationships between individuals in a sample, can also confound the association analysis due to non-independence and larger than expected phenotypic variance (Cheng et al., 2010; Voight & Pritchard, 2005). We estimated the Genetic Relatedness Matrix (GRM) from whole genome SNP data using mixmogam (a Python implementation of EMMAX) (Kang et al., 2010; Segura et al., 2012). We found while the majority of the 154 lines are genetically unrelated (**Figure S4A**), several pairs of lines showed higher levels of relatedness, e.g. RAL-350/RAL-358 and RAL-352/RAL-712 (**Figure S4B**). Next, we performed mixed linear model (MLM) regression to explicitly account for the cryptic relatedness as well as population stratification (Atwell et al., 2010; Yu et al., 2006). A permutation procedure specifically designed to preserve the phenotype covariance structure is used to establish a genome-wide 5% significance threshold (**Text S1**). The resulting p-value distribution is qualitatively similar to the linear regression analysis, and it identified *sfl* as significantly associated with the trait under a permutation-based 5% genome-wide threshold (**Figure S4E**). The most significant SNP (3L:6523119, dm3) has a raw *p*-value of 1.4×10^{-8}.

Below we will focus on identifying the gene(s) underlying the peak and genetically testing its association with the phenotype.

sulfateless (sfl) modifies eye area phenotype

The peak on chromosome 3L is confined to the third intron of the gene, *sfl* (**Figure 2D**). This intron also contains a nested gene (CG32396) lying close to the association peak. CG32396 is predicted to encode a protein with a probable tubulin beta-chain. To determine which of the two genes, or possibly both, is responsible for the association, we examined the expression pattern of each gene and also used RNAi to knock down gene expression. *sfl* is expressed in the eye-antennal imaginal disc and eye and brain in adults (**Figure S5, S6**). CG32396 has a testis-specific expression pattern in adults, with very low expression in the adult eye (**Figure S5**) and no detectable expression in eye imaginal discs by RT-PCR (**Figure S6, S7**).

RNAi knockdown of either *sfl* or CG32396 in the eye imaginal disc had no measurable effect on eye area. In contrast, RNAi against *sfl*, but not CG32396, significantly decreased mean eye area in the presence of $hINS^{C96Y}$ but not $hINS^{WT}$ (**Figure 3**). These results rule out CG32396 as the causal gene, and strongly implicate *sfl* as the genetic modifier of $hINS^{C96Y}$-induced eye degeneration.

To test if *sfl* also modifies the $hINS^{C96Y}$-induced phenotype in other tissues, we carried out RNAi knockdown of *sfl* in the developing wing (using a dpp-Gal4 driver) and notum (using an ap-Gal4 driver). In both experiments we observed more severe phenotypes than that caused by $hINS^{C96Y}$ alone (**Figure S8, S9**). However, the interpretation is made complicated by the fact that *sfl* knockdown alone causes mutant

316 phenotypes in these tissues, consistent with previous knowledge (Lin, 2004). At present
317 we cannot distinguish the alternative hypotheses of additive vs. epistatic interactions
318 between *sfl* and hINS^{C96Y}.
319
320 ***Heparan sulfate biosynthetic pathway modifies the hINS^{C96Y}-induced eye***
321 ***degeneration***
322 *Sulfateless* encodes a bi-functional enzyme in the heparin sulfate biosynthesis pathway.
323 An important component of the cell surface and extracellular matrix (Kirkpatrick &
324 Selleck, 2007), heparan sulfate-modified proteins, or proteoglycans (HSPG) regulate
325 signaling during development by influencing the levels and activity of growth factors and
326 morphogens at cell surfaces and in the extracellular matrix (Fujise et al., 2003; Giráldez
327 et al., 2002; Häcker et al., 1997; Kirkpatrick et al., 2004; Nakato et al., 1995). The
328 involvement of HSPGs in the cellular responses to misfolded proteins (proteostasis) has
329 not been previously described.
330 To further examine the hINS^{C96Y}-dependent interaction of *sfl*, we examined RNAi
331 knockdowns and mutants for two additional genes in the HS biosynthetic pathway: *ttv*
332 and *botv*, producing the glycosaminoglycan polymer that is modified by *sfl* (Lin, 2004).
333 SNPs in neither of the genes showed evidence of association in our GWAS (lowest
334 adjusted $P > 0.5$ in both loci, adjusted for multiple-testing using Bonferroni's method).
335 RNAi knockdown of both genes shows a hINS^{C96Y}-dependent effect on eye area in the
336 same direction as *sfl* RNAi (**Figure 4**). In addition, a mutant allele of *botv* also showed a
337 significant dominant enhancement of the eye degeneration phenotype. These results

implicate HSPGs in modifying the cellular response to misfolded proteins. Neither of the genes, however, was identified in the GWAS.

Intronic variation and sfl expression

We re-sequenced a 3 kb region containing the GWAS peak in *sfl* (and the nested gene CG32396) in 19 of the 154 DGRP lines and the transgenic hINS^{C96Y} stock to identify all the variants in this region. We found that the SNP achieving the lowest p-value genome-wide was an 18 bp/4 bp length polymorphism (relative to the *D. simulans* orthologous sequence) (**Figure 5A**). We also found three other insertion/deletion (INDEL) polymorphisms in this region, with sizes ranging from 4 – 30 bp and the minor alleles (deletion in all three cases) being present only once or twice in the sample. In contrast, the 18/4 bp polymorphism is present at 50% frequency in the DGRP sample. Below we will use the term "Single Feature Polymorphism" (SFP) to refer to both INDEL and single nucleotide polymorphism in the *sfl* locus.

A plot of haplotype structure surrounding the association peak (Haploview v4.2) pinpoints an LD block of 400 bp (block 66 in **Figure 5A**, chr3L:6523119-6523518). There are two major haplotypes in this block, each represented by two equal-sized groups among the 154 DGRP lines (**Figure 5B**). For convenience, we refer to these two haplotypes as the 18bp or 4bp allele, although it is worth noting that we don't have the ability to distinguish between the SFPs within this block, unless further recombinant individuals are sampled or generated.

Because all coding variants in *sfl* lie outside of this 400 bp LD block, we hypothesized that one or more of these intronic SFPs are the causal variant(s) and

modify the hINS^{C96Y}-induced eye phenotype by altering *sfl* expression. We tested this hypothesis by examining the correlation between the allelic states and the allele-specific expression level. We selected pairs of 4 bp and 18 bp lines from the respective phenotypic spectrum, crossed them to obtain F1 individuals heterozygous for the two alleles, and used pyro-sequencing to estimate the relative expression of the two alleles in eye imaginal discs. This method allowed us to measure the ratio of expression of *sfl* associated with each allele in the same animal, thereby controlling for both the trans-environment as well as experimental noise, resulting in highly reproducible results (**Figure S10**). Based on RNAi knock-down of *sfl*, which enhanced the hINS^{C96Y} phenotype, we expected the 4 bp allele (associated with more severe phenotypes in the GWAS) to produce less transcript than the 18 bp allele.

Allele-specific expression of *sfl* differed in both magnitude and direction among the 15 crosses (**Figure 6**). Seven crosses supported the hypothesis by exhibiting significantly greater expression from the 18 bp allele, with an 18/4 bp ratio ranging from 1.03 - 2.8 (median = 1.15). Two crosses, however, showed slightly greater expression from the 4 bp allele (18 bp/4 bp ratios of 0.94 and 0.96). The remaining six crosses showed no significant differences in expression of the two alleles in our test. While both more strains showed higher expression of the transcript linked to the 18 bp allele, and the difference in this direction is stronger, the small sample size and the modest correlation between the allelic states and the transcription level prevented us from drawing a conclusion. Proving the causal mutation(s) and identifying the mechanisms require further experiments making precise changes at the candidate loci and assaying the effects in the same genetic background.

Search for additional association by conditional analysis

In light of the above finding, we carried out a conditional analysis to identify variants that act independently of the 18 bp/4 bp SFP. To do so, we tested variants other than the 18/4 bp SFP, either within the *sfl* locus or genome-wide, by treating the 18/4 bp SFP as a covariate in a linear regression model. After accounting for multiple testing, we observed no significant signals in either case (**Figure S11**). The lack of significance genome-wide may be attributable to the lack of power after correcting for multiple testing. The analysis restricted to the 40 kb *sfl* locus reduces the burden of multiple testing by several orders of magnitude, but also fails to identify a significant association. Considering the large range of allele-specific expression differences between the 18 bp and 4 bp alleles observed in the 15 crosses, the additional *cis*-acting expression variants must either be low frequency alleles or have epistatic properties, two situations this analysis would be underpowered to detect.

DISCUSSION

sfl and hINS^{C96Y}-induced eye degeneration

Statistical (GWAS) and genetic (RNAi) evidence support a role for *sfl* as a natural genetic modifier for hINS^{C96Y}-induced eye degeneration. Although we conducted a GWAS for dominant-acting modifiers in a relatively small sample of lines (154) considering the large number of segregating common SNPs (1.6M), we found statistical support for a QTL in *sfl* in a mixed model analysis, which addresses effects of both population structure and genetic relatedness in the sample. One possible reason the *sfl*

QTL achieves statistical significance is because the two alternative alleles occur at approximately a 50% frequency in the sample, where GWAS is maximally powerful.

RNAi knockdown experiments showed that perturbation of *sfl* expression, and also two other genes in the HS biosynthesis pathway, has a measurable effect on eye degeneration, but only in the presence of $hINS^{C96Y}$ expression, indicating a specific interaction between protein misfolding and HS biosynthesis (also see (Park et al., 2013). RNAi against CG32396, the gene nested inside the intron of *sfl*, had no effect on eye area both in the absence and presence of $hINS^{C96Y}$, suggesting that the $hINS^{C96Y}$-induced eye degeneration phenotype is not simply a consequence of RNAi expression. We caution, however, that a genetic proof of *sfl* as modifying the phenotype in this population will require additional studies.

A direct test for *sfl* and the intronic variation being causal would be to genetically engineer two lines in the same genetic background, differing only at the *sfl* locus. A potential caveat of this approach lies in the assumption that the differential activity of the two alleles is independent of the genetic background (*i.e.,* no epistasis), which, if violated, will lead to a false negative result (Chandler et al., 2013). We used instead an indirect approach by examining the correlation between the allelic states and the expression level. To take into account the genetic background differences, we measured allele-specific gene expression of 18bp and 4bp *sfl* alleles in 15 different "controlled" genetic backgrounds, but keeping the background the same for the two alleles by comparing their expression ratios in heterozygotes. The results are mixed: the ratio of expression from the 18 bp/4 bp alleles differed in the 15 crosses, ranging from 2.8 to 0.94 (**Figure 6**); nearly half (7/15) showed greater expression from the allele

associated with the 18 bp variant, consistent with the expectation based on the RNAi result; two showed a small difference in the contrary direction (18bp/4bp ratio = 0.94 and 0.96); the remaining six were insignificant in our test. This marked heterogeneity in expression means we can neither accept nor reject the hypothesis of a causal role for the intronic variants and the expression level of *sfl*. Hence we are also not able to conclude that expression difference is the mechanism underlying the genotype-phenotype association, though it remains a possibility. Future experiments employing genome-editing technologies will allow better resolution of the mechanism(s) underlying the association (Gratz et al., 2013; Jinek et al., 2012; Ran et al., 2013).

Finally, we investigated whether additional eQTLs exist in *sfl* or in other genes acting epistatically with *sfl*. Likely due to lack of power, a conditional analysis failed to identify additional variants in the *sfl* locus or elsewhere in the genome. However, it is now well established that gene expression is a highly polygenic trait in *D. melanogaster*, with many eQTLs contributing to expression variability both in cis and trans (Brem et al., 2005; Brem et al., 2002; West et al., 2007), and intra-locus genetic complexity influencing a quantitative trait has long been known, as in the Adh example (King et al., 2012). In the 40 kb region spanning the *sfl* locus alone, 1,358 SNPs are present among the 14 lines used in this experiment, which individually or in combination could influence expression of the gene. Thus, predictions based on one or two strongly associated variant(s) is not adequate. A polygenic risk predictor may be needed to summarize contributions even from a single locus.

HSPG function and misfolded protein response

Our study identified the HS biosynthesis pathway (*sfl*, *ttv* and *botv*) as a modifier of eye degeneration induced by expression of a misfolded human proinsulin protein. Although we do not yet know whether this response is to a specific misfolded protein ($hINS^{C96Y}$) or whether it applies to a broader class of misfolded proteins, our discovery now implicates the HSPGs in the regulation of cellular proteostasis.

We propose that genetic variation in HS biosynthesis influences the response to misfolded protein through its biological activity in vesicular trafficking of misfolded protein. HS-modified proteins (heparin sulfate proteoglycans, HSPGs) are abundant components of cell surfaces and extracellular matrices, and are best understood for their roles in cell signaling and in functioning as co-receptors, processes integral to normal development (H\"{a}cker et al., 2005; Kirkpatrick & Selleck, 2007). HSPGs are also involved in endocytocis (Ren et al., 2009; Stanford et al., 2009) and vesicular trafficking (Nybakken & Perrimon, 2002; Sarrazin et al., 2011), roles that may link them to cellular response to misfolded proteins (Higashio & Kohno, 2002; Kim et al., 2009; Kimmig et al., 2012).

HSPGs may also influence membrane trafficking indirectly, perhaps by regulating signaling events that impinge on trafficking processes. The generation of phosphatidylinositol (3,4,5) triphosphate [PtdIns(3,4,5)P_3] by type I phosphoinositide (PI) 3-kinases is affected by a number of growth factors and cytokines, many of which are influenced by HSPGs as accessory molecules. PtdIns(3,4,5)P_3 affects a number of trafficking events, including endocytosis and autophagy (Downes et al., 2005).

In a yeast study of the mutant protein folding assistant, protein disulfide isomerase (Pdi1a'), the authors found that more than 50% of the 130 genes identified as synthetic-

lethal were related to vesicle trafficking, while only 10 belonged to the canonical unfolded protein response (UPR) pathway (Kim et al., 2009). In another study, Kimmig *et al* found an enrichment of vesicle-trafficking related genes among those that changed expression significantly after induction of ER-stress (Kimmig et al., 2009). Both studies indicate that a global regulation of vesicle trafficking is important to a cell's response to unfolded or misfolded protein. Activation of UPR has also been shown to affect ER-to-Golgi transport via stimulation of COPII vesicle formation from the ER (Higashio & Kohno, 2002). We propose that either natural variation or genetic perturbation of HS biosynthesis influences the global regulation of vesicle trafficking, which in turn affects the cell's ability to process an excess of unfolded or misfolded protein. Prolonged ER-stress may then lead to apoptosis.

Genetic architecture of the* hINS^{C96Y}*-induced eye degeneration phenotypes

Phenotypic heterogeneity that is dependent on the genetic background is a common phenomenon, and in humans, imposes a significant challenge in both diagnosis and treatment. Our fly model provides a tractable system for studying the genetic and molecular basis for such phenotypic heterogeneity, but with limitations imposed by the sample size of the study. To assess the power for identifying QTL using this population, we did a simple calculation for a t-test based statistic at $P = 0.05$ level, with Bonferroni's correction for multiple testing, which indicates that we have 66% power to identify a variant at 50% population frequency, with an effect-size of 1.0 (measured as the shift in phenotypic mean in units of standard deviation of the trait, see **Table S4**). This example was chosen to match the estimates for the 18 bp/4 bp indel polymorphism in the *sfl*

499 intron in the sample of 154 crosses. Any variant with a smaller effect-size and/or lower
500 frequency than the 18bp/4bp polymorphism would likely have been missed in this study.
501 *Sulfateless* was the only QTL identified as genome-wide significant in this study
502 (**Figure 2, S4**); its association with the trait is robust with respect to population structure
503 and cryptic relatedness (**Figure S3, S4**). This does not mean, however, that the genetic
504 architecture for the hINS^{C96Y}-induced eye phenotype involves a single locus. Rather, we
505 have several reasons to believe the genetic architecture must involve many loci. First,
506 the distribution of the phenotype, i.e. eye areas expressed as line means, suggests a
507 non-Mendelian genetic basis (**Figure 1**). Second, while ANOVA estimates that nearly
508 60% of the total phenotypic variance is between crosses, less than 20% within the 60%
509 (i.e. < 12% of the total variance) can be attributed to the *sfl* locus. Even this 20%
510 estimate, because it is derived from the same population used to identify the locus, is
511 liable to be an overestimate due to the Winner's Curse effect (Garner, 2007).
512 To estimate what percentage of the between-cross variance can be explained by the
513 additive effects of common variants combined, we applied the GCTA tool, which uses a
514 mixed linear model method, to the line means of the 154 crosses (Yang et al., 2011).
515 The result showed that 83% (standard error = 37%) of the variance between crosses
516 could be attributed to common, autosomal variants with minor allele frequencies greater
517 than 5%. Analysis using GEMMA (v0.94beta), which used a Bayesian method, achieved
518 nearly identical results (posterior mode = 0.83, s.e. = 0.41). We then did the same
519 analysis with GCTA, but including the 18bp/4bp indel polymorphism as a covariate to
520 remove the effect of *sfl*, in order to estimate the remaining additive heritability. As a
521 result, we got 62% (s.e. 47%). The large standard error as a result of the limited sample

size leaves the proportion of variance explained by all common SNPs undetermined. However, the estimates are encouraging and suggest that a potentially large proportion of phenotype variance may be explained by additional loci, which require larger sample size to identify.

Relationship to common, complex diseases

While our fly model is of a monogenic form of diabetes, it exhibits a complex genetic architecture when placed on a diverse set of genetic backgrounds. We posit that fly models of monogenetic disease are suitable subjects for the genetic dissection of common disorders in human.

One role of the Mendelian mutation is to sensitize the fly to allow phenotypic effects of background genetic modifiers to become visible. Although common disorders are normally considered as lacking a major mutation, a careful consideration suggests that this view is inaccurate. What common disorders lack are large-effect mutations shared by a substantial proportion of the affected individuals. For many diseases, perturbation may be required to boost the expressivity of additive genetic variation that would otherwise be cryptic, i.e., below a disease-causing threshold. Such a perturbation could be genetic, such as driver mutations in cancer, but could also be environmental, such as diet and life-style changes in the case of cardiovascular disease and type 2 diabetes. Consistent with this view, it has been proposed that recent genome evolution and rapid environmental as well as cultural changes in human history have decanalizing effects on physiology, which release cryptic genetic variation and underlie the rising incidence of common human disorders (Gibson, 2009).

A genetic screen for naturally occurring modifiers in a sensitized background, such as the one we employed here, should apply equally well in the study of Mendelian or complex disease. Were this not the case, two different classes of genetic modifiers would have to be posited. An intriguing question, which we found little empirical evidence for or against, could be addressed in the fly by constructing a series of sensitized backgrounds utilizing different disease-causing mutant hINS alleles of varying effect on disease (*e.g.,* neonatal diabetes *vs.* maturity-onset diabetes of the young (Støy et al., 2007)), and comparing the composition of naturally occurring modifiers.

Advantages of a fly model of complex disease

A primary mutation can manifest itself in different ways and with tissue-specific effects (Mefford et al., 2008), possibly a consequence of its interdependence with the individual's genetic background. The binary Gal4-UAS system enables the creation of a series of models using the same disease mechanism, but directed to different tissues with high tissue-specificity. The ability to construct and study multiple related models in parallel can provide insight into the basis of disease heterogeneity. In the accompanying paper we show, for example, that the developing eye and notum have different sets of genetic background modifiers of $hINS^{C96Y}$-dependent disease (Park et al., 2013). Sex-specific difference in disease risk and severity are also readily modeled in the fly. In both the fly and mouse model of $hINS^{C96Y}$-induced disease, males consistently show more severe disease phenotypes (Wang et al., 1999; Park et al., 2013).

Drosophila models of human disease provide a useful alternative to the study of complex disease in patient populations. First, many models of human disease have been established in the fly, most notably neurodegeneration and cancer (Bilen & Bonini, 2005; Gonzalez, 2013). We predict that natural variation will influence the severity of disease phenotypes in all of them. Second, many models of disease can be created by expression of a mutant allele, which makes them suitable for F1 screens between a tester stock and inbred population collections, such as we employed here. Our study shows that dominant genetic variation for disease severity is abundant. This outcrossing design also avoids unwanted effects of inbreeding on traits and better mimics the natural heterozygosity of low frequency variants. Third, this experimental design facilitates repeated measurement of a disease phenotype, thereby increasing the power to detect a causal association (Mackay et al., 2009). Fourth, LD is low in *D. melanogaster* and SNP are 20-40X more abundant than in humans. Finally, both forward and reverse genetics can be applied to investigate the biology and pathway genetics of candidate variants. For all these reasons we believe fly models will prove useful in understanding the genetic architecture of complex human disease.

ACKNOWLEDGMENTS

This work was funded by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK013914 and P30 DK020595), the National Institute of General Medical Sciences (GM081892), the Chicago Biomedical Consortium with support from the Searle Funds at The Chicago Community Trust, and a gift from the Kovler Family Foundation. SBS is supported by GM054832 and PJW is supported by

590 NSF MCB-1021398. We thank Dan Nicolae for technical help and advice on GWAS,
591 and Xiang Zhou and Matthew Stephens for advice on the mixed linear model approach
592 using Gemma. We thank Jian Yang and Peter Visscher for help with the interpretation
593 of the GCTA results. Joseph Coolon in the Wittkopp lab helped design the pyro-
594 sequencing assays, and Ellen Pederson at the DNA sequencing center at the University
595 of Michigan provided technical assistance. We also thank the anonymous reviewer and
596 Dr. Sabatti for helpful comments.
597

598 LITERATURE CITED

599 Atwell, S., Huang, Y. S., Vilhjalmsson, B. J., et al., 2010 Genome-wide association
600 study of 107 phenotypes in Arabidopsis thaliana inbred lines *Nature* **465**: 627--631
601 Bilen, J. and Bonini, N. M., 2005 Drosophila as a model for human neurodegenerative
602 disease. *Annual review of genetics* **39**: 153--171
603 Brem, R. B., Storey, J. D., Whittle, J., et al., 2005 Genetic interactions between
604 polymorphisms that affect gene expression in yeast. *Nature* **436**: 701--703
605 Brem, R. B., Yvert, G., Clinton, R., et al., 2002 Genetic dissection of transcriptional
606 regulation in budding yeast. *Science* **296**: 752--755
607 Chandler, C. H., Chari, S., and Dworkin, I., 2013 Does your gene need a background
608 check? How genetic background impacts the analysis of mutations, genes, and
609 evolution. *Trends in genetics :* **29**: 358--366
610 Cheng, R., Lim, J. E., Samocha, K. E., et al., 2010 Genome-wide association studies
611 and the problem of relatedness among advanced intercross lines and other highly
612 recombinant populations. *Genetics* **185**: 1033--1044
613 Chintapalli, V. R., Wang, J., and Dow, J. A. T., 2007 Using FlyAtlas to identify better
614 Drosophila melanogaster models of human disease. *Nat Genet* **39**: 715--720
615 Downes, C. P., Gray, A., and Lucocq, J. M., 2005 Probing phosphoinositide functions in
616 signaling and membrane trafficking. *Trends in cell biology* **15**: 259--268
617 Fujise, M., Takeo, S., Kamimura, K., et al., 2003 Dally regulates Dpp morphogen
618 gradient formation in the Drosophila wing. *Development* **130**: 1515--1522
619 Garner, C., 2007 Upward bias in odds ratio estimates from genome-wide association
620 studies. *Genet. Epidemiol.* **31**: 288--295
621 Gibson, G., 2009 Decanalization and the origin of complex disease *Nat Rev Genet* **10**:
622 134--140
623 Giráldez, A. J., Copley, R. R., and Cohen, S. M., 2002 HSPG modification by the
624 secreted enzyme Notum shapes the Wingless morphogen gradient. *Developmental*
625 *Cell* **2**: 667--676
626 Gonzalez, C., 2013 Drosophila melanogaster: a model and a tool to investigate
627 malignancy and identify new therapeutics. *Nature Reviews Cancer* **13**: 172--183
628 Gratz, S. J., Cummings, A. M., Nguyen, J. N., et al., 2013 Genome Engineering of
629 Drosophila with the CRISPR RNA-Guided Cas9 Nuclease *Genetics* **194**: 1029--
630 1035
631 Häcker, U., Lin, X., and Perrimon, N., 1997 The Drosophila sugarless gene modulates
632 Wingless signaling and encodes an enzyme involved in polysaccharide biosynthesis.
633 *Development* **124**: 3565--3573
634 Häcker, U., Nybakken, K., and Perrimon, N., 2005 Heparan sulphate proteoglycans:
635 the sweet side of development. *Nature Reviews Molecular Cell Biology* **6**: 530--541
636 Higashio, H. and Kohno, K., 2002 A genetic link between the unfolded protein response
637 and vesicle formation from the endoplasmic reticulum. *Biochemical and biophysical*
638 *research communications* **296**: 568--574
639 Jinek, M., Chylinski, K., Fonfara, I., et al., 2012 A Programmable Dual-RNA–Guided
640 DNA Endonuclease in Adaptive Bacterial Immunity *Science* **337**: 816--821

641 Kang, H. M. M., Sul, J. H. H., Service, S. K., et al., 2010 Variance component model to
642 account for sample structure in genome-wide association studies. *Nature genetics*
643 **42**: 348--354
644 Kim, J.-H. H., Zhao, Y., Pan, X., et al., 2009 The unfolded protein response is
645 necessary but not sufficient to compensate for defects in disulfide isomerization.
646 *The Journal of biological chemistry* **284**: 10400--10408
647 Kimmig, P., Diaz, M., Zheng, J., et al., 2012 The unfolded protein response in fission
648 yeast modulates stability of select mRNAs to maintain protein homeostasis. *eLife* **1**:
649 e00048
650 King, E. G., Merkes, C. M., McNeil, C. L., et al., 2012 Genetic dissection of a model
651 complex trait using the Drosophila Synthetic Population Resource. *Genome*
652 *research* **22**: 1558--1566
653 Kirkpatrick, C. A., Dimitroff, B. D., Rawson, J. M., et al., 2004 Spatial regulation of
654 Wingless morphogen distribution and signaling by Dally-like protein. *Developmental*
655 *Cell* **7**: 513--523
656 Kirkpatrick, C. A. and Selleck, S. B., 2007 Heparan sulfate proteoglycans at a glance
657 *Journal of Cell Science* **120**: 1829--1832
658 Kolde, R., 2011 pheatmap: Pretty Heatmaps (http://cran.r-
659 project.org/package=pheatmap)
660 Lin, X., 2004 Functions of heparan sulfate proteoglycans in cell signaling during
661 development *Development* **131**: 6009--6021
662 Listgarten, J., Lippert, C., Kadie, C. M., et al., 2012 Improved linear mixed models for
663 genome-wide association studies. *Nature methods* **9**: 525--526
664 Mackay, T. F. C., Stone, E. A., and Ayroles, J. F., 2009 The genetics of quantitative
665 traits: challenges and prospects *Nature Reviews Genetics* **10**: 565--577
666 Mefford, H. C., Sharp, A. J., Baker, C., et al., 2008 Recurrent rearrangements of
667 chromosome 1q21.1 and variable pediatric phenotypes. *The New England Journal*
668 *of Medicine* **359**: 1685--1699
669 Nakato, H., Futch, T. A., and Selleck, S. B., 1995 The division abnormally delayed
670 (dally) gene: a putative integral membrane proteoglycan required for cell division
671 patterning during postembryonic development of the nervous system in Drosophila.
672 *Development* **121**: 3687--3702
673 Nybakken, K. and Perrimon, N., 2002 Heparan sulfate proteoglycan modulation of
674 developmental signaling in Drosophila. *Biochimica et biophysica acta* **1573**: 280--
675 291
676 Park, S.-Y., Ludwig, M. Z., Tamarina, N. A., et al., 2013 A Drosophila Model for
677 Misfolded Protein Induced Neurodegeneration. *Genetics accepted*
678 Ran, F. A., Hsu, P. D., Lin, C.-Y., et al., 2013 Double Nicking by RNA-Guided CRISPR
679 Cas9 for Enhanced Genome Editing Specificity *Cell* **154**: 1380--1389
680 Ren, Y., Kirkpatrick, C. A., Rawson, J. M., et al., 2009 Cell type-specific requirements
681 for heparan sulfate biosynthesis at the Drosophila neuromuscular junction: effects on
682 synapse function, membrane trafficking, and mitochondrial localization. *The Journal*
683 *of neuroscience* **29**: 8539--8550
684 Roy, S., Ernst, J., Kharchenko, P. V., et al., 2010 Identification of Functional Elements
685 and Regulatory Circuits by Drosophila modENCODE *Science* **330**: 1787--1797

Sarrazin, S., Lamanna, W. C., and Esko, J. D., 2011 Heparan Sulfate Proteoglycans *Cold Spring Harbor Perspectives in Biology* **3**

Segura, V., Vilhjalmsson, B. J., Platt, A., et al., 2012 An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* **44**: 825--830

Stanford, K. I., Bishop, J. R., Foley, E. M., et al., 2009 Syndecan-1 is the primary heparan sulfate proteoglycan mediating hepatic clearance of triglyceride-rich lipoproteins in mice. *The Journal of clinical investigation* **119**: 3236--3245

Støy, J., Edghill, E. L., Flanagan, S. E., et al., 2007 Insulin gene mutations as a cause of permanent neonatal diabetes *Proceedings of the National Academy of Sciences* **104**: 15040--15044

Voight, B. F. and Pritchard, J. K., 2005 Confounding from cryptic relatedness in case-control association studies. *PLoS genetics* **1**: e32

Wang, J., Takeuchi, T., Tanaka, S., et al., 1999 A mutation in the insulin 2 gene induces diabetes with severe pancreatic beta-cell dysfunction in the Mody mouse. *The Journal of clinical investigation* **103**: 27--37

West, M. A., Kim, K., Kliebenstein, D. J., et al., 2007 Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* **175**: 1441--1450

Wittkopp, P. J., 2011 Using pyrosequencing to measure allele-specific mRNA abundance and infer the effects of cis- and trans-regulatory differences. *Methods in molecular biology (Clifton, N.J.)* **772**: 297--317

Yang, J., Lee, S. H., Goddard, M. E., et al., 2011 GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**: 76--82

Yu, J., Pressoir, G., Briggs, W. H., et al., 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness *Nature Genetics* **38**: 203--208

FIGURE LEGENDS

Figure 1 Distribution of eye area in $hINS^{C96Y}$ x DGRP crosses. Mean ± 1 s.d., sorted by the mean, is shown for crosses between the transgenic {GMR>>$hINS^{C96Y}$} line to 178 DGRP lines, and two randomly chosen DGRP inbred lines (red). Representative photographs of eyes from across the range of the distribution are shown. The rightmost image is of a non-transgenic wild type fly eye.

Figure 2 Genome-wide scan identifies candidate locus associated with the $hINS^{C96Y}$-induced phenotype. Quantile-Quantile (Q-Q) plot reveals an excess of small p-values on autosomes (A) but not on the X chromosome (B), which is not variable in the mapping population due to cross design. (C) Manhattan plot shows a strong peak (green) on chromosome 3L. The blue and red horizontal lines indicate raw $P < 10^{-5}$ and Bonferroni corrected $P < 0.05$, respectively. (D) UCSC browser view of the *sfl* locus containing the association peak. The intron containing the peak also contains a nested gene CG32396.

Figure 3 RNAi knockdown confirms *sfl* and excludes CG32396 as the causal gene. The effect of knocking down either CG32396 or *sfl* was tested in the absence ({UAS-RNAi} x {GMR-Gal4}) or presence ({UAS-RNAi} x {GMR-Gal4, UAS-$hINS^{C96Y}$}) of $hINS^{C96Y}$. Compared to the control crosses (first and third columns in both sexes), significant difference in mean eye area was observed only with RNAi against *sfl* and only in the presence of $hINS^{C96Y}$ (n=15, asterisks above a box plot indicate significant differences at 0.05 level determined by a student's t-test, with Bonferroni correction for multiple testing). In box plots, the median (black dot), interquartile (box) and 1.5 times

739 the interquartile range (whiskers) are indicated; data points outside the range are
740 represented by circles.
741
742 **Figure 4** RNAi and mutant analysis for heparin sulfate biosynthesis pathway genes.
743 The experimental design is the same as in Figure 3. Left panel shows the effect of RNAi
744 or mutant alleles in the absence of $hINS^{C96Y}$ expression; right panel shows the effect
745 when $hINS^{C96Y}$ is expressed in the eye imaginal disc. Mutants were tested in
746 heterozygous states for a dominant interaction with $hINS^{C96Y}$. Fifteen male flies are
747 measured for each group. The statistical significance of differences from the control
748 cross (gray, w1118) was determined by a two-sided student's t test. Those that are
749 significant at 0.05 level after Bonferroni correction are marked with a red arrowhead.
750
751 **Figure 5** Sequencing of a 3 kb region in *sfl* and the LD patterns in the region. (A)
752 Alignment of 19 DGRP sequences ordered by their eye degeneration phenotype (mean,
753 most severe on the bottom). The $hINS^{C96Y}$ transgenic line (asterisk) was also
754 sequenced. Red ticks and white spaces indicate SNPs and deletions relative to the
755 reference sequence. No insertions relative to the reference were found. The purple
756 track shows the $-\log_{10}$ of GWAS P-values. The bottom track shows the linkage blocks as
757 determined by Haploview (4.02) using the solid spine method with default settings (D' >
758 0.8). (B) Detailed haplotype block structures. Each numbered column represents a
759 polymorphic site, with the alleles colored as blue or red; each row represents a
760 haplotype with frequency > 0.01. An arrowhead marks the 18/4 bp indel polymorphism
761 (see text; 18 bp: blue; 4 bp: red). Finally, the number between any two blocks

762 represents the multi-allelic D', which quantifies the associations between adjacent
763 blocks.
764
765 **Figure 6** Pyro-sequencing measure of *sfl* allele-specific transcript ratio in 18 bp/4 bp
766 heterozygotes. (A) Schematic diagram of the pyro-sequencing approach. Colored lines
767 represent transcripts (mRNA) associated with either the 18 bp or the 4 bp allele,
768 expressed at different levels. Common primers were used to amplify both transcripts of
769 the gene of interest from the cDNA library made from eye imaginal disc tissues. Pyro-
770 sequencing was carried out on the amplified products. (B) A pyrogram of a heterozygote
771 with the polymorphic site (G/C) that is diagnostic for the 18 bp/4 bp indel highlighted.
772 The ratio of the two peaks (light intensity, y-axis) are used to calculate the relative ratio
773 of the two alleles. (E: enzyme, S: substrate, A/C/G/T: nucleotides). (C) Log2
774 transformed ratio of 18 bp/4 bp allele expression in 15 crosses between randomly
775 paired 18bp and 4bp lines. Estimates of the ratio and 95% confidence intervals are
776 plotted. The dotted line corresponds to equal expression from the two alternative alleles.
777

SUPPLEMENTARY FIGURES

Figure S1 Pyro-sequencing cross and assay design. (A) Cross design for pyro-sequencing. Six 18 bp and eight 4 bp lines were chosen from the 154 DGRP lines used in GWAS. The Bloomington center stock number is listed. In each cell, the order of the letter/number indicates the direction of the cross. For example, A1 indicates that males of #28240 were crossed to virgin females of #28190. (B) Pyro-sequencing assays. Four SNPs were selected within the transcribed regions to distinguish alleles associated with the 18/4 bp indel polymorphism.

Figure S2 Correlations of eye area between F1 males and females within the same cross. Mean ± 1 s.d. are plotted for a subset of 38 lines. The least square linear fit is indicated.

Figure S3 Population structure assessed through principal component analysis (PCA) using 900K autosomal SNPs after LD pruning. (A) 154 DGRP inbred lines projected onto the plane spanned by the first two principal components (PC1, PC2). The points are colored according to the phenotype severity in the $hINS^{C96Y}$ crosses (red: severe, or first 25%; blue: intermediate, 25%-75%; green: mild, 75%-100%, percentiles in eye area distribution from small to large). (B) projection onto PC1 grouped by the severity of the eye phenotype showed no correlation between the two.

Figure S4 Mixed linear model regression accounting for cryptic relatedness. (A) The heat map shows a 154 x 154 matrix representing the centered genetic relatedness

matrix (GRM) estimated using EMMAX. The boxed area is shown in detail in (B), with their line ID (RAL#) indicated on the right and bottom. The GRM was used in a mixed linear model to perform genome wide association in the 154 lines. And the resulting p-values for autosomal and X-linked variants are plotted as Q-Q plot in (C) and (D), with the red line indicating matches between the data and the null (uniform) p-value distribution. (E) Manhattan plot showing the –log10 p-values against the chromosomal coordinates. No association is expected on the X chromosome. The blue dotted line indicates a Bonferroni corrected $P < 0.05$, while the red solid line indicates a 5% genome-wide significant level based on 500 permutations.

Figure S5 FlyAtlas expression report for CG32396 and sfl. (A) CG32396 (B) *sfl*. Figure obtained through FlyBase.

Figure S6 Quantitative RT-PCR quantification of mRNA levels for CG32396 and *sfl* in eye imaginal disc samples. Two DGRP lines were chosen and eye imaginal discs were prepared from either six male or six female larvae, resulting in four biological samples. qRT-PCR were performed for each sample and three genes (RP49 -- red curve, *sfl* -- yellow, and CG32396 -- green). Shown is the amplification plot: x-axis -- cycle number; y-axis -- base-line corrected relative fluorescence intensity proportional to the amount of amplicons. Both RP49 and *sfl* were first detected in the 18-20th cycle, whereas the appearance of CG32396 did not occur until after 32 cycle. In addition, multiple melting points were detected for CG32396 assays, but not in the other two genes.

Figure S7 Relative quantity of mRNA quantified by qRT-PCR in male and female larvae. In each category, the first three bars represent three independent female larvae sample (whole larva), each assayed with three technical replicates. The heights of the bars represent the mean; the full range of RQ values are indicated by the error bars. The next three bars correspond to three independent male larvae assayed for the same gene. kl-3 and Pp1-Y2 are both located on the Y-chromosome and are known to have testis-specific expression. The RQ values were measured using RP49 gene as the internal control, with the first female larva sample (F-1) as the reference and RQ set to one.

Figure S8 Depletion of *sfl* by RNAi in the developing wing expressing hINS^{C96Y} driven by dpp-Gal4. For both females and males, dpp >> hINS^{C96Y} or Dpp >> *sfl* RNAi expression alone reduces wing area between the L2 and L4 longitudinal veins relative to the posterior-most sector of the wing (bordered by L5). This reduction is more severe in the *sfl* knockdown genotype than in the hINA^{C96Y}-expressing genotype. Co-expression of *sfl* RNAi and hINS ^{C96Y} by dpp-Gal4 results in the obliteration of the L3 vein and further relative reduction of the L2-L4 area. (A): Wild type wing showing the measured regions of the wing used to quantify the effects of both *sfl* RNAi and hINS^{C96Y} expression in dpp-Gal4 domain (L3-L4 intervein sector). Quantification of the (B) female or (E) male wing phenotypes generated by transgenes dpp-Gal4; dpp-Gal4 >UAS-hINS^{C96Y}; (C, G) dpp-Gal4 >> UAS-*sfl* RNAi; and (D, H) dpp-Gal4 >>UAS-*sfl* RNAi; UAS-hINS^{C96Y}. The values represent the ratio of the third posterior cell (in pink

color) divided by the L2-L4 intervein sector (in green color) wing area. ***, P < 0.001; Mann-Whitney U test.

Females: dpp-Gal4 (n= 15; Mean= 0.62), dpp-Gal4 >UAS-hINS^{C96Y} (n= 15; Mean=0.65), dpp-Gal4 >> UAS-*sfl* RNAi (n= 23; Mean=1.3) and dpp-Gal4 >>UAS-*sfl* RNAi; UAS-hINS^{C96Y} (n= 22; Mean=1.76).

Males: dpp-Gal4 (n= 15; Mean=0.59), dpp-Gal4 >UAS-hINS^{C96Y} (n= 15; Mean=0.64), dpp-Gal4 >> UAS-*sfl* RNAi (n=23; Mean=1.2) and dpp-Gal4 >>UAS-*sfl* RNAi; UAS-hINS^{C96Y} (n= 29; Mean=1.68).

Figure S9 Depletion of *sfl* by RNAi in the developing notum expressing hINS^{C96Y} driven by ap-Gal4. For both females and males, ap > hINS^{C96Y} or ap > *sfl* RNAi expression alone reduces notum area and causes loss of dorsal macrochaetae. Co-expression of *sfl* RNAi and hINS^{C96Y} by ap-Gal4 results in greater destruction of the notum and macrochaetae in both sexes. However, in the male the notum and additional dorsal structures are absent and this phenotype is lethal.

ap-Gal4 > hINS^{C96Y} (A) female and (D) male;

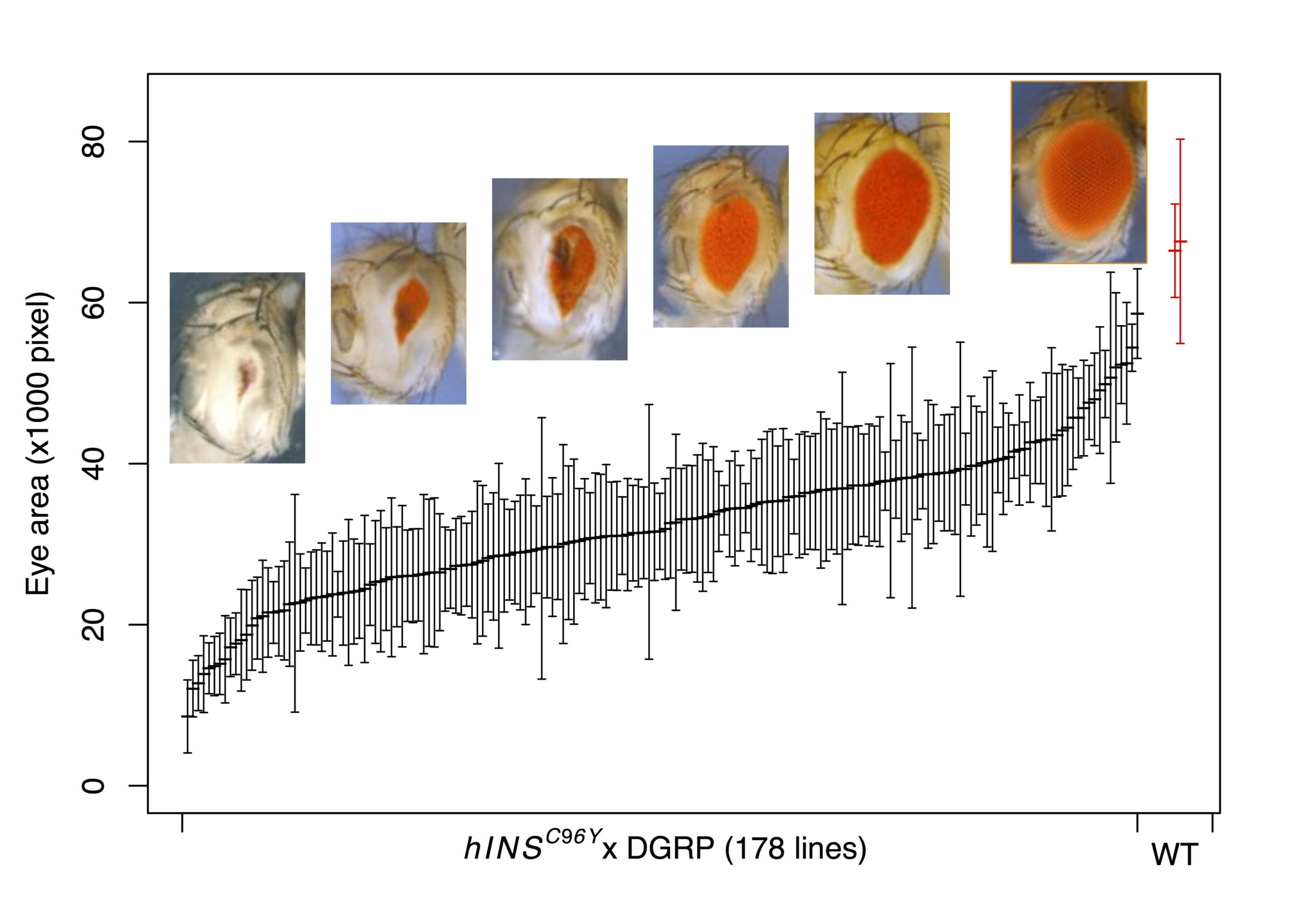
ap-Gal4 > *sfl* RNAi (B) female and (E) male;

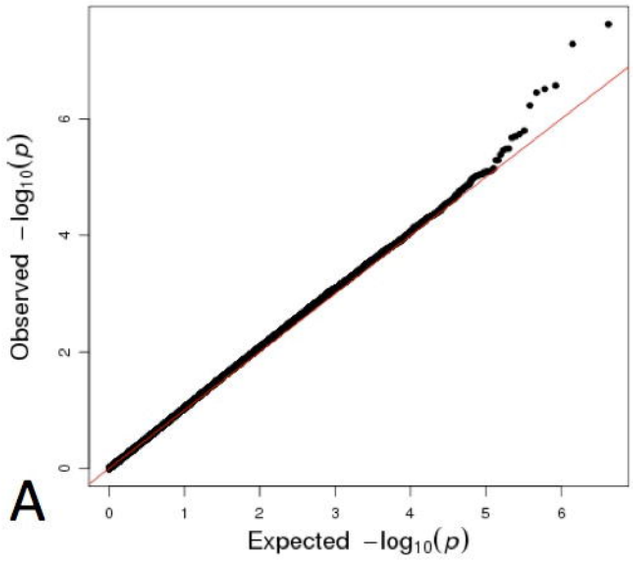
ap-Gal4>> hINS^{C96Y}, *sfl* RNAi (C) female (F) male

Figure S10 log2 transformed ratios between transcript levels associated with 18 bp/4 bp alleles. The allele-specific expression ratios were measured in each F1 heterozygote by pyro-sequencing, with three (or four) biological replicates and four (or three) pyro-technical replicates, to obtain a total of 12 measurements. In each of the 15 crosses, the

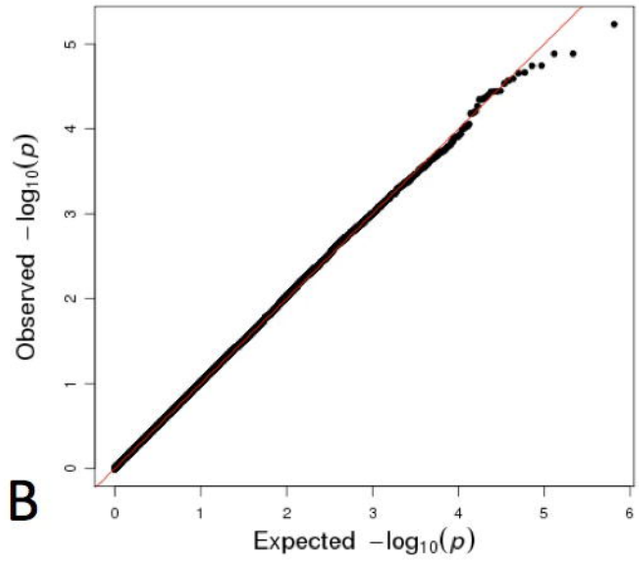
technical replicates were plotted in a single column, with different columns representing the biological replicates. In the titles of each panel, the last three digits in the stock number were shown for lines used in the cross.

Figure S11 Conditional regression analysis to detect additional SNPs associated with the phenotype of interest. (A) within the *sfl* locus; (B) all chromosomes. The intronic 18 pb/4 bp polymorphism in *sfl* is included in the linear model as a covariate. The two dotted lines in (A) correspond to a single test 0.05 level (red) and the multiple testing corrected 0.05 level using Bonferroni's method (blue). The red line in (B) represents the Bonferroni corrected 0.05 level.

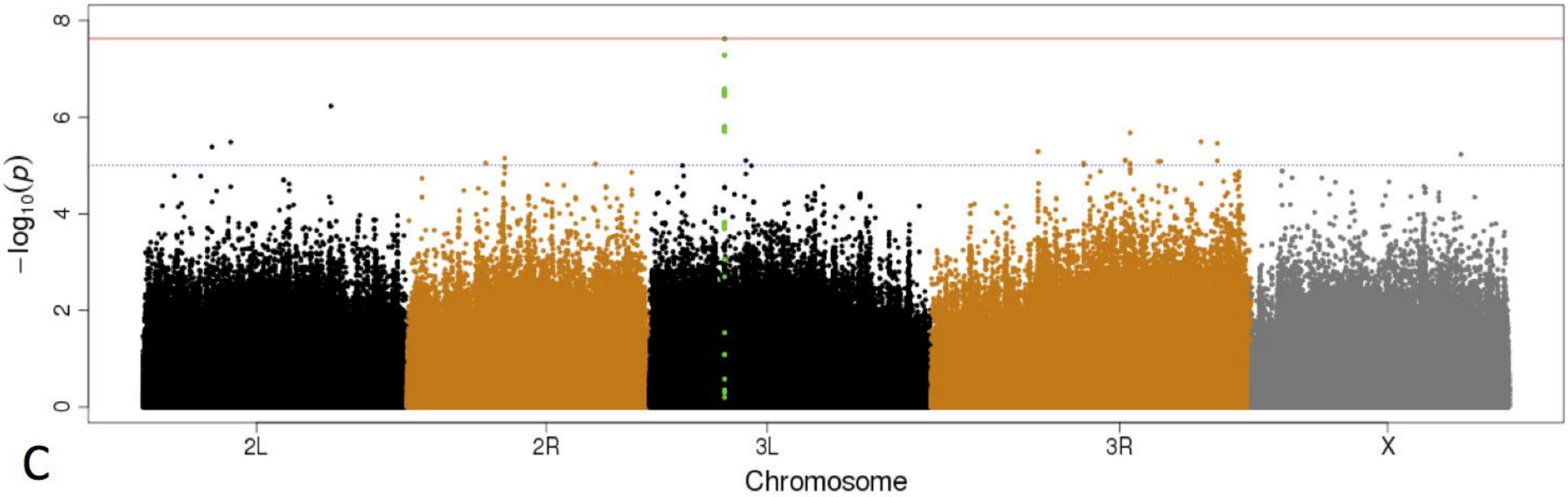

Eye area (x1000 pixel)
0
20
40
60
80
hINS^C96Y x DGRP (178 lines)
WT


Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

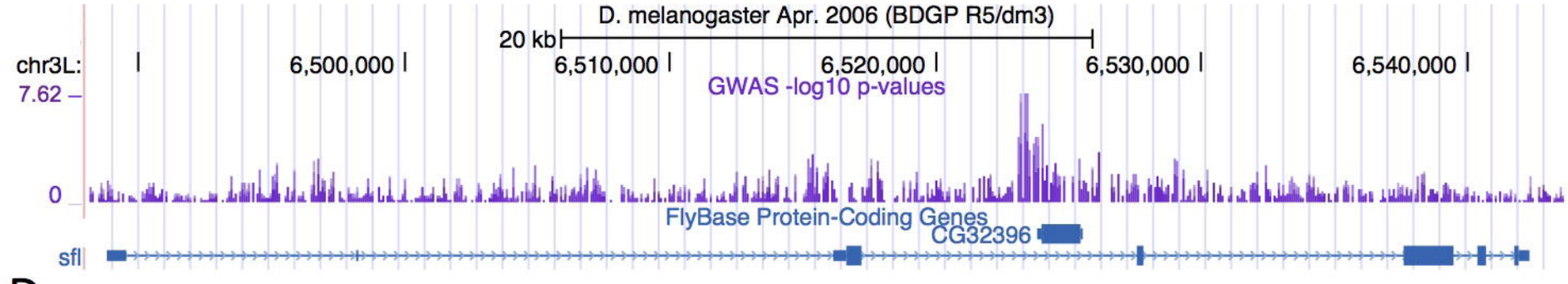
A


Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

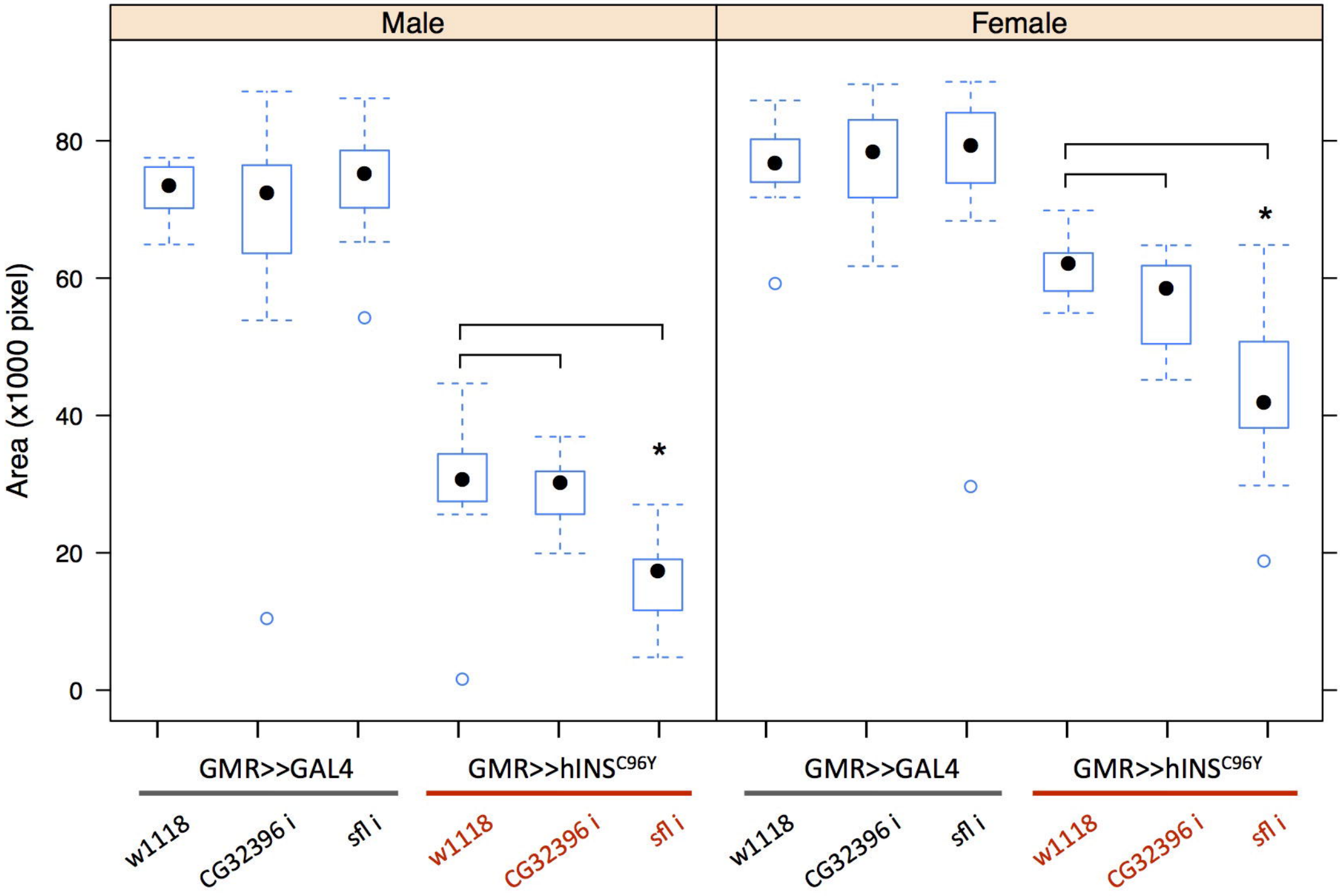
B

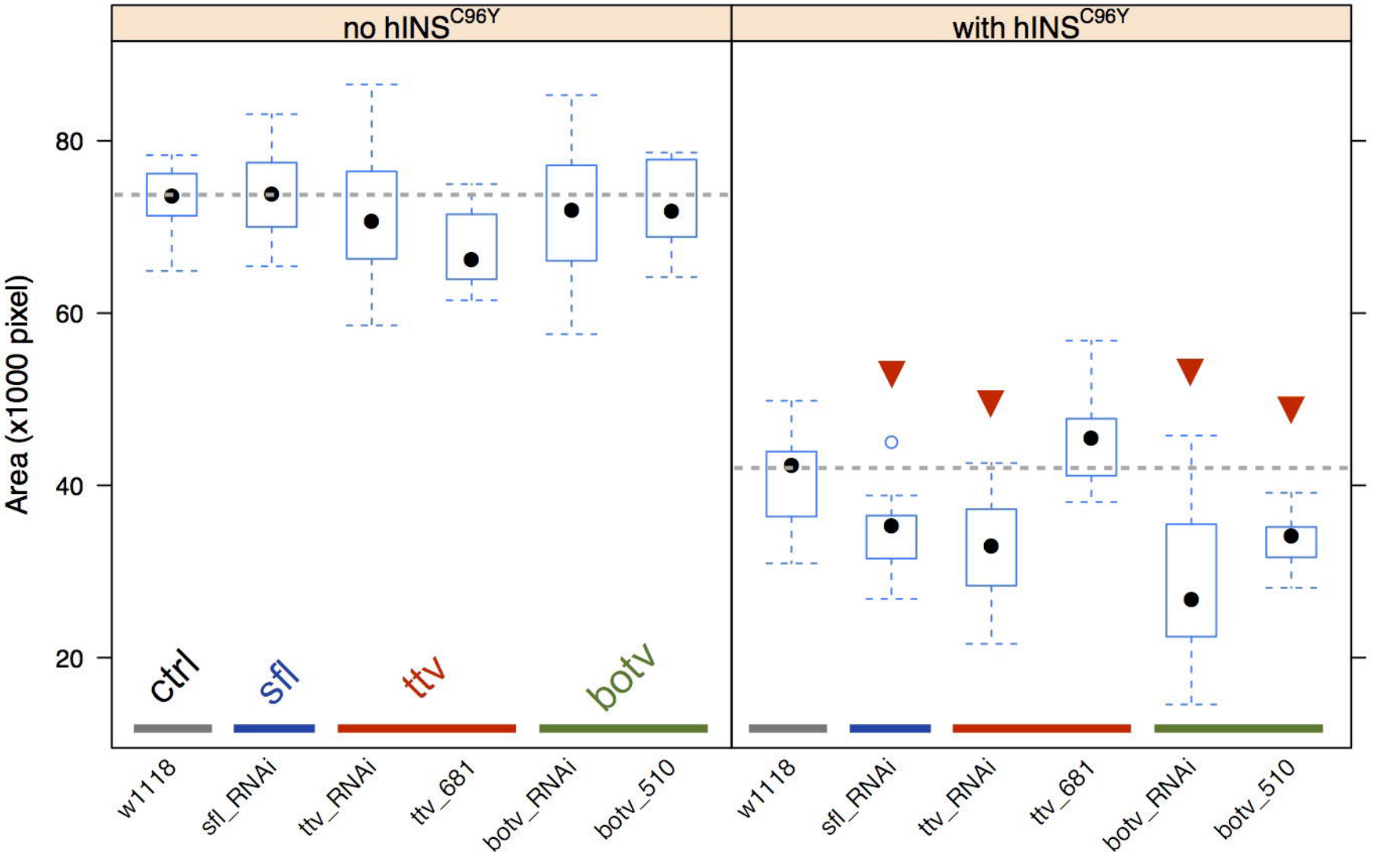

$-\log_{10}(p)$
2L
2R
3L
3R
X
Chromosome

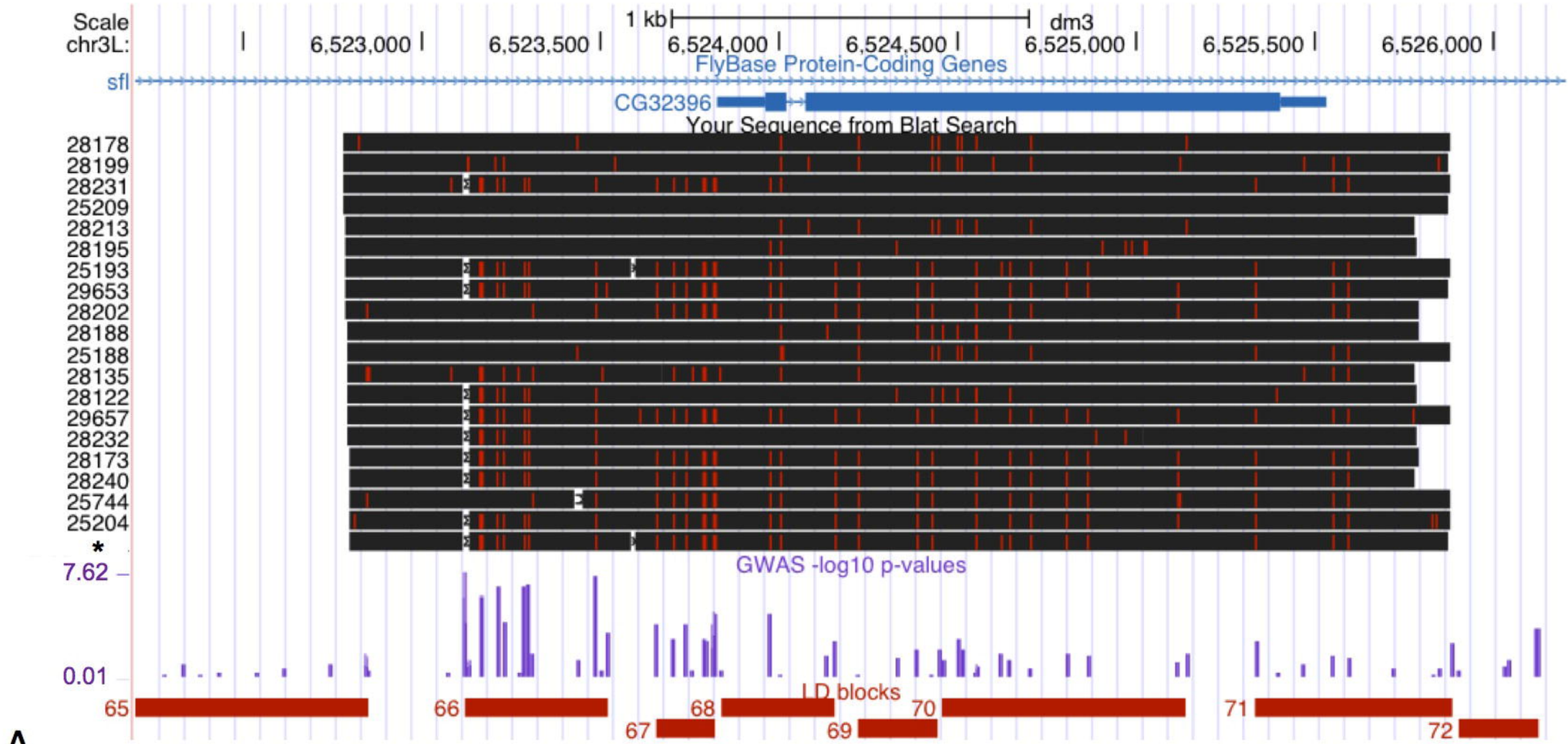
C


D. melanogaster Apr. 2006 (BDGP R5/dm3)
20 kb
chr3L:
6,500,000
6,510,000
6,520,000
6,530,000
6,540,000
7.62
GWAS -log10 p-values
0
FlyBase Protein-Coding Genes
CG32396
sfl

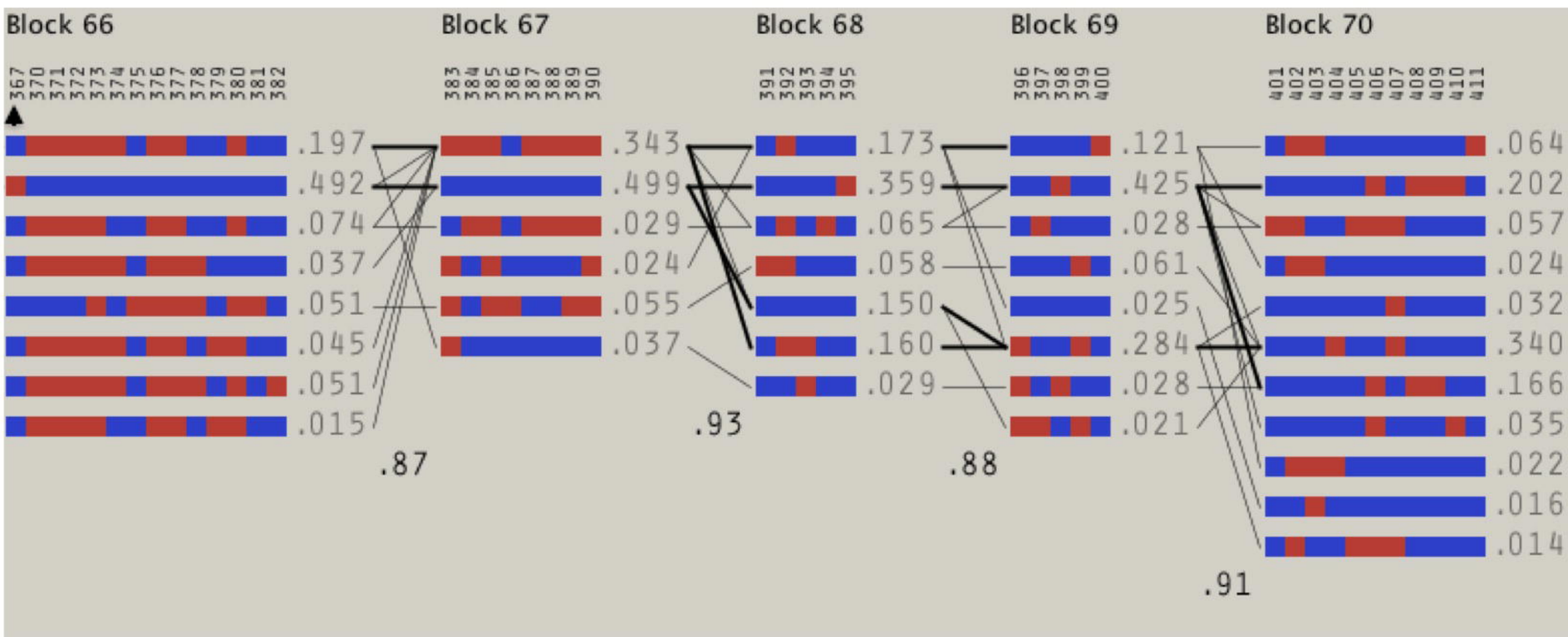
D


Male
Female
Area (x1000 pixel)
0
20
40
60
80
*
*
GMR>>GAL4
GMR>>hINS^C96Y
GMR>>GAL4
GMR>>hINS^C96Y
w1118
CG32396 i
sfl i
w1118
CG32396 i
sfl i
w1118
CG32396 i
sfl i
w1118
CG32396 i
sfl i

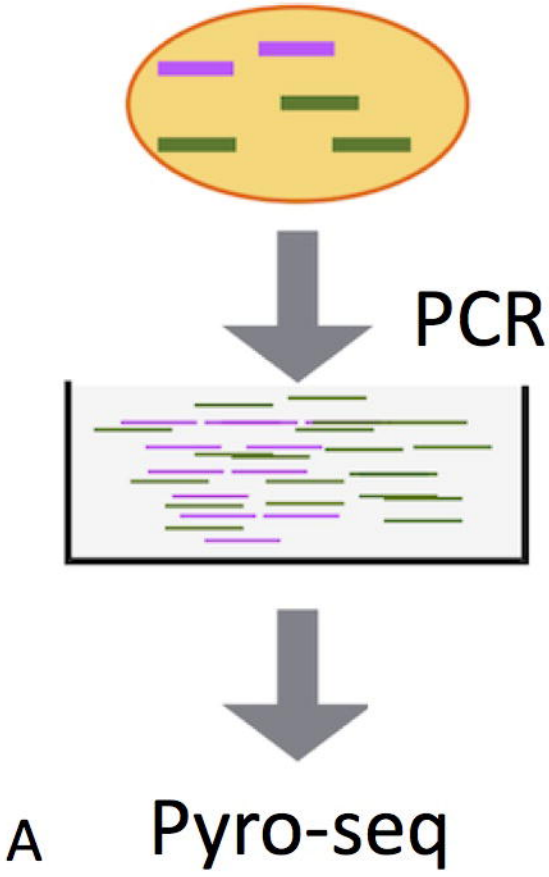

no hINS^C96Y
with hINS^C96Y
Area (x1000 pixel)
80
60
40
20
ctrl
sfl
ttv
botv
w1118
sfl_RNAi
ttv_RNAi
ttv_681
botv_RNAi
botv_510
w1118
sfl_RNAi
ttv_RNAi
ttv_681
botv_RNAi
botv_510


Scale
chr3L:
1 kb
dm3
6,523,000
6,523,500
6,524,000
6,524,500
6,525,000
6,525,500
6,526,000
FlyBase Protein-Coding Genes
sfl
CG32396
Your Sequence from Blat Search
28178
28199
28231
25209
28213
28195
25193
29653
28202
28188
25188
28135
28122
29657
28232
28173
28240
25744
25204
*
GWAS -log10 p-values
7.62
0.01
LD blocks
65
66
67
68
69
70
71
72

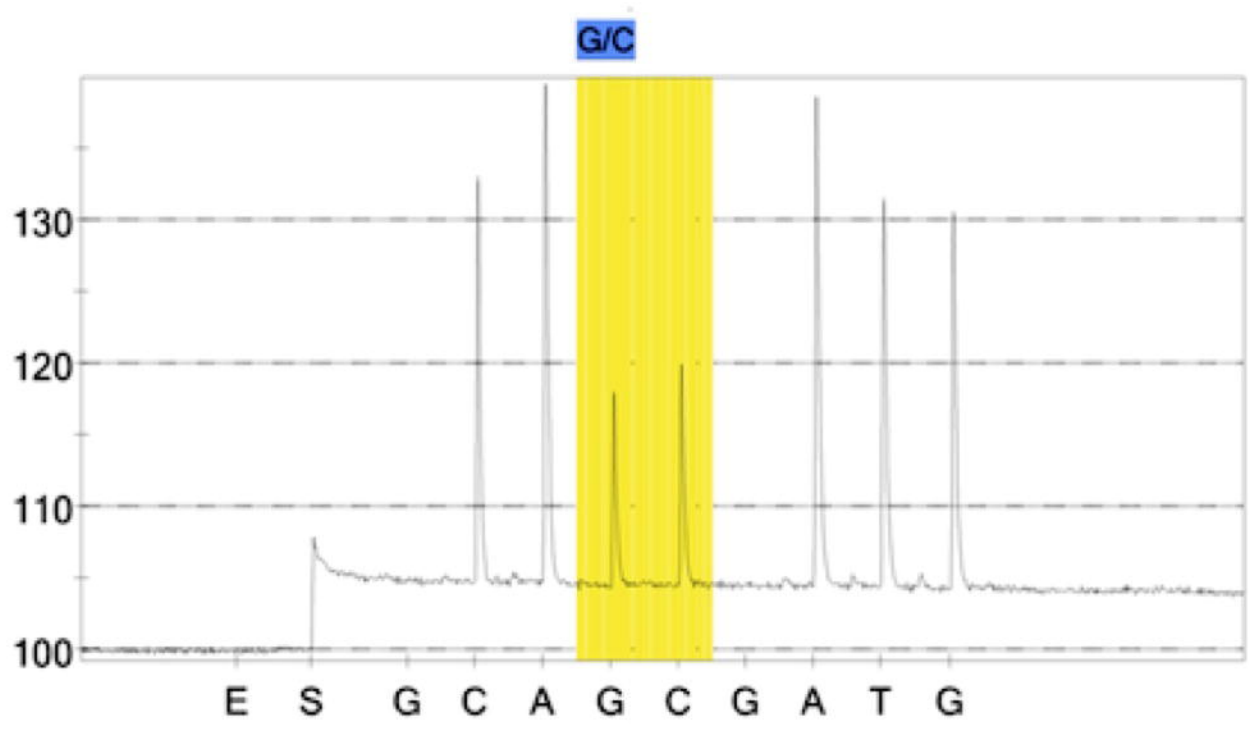
A


Block 66
Block 67
Block 68
Block 69
Block 70
.197
.492
.074
.037
.051
.045
.051
.015
.87
.343
.499
.029
.024
.055
.037
.93
.173
.359
.065
.058
.150
.160
.029
.88
.121
.425
.028
.061
.025
.284
.028
.021
.064
.202
.057
.024
.032
.340
.166
.035
.022
.016
.014
.91

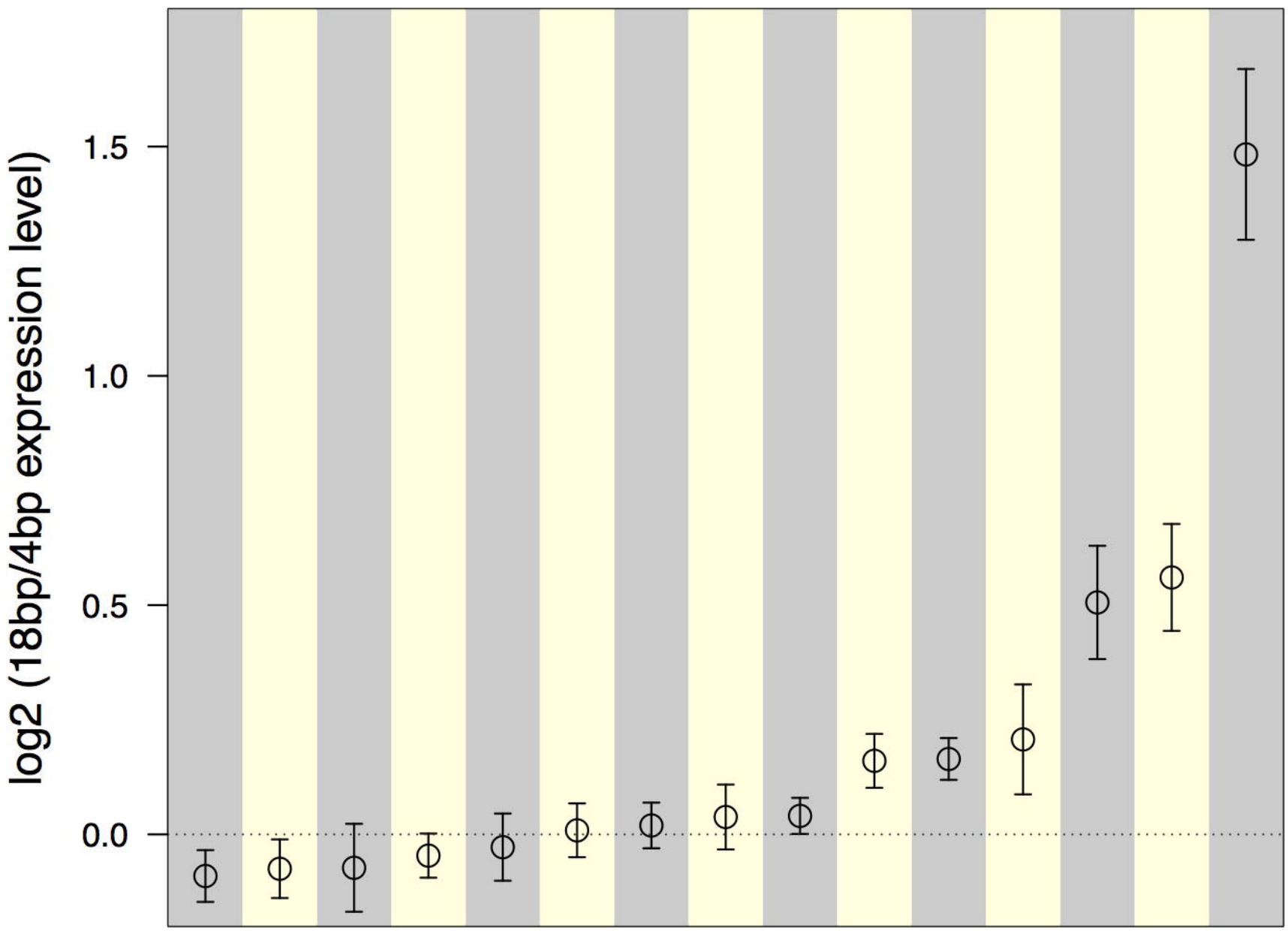
B


PCR
Pyro-seq

A


G/C
130
120
110
100
E S G C A G C G A T G

B


log2 (18bp/4bp expression level)
1.5
1.0
0.5
0.0
Cross

C