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,[†] Pengyao Jiang,* Soo-Young Park,[‡] Natalia A. Tamarina,[‡] Scott B. Selleck,[§] Patricia J. Wittkopp, Graeme I. Bell,[‡,††] and Martin Kreitman*,[2]**

*Department of Ecology and Evolution, [‡]Department of Medicine, and [††]Department of Human Genetics, The University of Chicago, Chicago, Illinois 60637, [†]Department of Epidemiology, and Department of Biostatistics, Harvard School of Public Health, Harvard University, Boston, Massachusetts 02115, [§]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, and **Department of Ecology and Evolutionary Biology, and Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, Michigan 48109

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 400-bp linkage block within an intron of the gene *sulfateless* (*sfl*). RNAi knockdown 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 heterozygotes. 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, and 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.

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 article (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.

doi: 10.1534/genetics.113.157800
Manuscript received September 21, 2013; accepted for publication November 21, 2013; published Early Online November 26, 2013.
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157800/-/DC1.
[1]Present address: FAS Center for Systems Biology, Harvard University, 52 Oxford St., Cambridge, MA 02138.
[2]Corresponding authors: Department of Ecology and Evolution, The University of Chicago, 1101 E. 57th St., Chicago, IL 60637-1573. E-mail: martinkreitman@gmail.com; and FAS Center for Systems Biology, Harvard University, 52 Oxford St., Cambridge, MA 02138. E-mail: binhe@fas.harvard.edu

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 hundreds of base pairs, 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 second chromosome, which was balanced over CyO. DGRP lines were obtained from the Bloomington Stock Center. RNAi lines against *sfl* (GD5070), *ttv* (GD4871), and *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°. 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 are available in Supporting Information, Table S1.

Principal Component Analysis

The whole-genome SNP data set for the 154 DGRP lines used for genome-wide association study (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 v. 1.07, with parameter–indep-pairwise 50 5 0.5) were used to identify the top 15 principal components (PCs) (SmartPCA software in Eigensoft v. 3.0, 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). On the X chromosome, 1,616,121 autosomal and 256,948 SNPs 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 (Figure S4). To assess the genome-wide significance threshold while accounting for both the relatedness structure in the data as well as the nonindependence between SNPs due to LD, we performed a permutation procedure (details in File 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, an 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 (v. 1.0) to estimate the genetic relatedness matrix with all SNPs with minor allele frequency >5%

(–MAF 0.05). We then used the restricted maximum-likelihood (REML) 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 third-instar larvae. The individual larva was sexed and dissected in 1× 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 endogenous reference (SYBR-Green assay). Primers used for qRT–PCR are listed in Table S3.

RNAi and validation studies

All RNAi lines were originally from the Vienna Drosophila RNAi Center as *P*-element insertion lines on a co-isogenic w1118 background. Each RNAi line was first tested to determine whether it alone had an effect on eye development by crossing it to GMR–Gal4 and comparing the eye area of the F1 males (or females) to the control cross between w1118 and GMR–Gal4. In all crosses, GMR–Gal4 was used as the maternal parent. To test its effect on the $hINS^{C96Y}$-induced eye-degeneration phenotype, the RNAi line was crossed to the GMR >> $hINS^{C96Y}$ line (used as maternal parent), so that both $hINS^{C96Y}$ and the RNAi constructs are driven by GMR–Gal4. The resulting phenotype was compared to the cross between $hINS^{C96Y}$ females and w1118 males. At least 10 individual flies were measured per cross and a *t*-test was used to determine significance at 0.05 level with multiple testing correction. For mutant lines, GMR–Gal4 was replaced with w1118 in the first test and used as a control. The same scheme was used for the second test. It is worth noting that because the mutants were tested in heterozygous states, only dominant interaction with $hINS^{C96Y}$ are revealed.

sfl expression studies

Six lines carrying the 18-bp indel allele and eight carrying the 4-bp allele were chosen and paired to form 15 crosses (Figure S1A). Three sets of 10 late third-instar (wandering stage) larvae were collected from each cross and dissected in 1× 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 that could be used to distinguish the two alleles in each cross was identified (Figure S1B). Four such SNPs were chosen and pyrosequencing assays were designed (primers listed in Table S3). Pyrosequencing was performed as previously described (Wittkopp 2011). Briefly, each of the three cDNA and one gDNA sample per cross were 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 $y_{ij} = \alpha + L_i + \varepsilon_{ij}$, 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 $y_i = \alpha + \varepsilon_i$, 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 10 male progeny from each $hINS^{C96Y}$ × 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, p. 115)], 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, nor 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

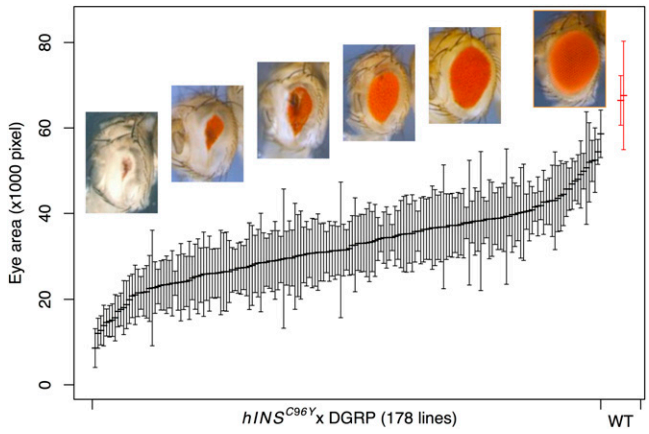

Figure 1 Distribution of eye area in hINS^{C96Y} × DGRP crosses. Mean ±1 SD, sorted by the mean, is shown for crosses between the transgenic {GMR >> hINS^{C96Y}} line to 178 DGRP lines, and two randomly chosen DGRP inbred lines (red). Representative photographs of eyes from across the range of the distribution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

consistent with Gal4 protein measurements and the fact that the eye-degeneration phenotype is insensitive to GMR–Gal4 gene dose when hINS^{C96Y} is present in single copy (Park *et al.* 2013, Figure 3). Finally, when we expressed hINS^{C96Y} in the notum (rather than the eye) and measured the loss of macrochaetae in F1 crosses to 38 DGRP lines for which we also collected eye-degeneration data, we observed no correlation between the two traits, indicating that the degeneration phenotypes are not caused by line-specific differences in mutant insulin expression (Park *et al.* 2013).

Genome-wide association analysis

To identify candidate genetic loci and variants underlying the phenotypic variation, we carried out GWAS on the F1 males from the crosses of hINS^{C96Y} and 154 DGRP lines. We used mean eye area as a quantitative trait to perform single-marker regression for 1.6 million autosomal SNPs, restricted to biallelic sites for which the minor allele frequency is at least 5%. The result revealed a strong peak on chromosome 3L and minor ones on other major chromosome arms (Figure 2C). The most significant SNP underlying the chromosome 3L peak has a raw *P*-value of 2.4×10^{-8} (*t*-test); the Bonferroni-corrected $P = 0.04$.

Population stratification is a potential confounder for GWAS—it can inflate the test statistic for nonassociated variants if the population structure correlates with the phenotype. We assessed its impact in our study in three ways. First, we evaluated the Q–Q plots for autosomal and X-linked variants. Neither showed a systematic shift toward low *P*-values compared to the null expectation, which would be expected if population structure induces false association signals (Figure 2, A and B). 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* and 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 found only an excess of low *P*-values in autosomal variants, but not in X-linked ones (Figure 2, A and B). The above analyses suggest 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 nonindependence and larger than expected phenotypic variance (Voight and Pritchard 2005; Cheng *et al.* 2010). We estimated the GRM from whole-genome SNP data using mixmogam (a Python implementation of EMMAX) (Kang *et al.* 2010; Segura *et al.* 2012). We found that 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 (Yu *et al.* 2006; Atwell *et al.* 2010). A permutation procedure specifically designed to preserve the phenotype covariance structure is used to establish a genome-wide 5% significance threshold (File 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 focus on identifying the gene(s) underlying the peak and genetically testing its association with the phenotype.

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 β-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 and Figure 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 and Figure 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*

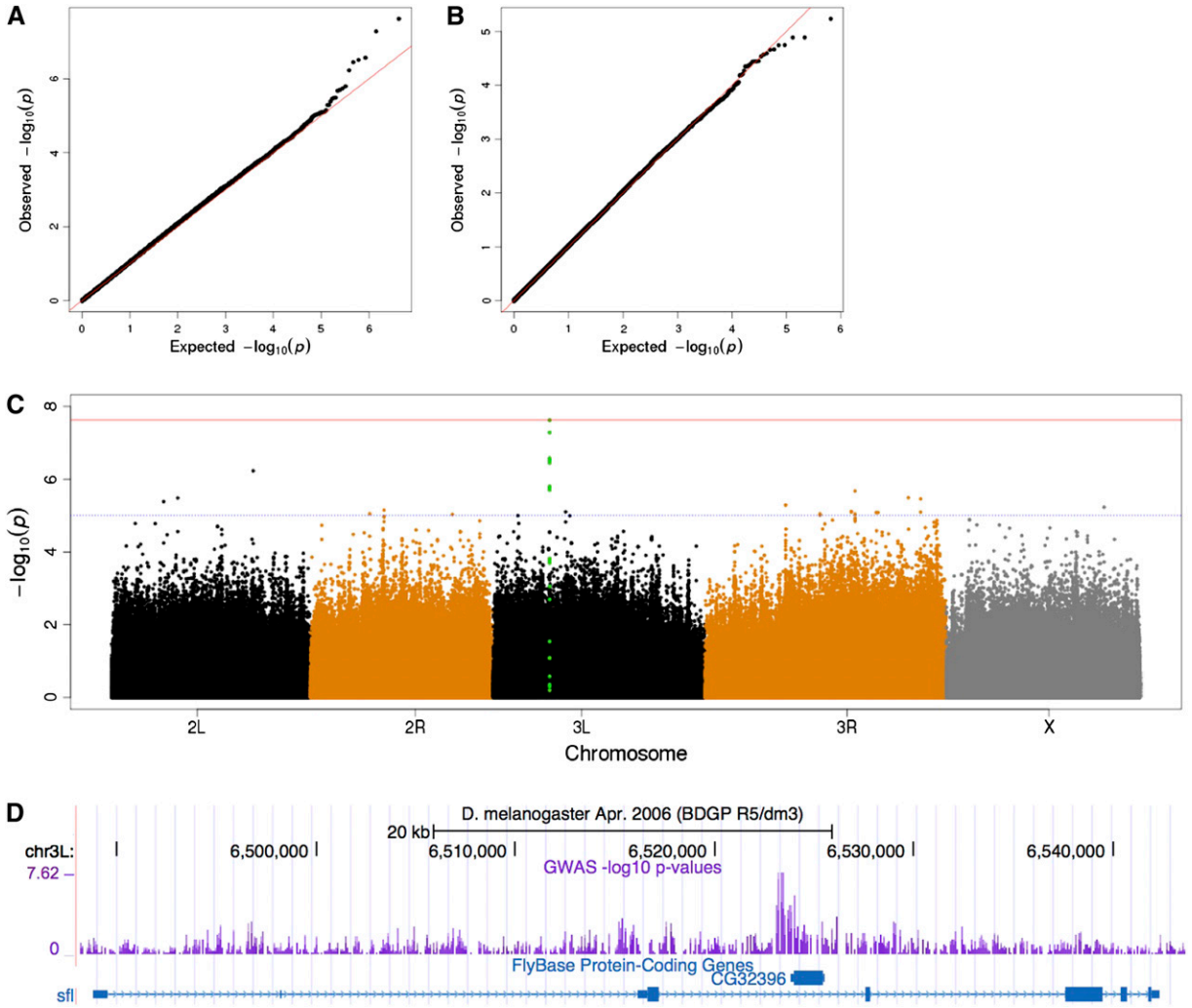

Figure 2 Genome-wide scan identifies candidate locus associated with the hINS^{C96Y}-induced phenotype. Quantile–quantile (Q–Q) plot reveals an excess of small *P*-values on autosomes (A) but not on the X chromosome (B), which is not variable in the mapping population due to cross design. (C) Manhattan 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.

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 and Figure S9). However, the interpretation is complicated by the fact that *sfl* knockdown alone causes mutant phenotypes in these tissues, consistent with previous knowledge (Lin 2004). At present we cannot distinguish the alternative hypotheses of additive *vs.* epistatic interactions between *sfl* and hINS^{C96Y}.

Heparan sulfate biosynthetic pathway modifies the hINS^{C96Y}-induced eye degeneration

Sulfateless encodes a bifunctional enzyme in the heparin sulfate biosynthesis pathway. An important component of the cell surface and extracellular matrix (Kirkpatrick and Selleck 2007), HSPGs regulate signaling during development by influencing the levels and activity of growth factors and morphogens at cell surfaces and in the extracellular matrix (Nakato *et al.* 1995; Häcker *et al.* 1997; Giráldez *et al.* 2002; Fujise *et al.* 2003; Kirkpatrick *et al.* 2004). The involvement of HSPGs in the cellular responses to misfolded proteins (proteostasis) has not been previously described.

To further examine the hINS^{C96Y}-dependent interaction of *sfl*, we examined RNAi knockdowns and mutants for two additional genes in the HS biosynthetic pathway, *ttv* and *botv*, producing the glycosaminoglycan polymer that is modified by *sfl* (Lin 2004). SNPs in neither of the genes showed evidence of association in our GWAS (lowest adjusted $P >$ 0.5 in both loci, adjusted for multiple-testing using Bonferroni's method). RNAi knockdown of both genes shows a hINS^{C96Y}-dependent effect on eye area in the same direction as *sfl* RNAi (Figure 4). In addition, a mutant allele of *botv* also showed a 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.

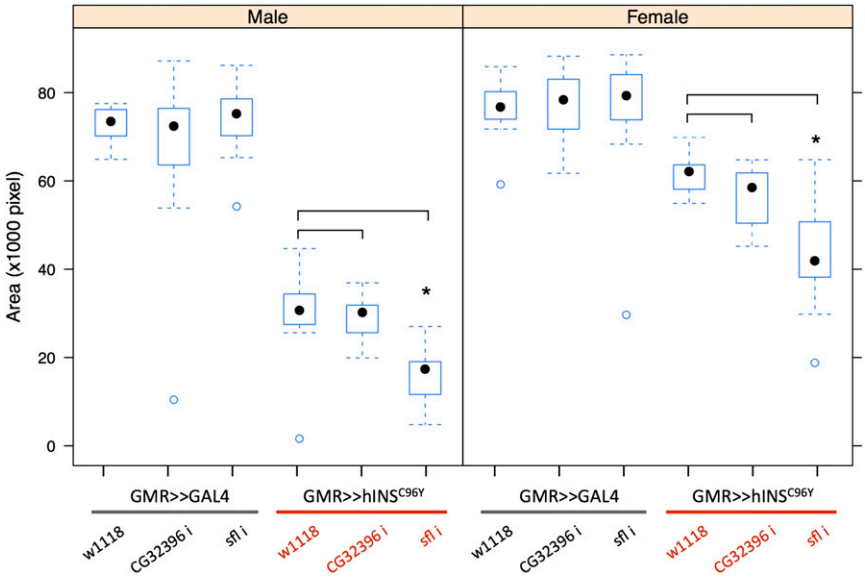

Figure 3 RNAi knockdown confirms *sfl* and excludes CG32396 as the causal gene. The effect of knocking down either CG32396 or *sfl* was tested in the absence ({UAS–RNAi} × {GMR–Gal4}) or presence ({UAS–RNAi} × {GMR–Gal4, UAS–hINS^{C96Y}}) of hINS^{C96Y}. Compared to the control crosses (first and third 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 the interquartile range (whiskers) are indicated; data points outside the range are represented by circles.

Intronic variation and sfl expression

We resequenced 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 *Drosophila simulans* orthologous sequence) (Figure 5A). We also found three other insertion/deletion (INDEL) polymorphisms in this region, with sizes ranging from 4 to 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 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 v. 4.2) 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 18-bp or 4-bp allele, although it is worth noting that we do not 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- and 18-bp lines from the respective phenotypic spectrum, crossed them to obtain F1 individuals heterozygous for the two alleles, and used pyrosequencing 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 to 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 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-/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

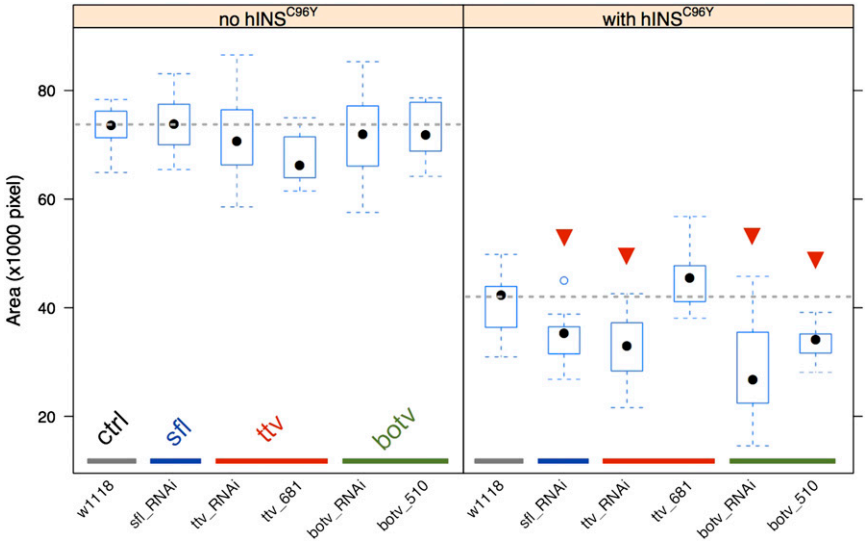

Figure 4 RNAi and mutant analysis for heparin sulfate biosynthesis pathway genes. The experimental design is the same as in Figure 3. Left: the effect of RNAi or mutant alleles in the absence of hINS^{C96Y} expression. Right: the effect when hINS^{C96Y} is expressed in the eye imaginal disc. Mutants were tested in heterozygous states for a dominant interaction with hINS^{C96Y}. Fifteen male flies are measured for each group. The statistical significance of differences from the control cross (gray, w1118) was determined by a two-sided student's *t*-test. Those that are significant at 0.05 level after Bonferroni correction are marked with a red arrowhead.

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- 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 (1.6 million), 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 that the *sfl* QTL achieves statistical significance is because the two alternative alleles occur at 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 in both 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 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 18- and 4-bp *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-/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 (18-/4-bp ratio = 0.94 and 0.96); and 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, although it remains a possibility. Future experiments employing genome-editing technologies will allow better resolution of the mechanism(s) underlying the association (Jinek *et al.* 2012; Gratz *et al.* 2013; 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 *Drosophila melanogaster*, with many eQTLs contributing to expression variability both in *cis* and in *trans* (Brem *et al.* 2002, 2005; West *et al.* 2007), and intralocus genetic complexity influencing a quantitative trait has long been known, as in the Adh example (King *et al.*

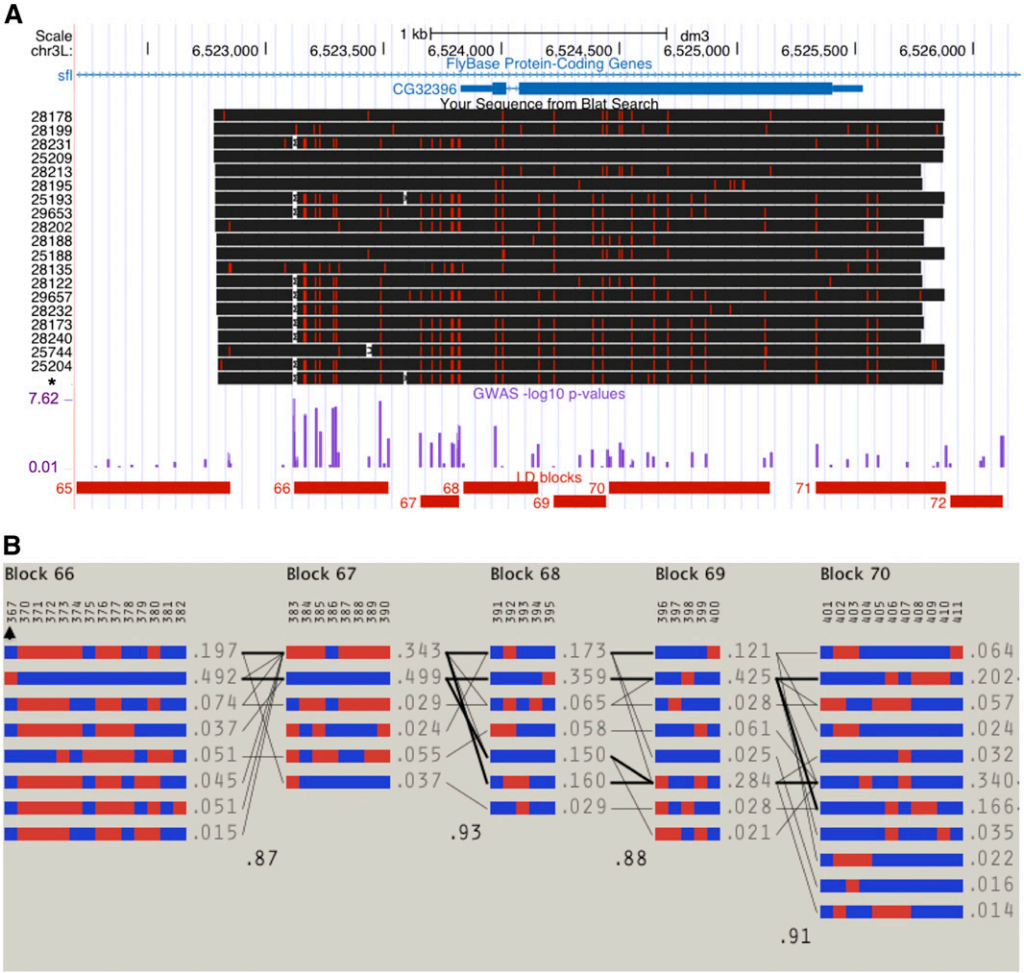

Figure 5 Sequencing of a 3-kb region in *sfl* and the LD patterns in the region. (A) Alignment of 19 DGRP sequences ordered by their eye-degeneration phenotype (mean, most severe on the bottom). The hINS^{C96Y} transgenic line (asterisk) was also sequenced. Red ticks and white spaces indicate SNPs and deletions relative to the reference sequence. No insertions relative to the reference were found. The purple track shows the $-\log_{10}$ of GWAS *P*-values. The bottom track shows the linkage blocks as determined by Haploview (4.02) using the solid spine method with default settings ($D' > 0.8$). (B) Detailed haplotype block structures. Each numbered column represents a polymorphic site, with the alleles colored as blue or red; each row represents a haplotype with frequency >0.01. An arrowhead marks the 18-/4-bp indel polymorphism (see text; 18 bp, blue; 4 bp, red). Finally, the number between any two blocks represents the multiallelic *D'*, which quantifies the associations between adjacent blocks.

2012). In the 40-kb region spanning the *sfl* locus alone, 1358 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 (HSPGs) are abundant components of cell surfaces and extracellular matrices and are best understood for their roles in cell signaling and in functioning as coreceptors, processes integral to normal development (Häcker *et al.* 2005; Kirkpatrick and Selleck 2007). HSPGs are also involved in endocytosis (Ren *et al.* 2009; Stanford *et al.* 2009) and vesicular trafficking (Nybakken and Perrimon 2002; Sarrazin *et al.* 2011), roles that may link them to cellular response to misfolded proteins (Higashio and 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.* (2012) found an enrichment of vesicle-trafficking-related genes among those that changed expression significantly after induction of ER stress. Both studies indicate that a global regulation of vesicle trafficking is important to a cell's response to unfolded or

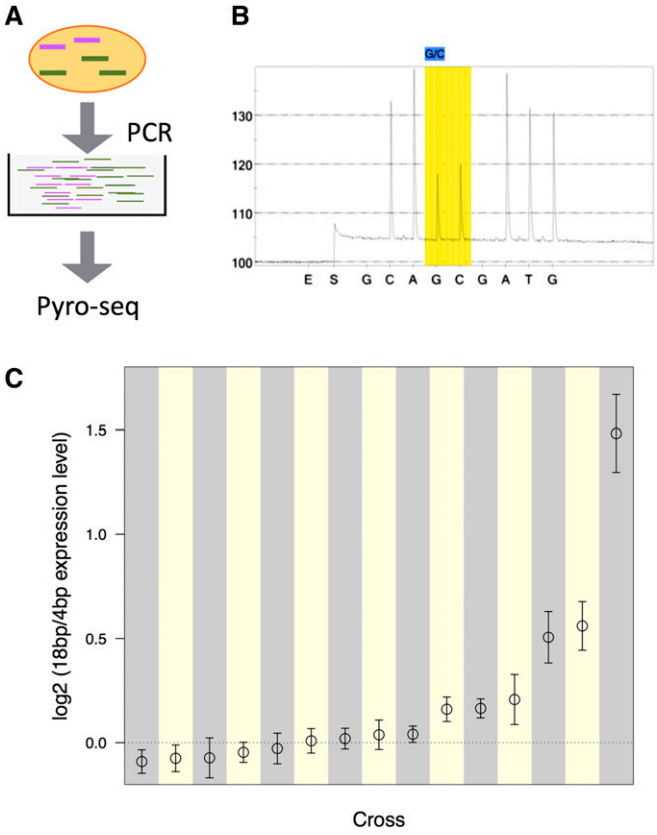

Figure 6 Pyro-sequencing measure of *sfl* allele-specific transcript ratio in 18-/4-bp heterozygotes. (A) Schematic diagram of the pyrosequencing approach. Colored lines represent transcripts (mRNA) associated with either the 18 or the 4bp allele, expressed at different levels. Common primers were used to amplify both transcripts of the gene of interest from the cDNA library made from eye imaginal disc tissues. Pyrosequencing was carried out on the amplified products. (B) A pyrogram of a heterozygote with the polymorphic site (G/C) that is diagnostic for the 18-/4-bp indel highlighted. The ratio of the two peaks (light intensity, *y*-axis) are used to calculate the relative ratio of the two alleles. (E, enzyme; S, substrate; A/C/G/T, nucleotides). (C) Log2-transformed ratio of 18-/4-bp allele expression in 15 crosses between randomly paired 18- and 4-bp lines. Estimates of the ratio and 95% confidence intervals are plotted. The dotted line corresponds to equal expression from the two alternative alleles.

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 and 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-/4-bp indel polymorphism in the *sfl* intron in the sample of 154 crosses. Any variant with a smaller effect size and/or lower frequency than the 18-/4-bp polymorphism would likely have been missed in this study.

Sulfateless was the only QTL identified as genome-wide significant in this study (Figure 2 and Figure S4); its association with the trait is robust with respect to population structure and cryptic relatedness (Figure S3 and Figure S4). This does not mean, however, that the genetic architecture for the hINS^{C96Y}-induced eye phenotype involves a single locus. Rather, we have several reasons to believe that the genetic architecture must involve many loci. First, the distribution of the phenotype, *i.e.*, eye areas expressed as line means, suggests a non-Mendelian genetic basis (Figure 1). Second, while ANOVA estimates that nearly 60% of the total phenotypic variance is between crosses, <20% within the 60% (*i.e.*, <12% of the total variance) can be attributed to the *sfl* locus. Even this 20% estimate, because it is derived from the same population used to identify the locus, is liable to be an overestimate due to the Winner's curse effect (Garner 2007).

To estimate what percentage of the between-cross variance can be explained by the additive effects of common variants combined, we applied the GCTA tool, which uses a mixed linear model method, to the line means of the 154 crosses (Yang *et al.* 2011). The result showed that 83% (standard error 37%) of the variance between crosses could be attributed to common, autosomal variants with minor allele frequencies >5%. Analysis using GEMMA (v. 0.94beta), which used a Bayesian method, achieved nearly identical results (posterior mode 0.83, SE 0.41). We then did the same analysis with GCTA, but including the 18-/4-bp indel polymorphism as a covariate to remove the effect of *sfl*, to estimate the remaining additive heritability. As a result, we got 62% (SE 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 humans.

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 iseases, 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 lifestyle 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 article 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 differences 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 and 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 20–40× 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

We thank Dan Nicolae for technical help and advice on GWAS and Xiang Zhou and Matthew Stephens for advice on the mixed linear model approach using Gemma. We thank Jian Yang and Peter Visscher for help with the interpretation of the GCTA results. Joseph Coolon in the Wittkopp lab helped design the pyro-sequencing assays, and Ellen Pederson at the DNA sequencing center at the University of Michigan provided technical assistance. We also thank the anonymous reviewer and Dr. Sabatti for helpful comments. 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. S.B.S. is supported by GM054832 and P.J.W. is supported by National Science Foundation MCB-1021398.

Note added in proof: See Park *et al.* 2014 (pp. 539–555) in this issue for a related work.

Literature Cited

Atwell, S., Y. S. Huang, B. J. Vilhjalmsson, G. Willems, M. Horton *et al.*, 2010 Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627–631.

Bilen, J., and N. M. Bonini, 2005 Drosophila as a model for human neurodegenerative disease. Annu. Rev. Genet. 39: 153–171.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752–755.

Brem, R. B., J. D. Storey, J. Whittle, and L. Kruglyak, 2005 Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436: 701–703.

Chandler, C. H., S. Chari, and I. Dworkin, 2013 Does your gene need a background check?: how genetic background impacts the analysis of mutations, genes, and evolution. Trends Genet. 29: 358–366.

Cheng, R., J. E. Lim, K. E. Samocha, G. Sokoloff, M. Abney *et al.*, 2010 Genome-wide association studies and the problem of relatedness among advanced intercross lines and other highly recombinant populations. Genetics 185: 1033–1044.

Chintapalli, V. R., J. Wang, and J. A. T. Dow, 2007 Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat. Genet. 39: 715–720.

Downes, C. P., A. Gray, and J. M. Lucocq, 2005 Probing phosphoinositide functions in signaling and membrane trafficking. Trends Cell Biol. 15: 259–268.

Fujise, M., S. Takeo, K. Kamimura, T. Matsuo, T. Aigaki, S. Izumi *et al.*, 2003 Dally regulates Dpp morphogen gradient formation in the Drosophila wing. Development 130: 1515–1522.

Garner, C., 2007 Upward bias in odds ratio estimates from genome-wide association studies. Genet. Epidemiol. 31: 288–295.

Gibson, G., 2009 Decanalization and the origin of complex disease. Nat. Rev. Genet. 10: 134–140.

Giráldez, A. J., R. R. Copley, and S. M. Cohen, 2002 HSPG modification by the secreted enzyme Notum shapes the Wingless morphogen gradient. Dev. Cell 2: 667–676.

Gonzalez, C., 2013 *Drosophila melanogaster*: a model and a tool to investigate malignancy and identify new therapeutics. Nat. Rev. Cancer 13: 172–183.

Gratz, S. J., A. M. Cummings, J. N. Nguyen, D. C. Hamm, L. K. Donohue *et al.*, 2013 Genome Engineering of *Drosophila* with the CRISPR RNA-Guided Cas9 Nuclease. Genetics 194: 1029–1035.

Häcker, U., X. Lin, and N. Perrimon, 1997 The Drosophila sugarless gene modulates Wingless signaling and encodes an enzyme involved in polysaccharide biosynthesis. Development 124: 3565–3573.

Häcker, U., K. Nybakken, and N. Perrimon, 2005 Heparan sulphate proteoglycans: the sweet side of development. Nat. Rev. Mol. Cell Biol. 6: 530–541.

Higashio, H., and K. Kohno, 2002 A genetic link between the unfolded protein response and vesicle formation from the endoplasmic reticulum. Biochem. Biophys. Res. Commun. 296: 568–574.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, 2012 A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science 337: 816–821.

Kang, H. M. M., J. H. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42: 348–354.

Kim, J.-H. H., Y. Zhao, X. Pan, X. He, and H. F. Gilbert, 2009 The unfolded protein response is necessary but not sufficient to compensate for defects in disulfide isomerization. J. Biol. Chem. 284: 10400–10408.

Kimmig, P., M. Diaz, J. Zheng, C. C. Williams, A. Lang *et al.*, 2012 The unfolded protein response in fission yeast modulates stability of select mRNAs to maintain protein homeostasis. eLife 1: e00048

King, E. G., C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen, K. W. Broman *et al.*, 2012 Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res. 22: 1558–1566.

Kirkpatrick, C. A., and S. B. Selleck, 2007 Heparan sulfate proteoglycans at a glance. J. Cell Sci. 120: 1829–1832.

Kirkpatrick, C. A., B. D. Dimitroff, J. M. Rawson, and S. B. Selleck, 2004 Spatial regulation of Wingless morphogen distribution and signaling by Dally-like protein. Dev. Cell 7: 513–523.

Kolde, R., 2011 pheatmap: Pretty Heatmaps. http://cran.r-project.org/package=pheatmap.

Lin, X., 2004 Functions of heparan sulfate proteoglycans in cell signaling during development. Development 131: 6009–6021.

Listgarten, J., C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, and D. Heckerman, 2012 Improved linear mixed models for genome-wide association studies. Nat. Methods 9: 525–526.

Mackay, T. F. C., E. A. Stone, and J. F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. Nat. Rev. Genet. 10: 565–577.

Mefford, H. C., A. J. Sharp, C. Baker, A. Itsara, Z. Jiang *et al.*, 2008 Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N. Engl. J. Med. 359: 1685–1699.

Nakato, H., T. A. Futch, and S. B. Selleck, 1995 The division abnormally delayed (dally) gene: a putative integral membrane proteoglycan required for cell division patterning during postembryonic development of the nervous system in Drosophila. Development 121: 3687–3702.

Nybakken, K., and N. Perrimon, 2002 Heparan sulfate proteoglycan modulation of developmental signaling in Drosophila. Biochim. Biophys. Acta 1573: 280–291.

Park, S.-Y., M. Z. Ludwig, N. A. Tamarina, B. Z. He, S. Carl *et al.*, 2013 Genetic complexity in a *Drosophila* model of diabetes-associated misfolded human proinsulin. Genetics 539–555.

Ran, F. A., P. D. Hsu, C.-Y. Lin, J. S. Gootenberg, S. Konermann *et al.*, 2013 Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell 154: 1380–1389.

Ren, Y., C. A. Kirkpatrick, J. M. Rawson, M. Sun, and S. B. Selleck, 2009 Cell type-specific requirements for heparan sulfate biosynthesis at the Drosophila neuromuscular junction: effects on synapse function, membrane trafficking, and mitochondrial localization. J. Neurosci. 29: 8539–8550.

Roy, S., J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton *et al.*, 2010 Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330: 1787–1797.

Sarrazin, S., W. C. Lamanna, and J. D. Esko, 2011 Heparan sulfate proteoglycans. Cold Spring Harb. Perspect. Biol. DOI: 10.1101/cshperspect.a004952

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

Stanford, K. I., J. R. Bishop, E. M. Foley, J. C. Gonzales, I. R. Niesman *et al.*, 2009 Syndecan-1 is the primary heparan sulfate proteoglycan mediating hepatic clearance of triglyceride-rich lipoproteins in mice. J. Clin. Invest. 119: 3236–3245.

Støy, J., E. L. Edghill, S. E. Flanagan, H. Ye, V. P. Paz *et al.*, 2007 Insulin gene mutations as a cause of permanent neonatal diabetes. Proc. Natl. Acad. Sci. USA 104: 15040–15044.

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

Wang, J., T. Takeuchi, S. Tanaka, S. K. Kubo, T. Kayo *et al.*, 1999 A mutation in the insulin 2 gene induces diabetes with severe pancreatic beta-cell dysfunction in the Mody mouse. J. Clin. Invest. 103: 27–37.

West, M. A., K. Kim, D. J. Kliebenstein, H. van Leeuwen, R. W. Michelmore *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 Mol. Biol. 772: 297–317.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88: 76–82.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

Communicating editor: C. Sabatti

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157800/-/DC1

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

Bin Z. He, Michael Z. Ludwig, Desiree A. Dickerson, Levi Barse, Bharath Arun, Bjarni J. Vilhjálmsson, Pengyao Jiang, Soo-Young Park, Natalia A. Tamarina, Scott B. Selleck, Patricia J. Wittkopp, Graeme I. Bell, and Martin Kreitman

DOI: 10.1534/genetics.113.157800

A

	18bp	28190	28141	28178	28144	28135	28171
4bp		A	B	C	D	E	F
28240	1	**A1**	**1B**				
28231	2		**B2**	**2C**			
28138	3			**C3**			**3F**
25204	4				**D4**	**4E**	
28211	5	**5A**				**E5**	**5F**
28227	6				**6D**		**F6**
28139	7						**F7**
28122	8			**C8**			

B

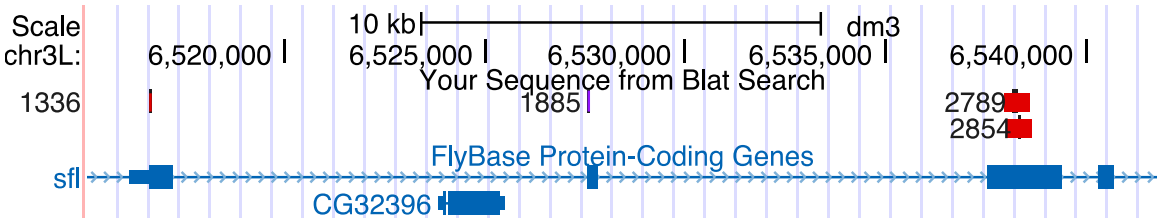

Figure S1 Pyro-sequencing cross and assay design. (A) Cross design for pyro-sequencing. Six 18bp and eight 4bp lines were randomly chosen from the 154 DGRP lines used in GWAS. The Bloomington center stock number is listed. In each cell, the order of the letter/number indicate the direction of the cross.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

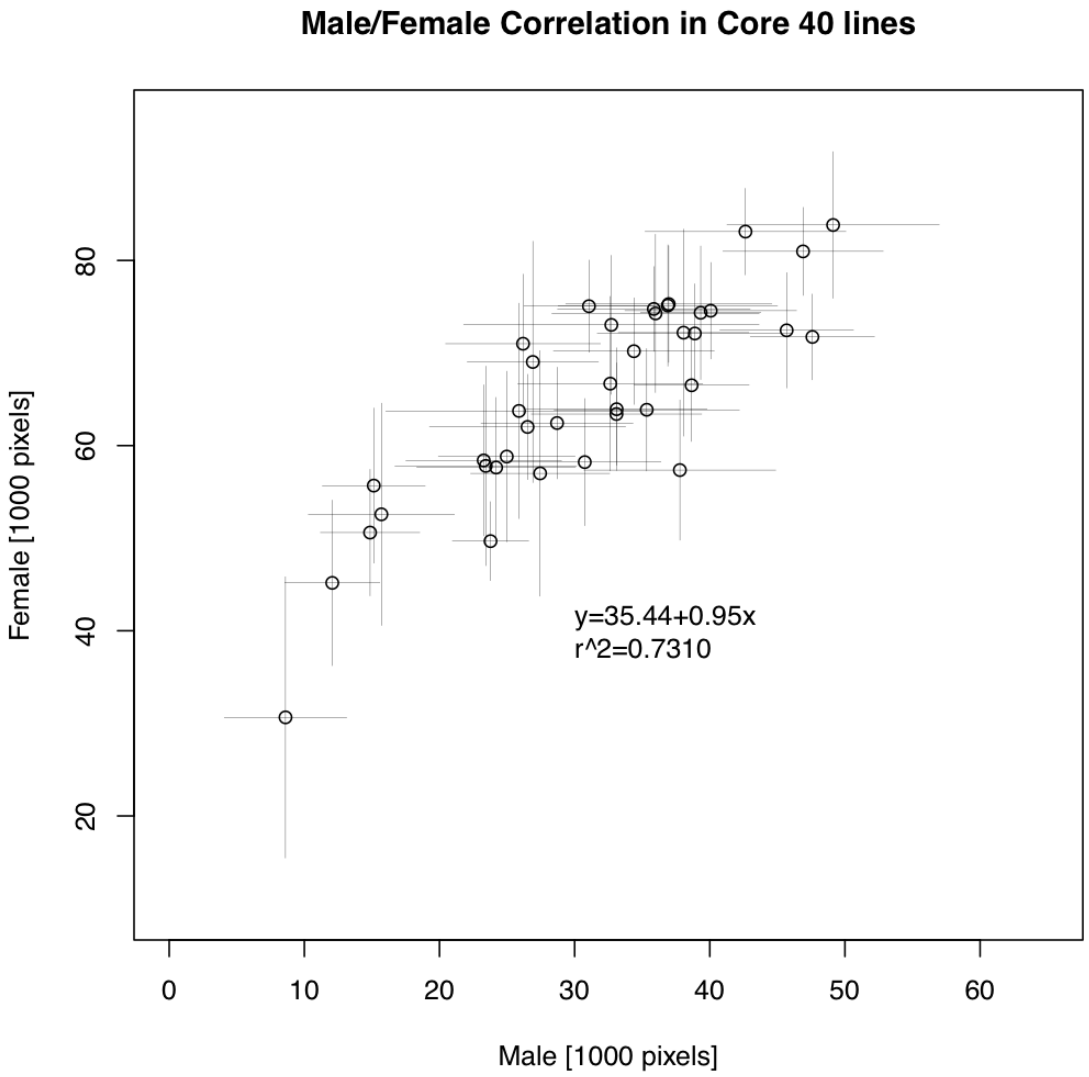

Figure S2 Correlations of eye area between F1 males and females within the same cross. Mean ± 1 s.d. are plotted for a subset of 38 lines. The least square linear fit is indicated.

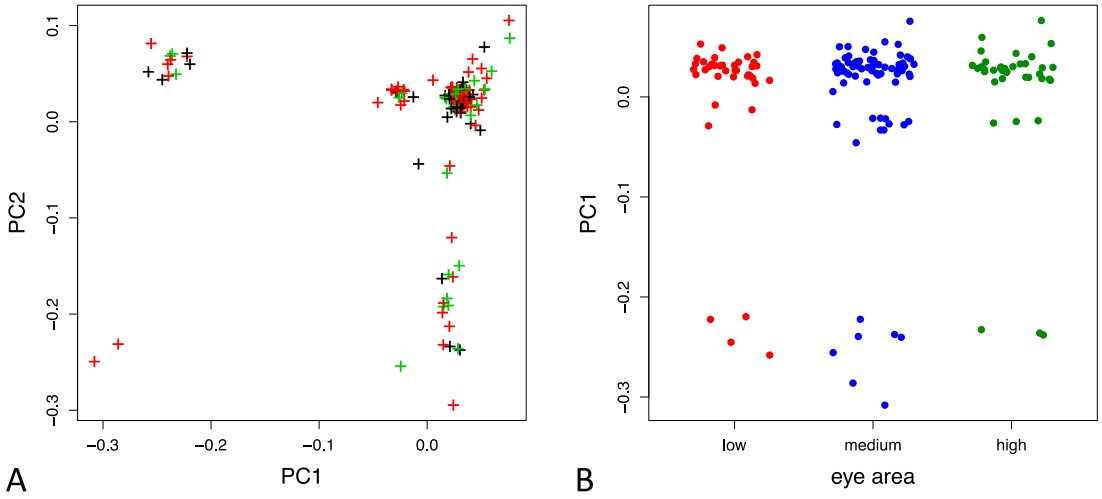

Figure S3 Population structure assessed through principal component analysis (PCA) using 900K autosomal SNPs after LD pruning. (A) 154 DGRP inbred lines projected onto the plane spanned by the first two principal components (PC1, PC2). The points are colored according to the phenotype severity in the 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 their phenotype severity showed no correlation between the two.

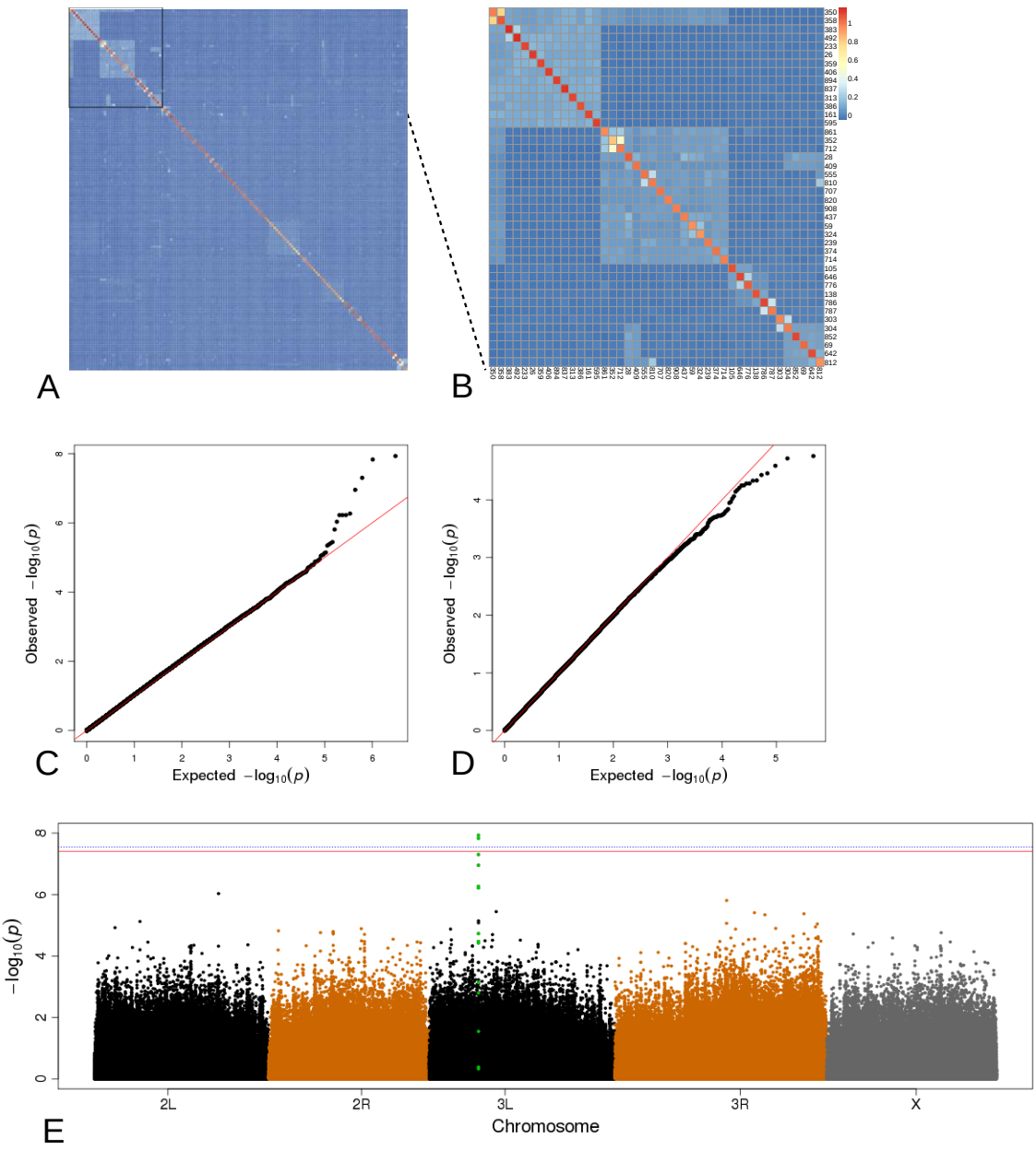

Figure S4 Mixed linear model regression accounting for cryptic relatedness. (A) The heat map shows a 154 x 154 matrix representing the centered genetic relatedness matrix (GRM) estimated using EMMAX. The boxed area is shown in detail in (B), with their line ID (RAL#) indicated on the right and bottom. 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.

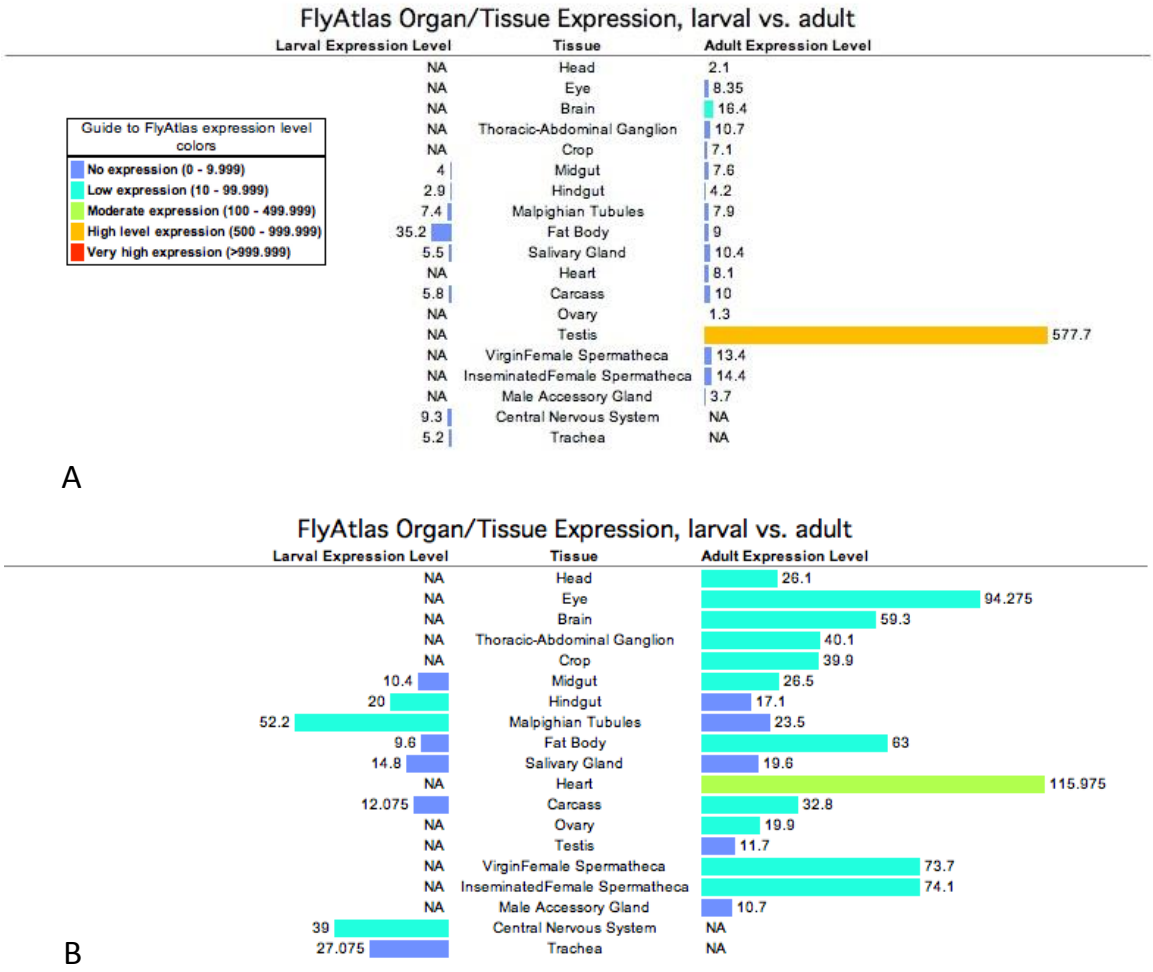

Figure S5 FlyAtlas expression report for CG32396 and sfl. (A) CG32396 (B) sfl. Figure obtained through FlyBase.

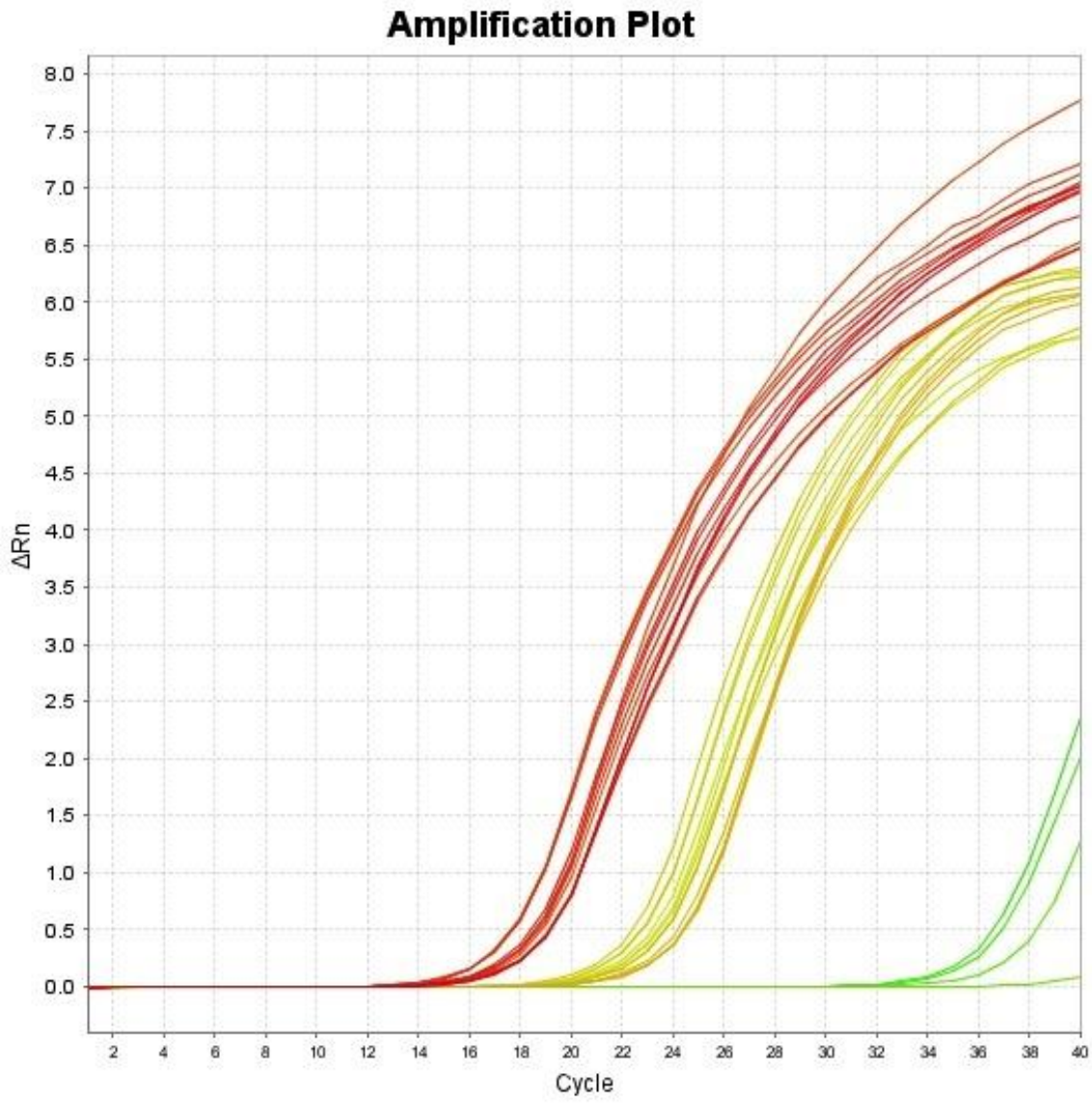

Figure S6 qRT-PCR quantification of mRNA levels for CG32396 and *sfl* in eye imaginal disc samples. Two inbred lines from DGRP were randomly chosen and eye imaginal disc samples were prepared from either 6 male or 6 female larvae, resulting in 4 biological samples. qRT-PCR were performed for each sample 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 detected starting in the 18-20th cycle, while amplification didn't happen for CG32396 until after 32 cycle. In addition, multiple melting points were detected for CG32396 assays, but not in the other two genes.

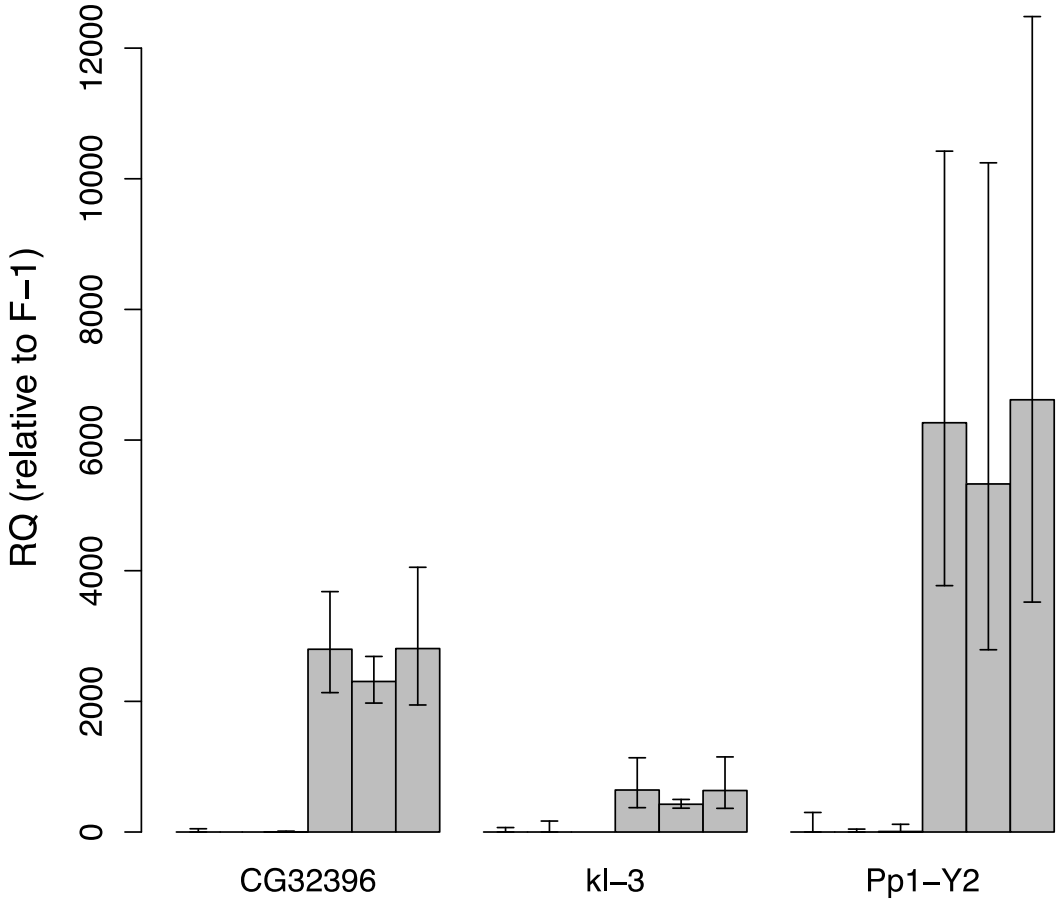

Figure S7 Relative quantity of mRNA quantified by qRT-PCR in male and female larvae. In each category, the first three bars represent three independent female larvae sample (whole larva), each assayed with three technical replicates. The height of the bar represent the mean and the full range of RQ values were 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 a testis-specific expression level. The RQ values were measured using RP49 gene as the internal control, and the first female larva sample (F-1) as the reference, whose RQ is set to one.

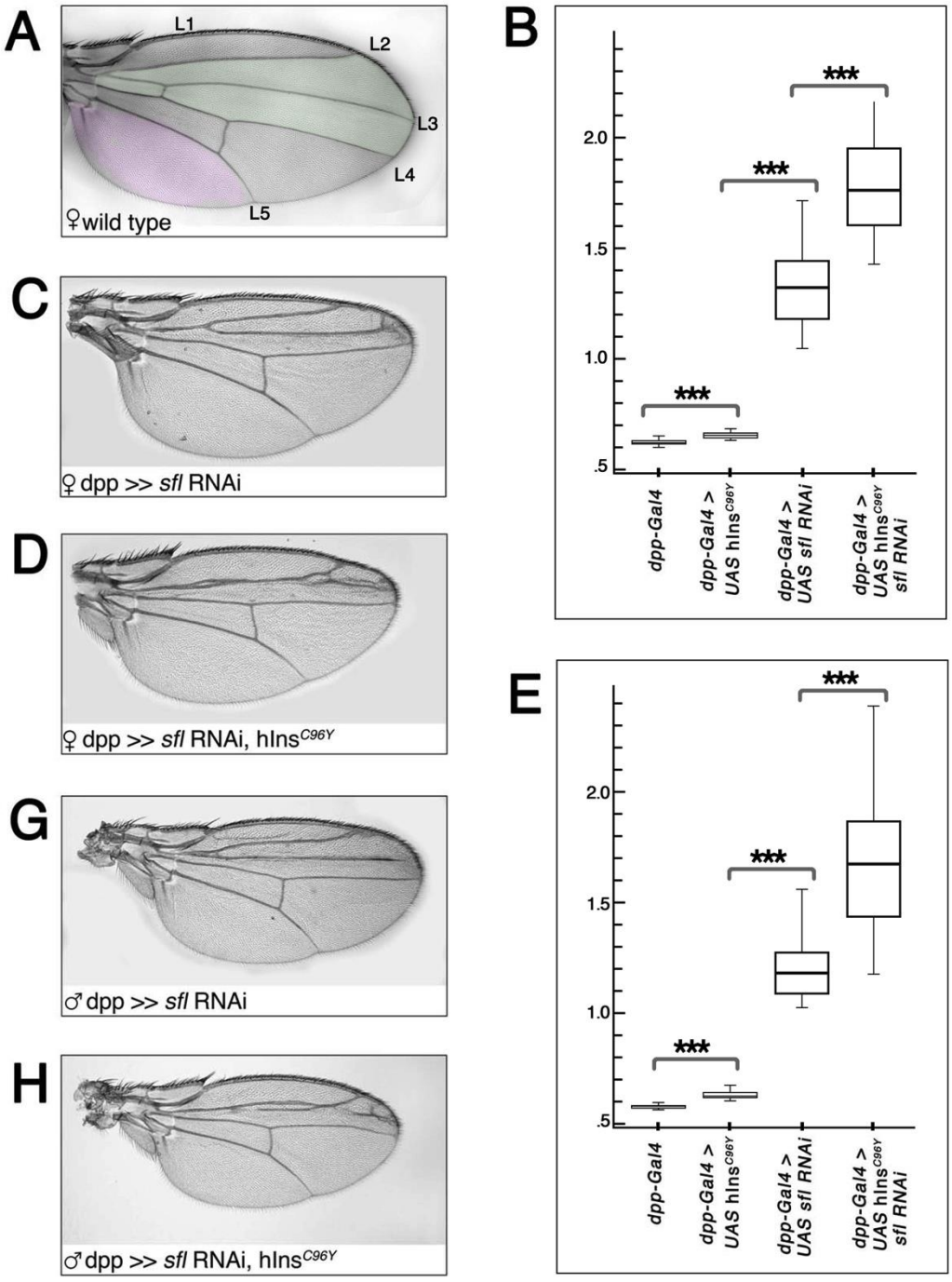

Figure S8 Depletion of *sfl* by RNAi in the developing wing expressing hINS^{C96Y} driven by dpp-Gal4. For both females and males, dpp >> hINS^{C96Y} or Dpp >> *sfl* RNAi expression alone reduces wing area between the L2 and L4 longitudinal veins relative to the posterior-most sector of the wing (bordered 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 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).

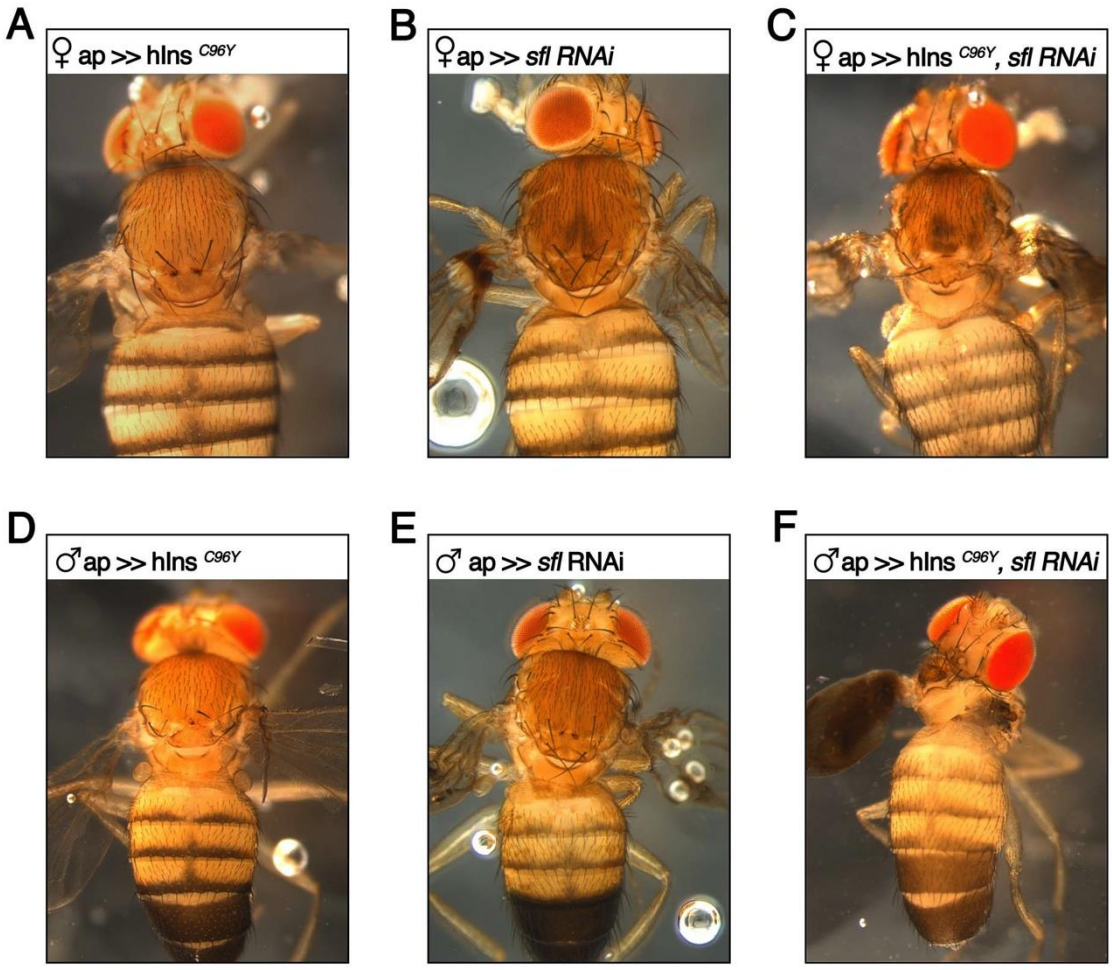

Figure S9 Depletion of *sfl* by RNAi in the developing notum expressing hINS^{C96Y} driven by ap-Gal4. For both females and males, ap > hINS^{C96Y} or ap > *sfl* RNAi expression alone reduces notum area and causes loss of dorsal macrochaetae. Co-expression of *sfl* RNAi and hINS^{C96Y} by ap-Gal4 results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

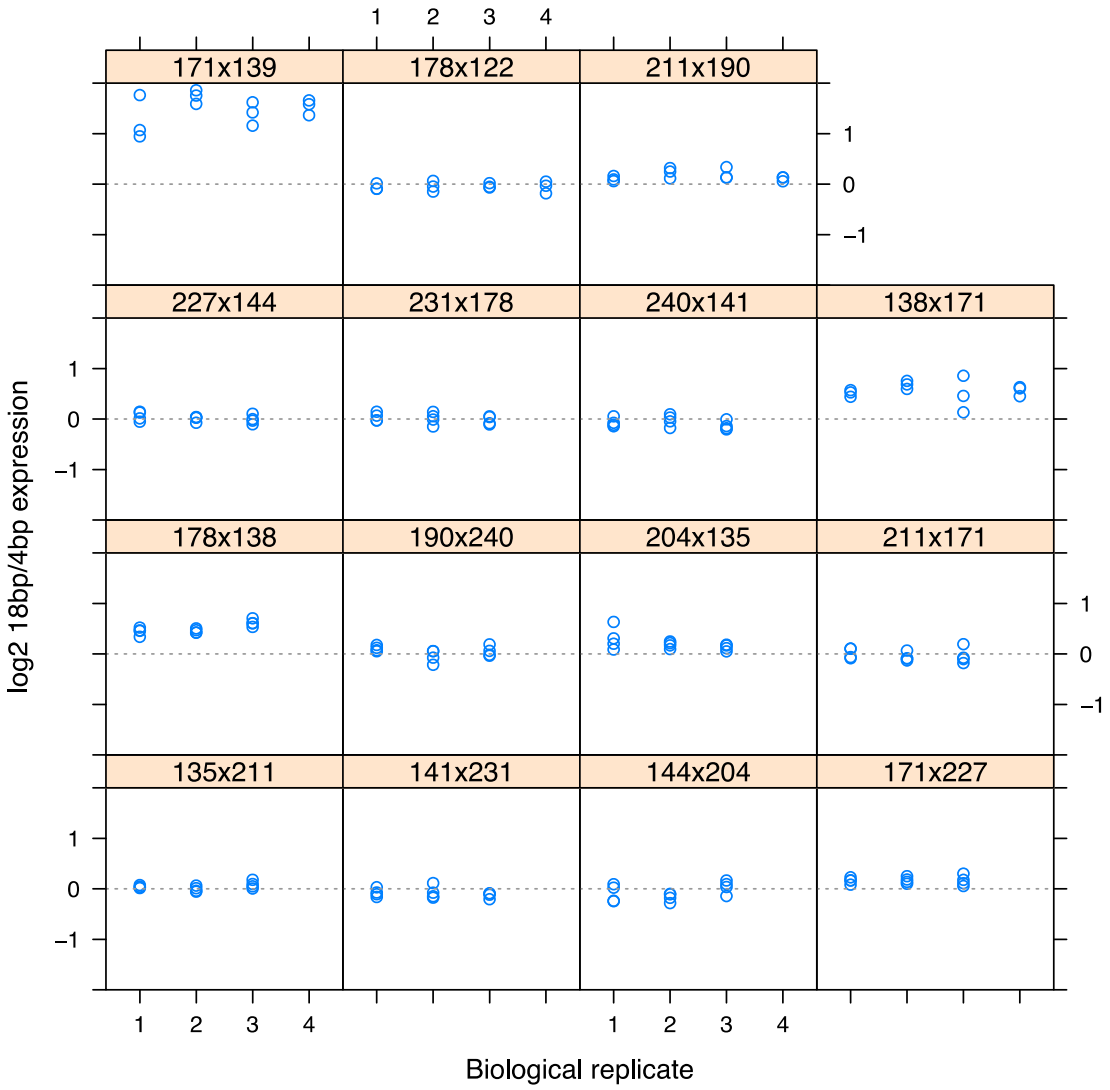

Figure S10 log2 transformed ratios between transcript levels associated with 18bp/4bp alleles. The allele-specific expression ratios were measured in F1 hybrid individuals by pyro-sequencing, with three (or four) biological replicates and four (or three) pyro-technical replicates, to obtain a total of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

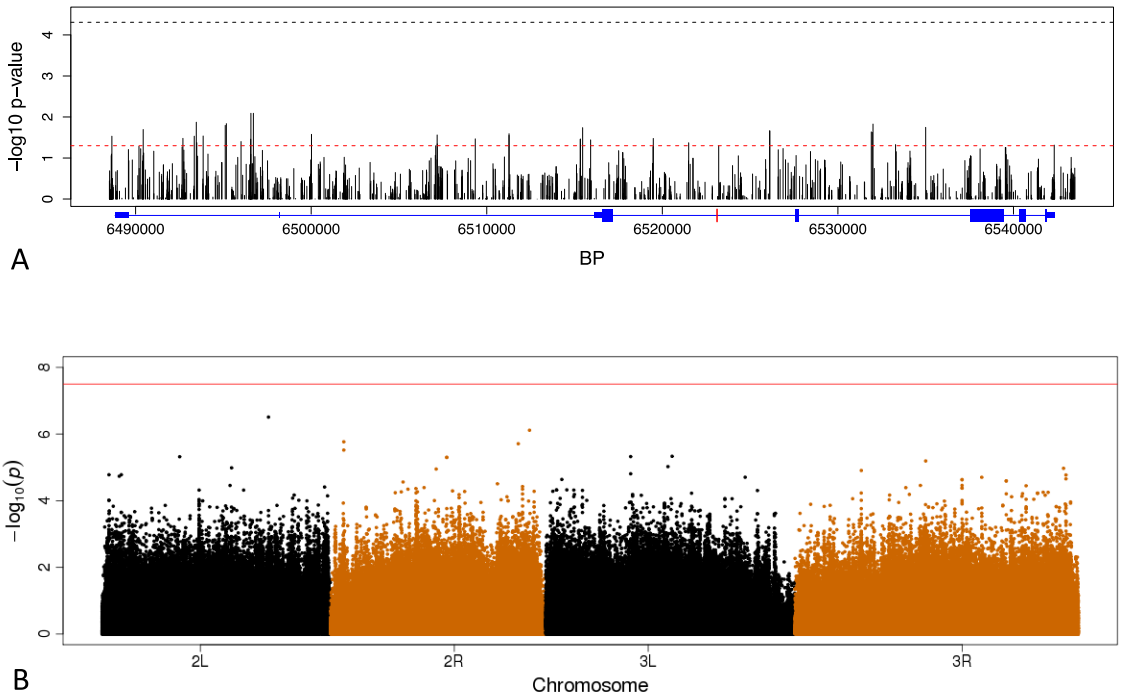

Figure S11 Conditional regression analysis to detect additional SNPs associated with the phenotype of interest. (A) within the *sfl* locus; (B) all chromosomes. The intronic 18/4bp polymorphism in *sfl* is included in the linear model as a covariate. The two dotted lines in (A) correspond to a single 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.

File S1

Mixed Model Permutation Test

When (cryptic) relatedness or population structure is present in a sample, then naïve permutation test that randomizes the phenotype values can result in inflated type-1 error (Churchill & Doerge, 2008). To address this concern we employ a permutation scheme that preserves an estimated phenotypic covariance structure as estimated using a mixed model. The idea, which is inspired by (Müller et al., 2011), is to apply a transformation to the phenotypes so that they become (approximately) independent, permute them, and then transform them back. We can show that under the mixed model assumptions, this transformation is the Cholesky decomposed inverse phenotypic covariance matrix, as estimated from using a mixed model. Hence, we transform the phenotypes as follows:

$$Y^* = cholesky(V^{-1})'Y,$$

where *Y* denotes the phenotype vector and the *V* the estimated phenotypic covariance matrix. Under the model, $Var(Y^*) = I$, which allows us to permute those values, and then apply the inverse transformation to obtain permuted phenotypes that preserve the estimated structure as follows:

$$Y_{perm} = cholesky(V)'{Y^*}_{perm}.$$

Interestingly, this approach is similar to the approach of (Aulchenko et al., 2007), where they permuted the residuals after regressing out the genomic BLUP. The difference is that we do not attempt to remove the effects of family and population structure (as inferred by a mixed models) but instead apply a transformation that preserves the (estimated) phenotypic covariance structure. Finally, in the context of mixed model association mapping it is possible to perform the permutation test very efficiently by applying this transformation to the genotypes as well. Then the least square estimate using these transformed quantities (phenotypes and genotypes) is (trivially) identical to the generalized least square estimate as obtained from EMMAX (Kang et al., 2010). For obtaining a 5% genome-wide significance threshold we performed 500 permutations and redid the genome-wide association using the EMMAX algorithm. This permutation test is implemented in the mixmogam software (Segura et al., 2012).

References

Aulchenko, Y. S., de Koning, D.-J., and Haley, C., 2007 Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis *Genetics* **177**: 577--585

Churchill, G. A. and Doerge, R. W., 2008 Naive application of permutation testing leads to inflated type I error rates. *Genetics* **178**: 609--610

Kang, H. M. M., Sul, J. H. H., Service, S. K., et al., 2010 Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42**: 348--354

Müller, B. U., Stich, B., and Piepho, H.-P. P., 2011 A general method for controlling the genome-wide type I error rate in linkage and association mapping experiments in plants. *Heredity* **106**: 825--831

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

Tables S1-S2
Available for download at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.157800/-/DC1

Table S1 Raw data

Table S2 DRRP lines used in this study

Table S3 Sequence primers used in this study

Name	Sequence (5'->3')
	qRT-PCR
sfl_F1	TCGATACGGGCGTGTTTAATGGAC
sfl_R1	TTGATAATGGGTGCGGGATGCG
CG32396_F1	AGCGGAGATTGGGTCGAAATGAG
CG32396_R1	CATGTGAAATCACGTGCCAGAAAG
kl-3F1	ATGGCAAACGTAGACCCACCTC
kl-3R1	GTACCGGCGGACGATTCTTTAG
Pp1-Y2F1	TTTGTTGTCACGGCGGTCTCAG
Pp1-Y2R1	ACGTCACATGGTCGGGCTAATTG
RP49-F1	CGGATCGATATGCTAAGCTGT
RP49-R1	GCGCTTGTTCGATCCGTA
	Pyro-seq
1336F1	CGGGCGGCAATCAACATAA
1336R1	CGGTCACGGAGCTACCAAATT
1336S1	CTCATTAAGCAGCCG
2789F1	GACTGCGACCAGATGATGTGAG
2789R1	CTTCCCTCGTGCCATGATGATA
2789S1	TTCCCGAGAATCCCA
2854F1	CGGGAAAATACTATCATCATGGC
2854R1	GTGCGAAAACCAGTTGAACTC
2854S1	TCCTGAACGTTCTGC
1885F1	TAATGGACTTATTCAACGCGACAC
1885R1	TGTGTTTGCCACCAGAGTTG
1885S1	CGGCAGTTGATAATGG

Table S4 Power calculation for GWAS with 154 lines

Minor Allele Frequency	**Effect Size***				
	0.75	**1**	**1.25**	**1.5**	**2**
0.01	0.00	0.00	0.00	0.00	0.00
0.05	0.00	0.00	0.01	0.03	0.26
0.1	0.00	0.02	0.12	0.39	0.94
0.2	0.02	0.19	0.63	0.94	1.00
0.3	0.06	0.45	0.90	1.00	1.00
0.4	0.11	0.60	0.96	1.00	1.00
0.5	0.13	0.66	0.97	1.00	1.00

* Effect size is measured as the shift in the phenotype mean in units of s.d. for the trait

The calculation is done using the t-distribution. The R-code is attached below:

```
myPower.t <- function(effect.size=1,alpha=0.05,m,n){
  ## Power for GWAS t test
  ## calculate power for a t test comparing two populations with equal variance but unequal sample sizes
  ## m, n: sample size of each allele class, not to be confused with m above
  df = m+n-2
  A = 1/sqrt(1/m+1/n) ## factor for calculating t statistics
  T = qt(1-alpha/2,m+n-2)
  T1 <- T-effect.size*A
  beta <- pt(T1,m+n-2)
  return(1-beta)
}
## plot power of GWAS t test ##
alpha1=.05/1.37e6
power <- NULL
effect.size <- c(0.75,1,1.25,1.5,2)
freq <- c(0.01,0.05,0.1,0.2,0.3,0.4,0.5)
N = 154 # size of GWAS mapping population
for(p in freq){
  m = as.integer(N*p)
  n = N-m
  power <- rbind(power, sapply(effect.size,function(x) myPower.t(x,alpha1,m,n)))
}
dimnames(power) <- list("freq"=freq,"effect.size"=effect.size)
```