The *Drosophila* gut response to infection: a systems approach

THÈSE N^{O} 7151 (2016)

PRÉSENTÉE LE 22 SEPTEMBRE 2016

À LA FACULTÉ DES SCIENCES DE LA VIE
UNITÉ DU PROF. LEMAITRE
PROGRAMME DOCTORAL EN APPROCHES MOLÉCULAIRES DU VIVANT

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Maroun BOU SLEIMAN

acceptée sur proposition du jury:

Prof. P. Gönczy, président du jury
Prof. B. Lemaitre, Prof. B. Deplancke, directeurs de thèse
Prof. T. Flatt, rapporteur
Dr R. Guigó, rapporteur
Prof. J. Fellay, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2016

Acknowledgements

First and foremost, I would like to extend my warm gratitude to Profs. Bruno Lemaitre and Bart Deplancke for accepting me in their labs. I will forever be grateful for their mentorship and advice in academic and non-academic matters. Prof. Lemaitre has given me the chance to intern in his lab and exposed me to the great scientific environment at EPFL. I admire his humanistic approach in dealing with others and the holistic views he has formed of science and society. Perhaps the most enjoyable moments I had with him were the times when we discussed over lunch-breaks about everything but science, from psychology to mundane matters. I sincerely wish him the best in research as well as in his wonderful family life. Prof. Deplancke inspires me on so many levels. It is rare to find successful people of this intellectual caliber who enjoy life and treat everyone around them as friends rather than colleagues. The fact that sometimes I feel I am older (but not wiser) resonates loudly in my mind and inspires me to become like him in the future. I hope we keep the contact, as well as collaborations, in the years to come. I would like to thank the members of my jury, Prof. Pierre Gönczy, Prof. Jacques Fellay, Prof. Thomas Flatt, and Prof. Roderic Guigo, for all the constructive comments and discussions on my work.

My greatest gratitude goes to Dani Osman (now Prof. Osman). Dani very quickly became my colleague and friend. His departure felt like losing a part of myself. Dani launched me into the world of scientific research. He laid the tracks on which my PhD progressed and inspired me to keep on pushing. He is a fantastic colleague to work with. I cannot remember a single time we did not agree on matters. His energy is truly inspiring and I hope he will keep on moving forward in his scientific endeavours.

I would like to collectively thank my colleagues in the Lemaitre and Deplancke labs. Throughout the years, the friendly environment in both labs helped me establish solid friendships. I will need a complete volume to individually express my gratitude to all these amazing people. However, I would specifically single out Gonzalo, who has become as close to me as a brother. I sincerely hope that our relationship will endure the test of time and wish him the best. I would also like to thank Alfred for the collaborations and discussions we had and wish him the best of luck in his career. My deep appreciation goes to Christophe for his assistance in the laboratory as well as for the coffee-break discussions we had. I also thank Sveta for being an excellent and dedicated colleague and Claudine for just being a decent and great person. Juan has always been the one bringing people together, and I am thankful to have known him. I hope that his future project in Kenya brings him satisfaction and scientific prosperity. I would equally like to thank Jeremy and Samuel

for all the serious discussions as well as the fun we had together. As for the members of the Deplancke lab, it has been truly an honour to work with them. Michael has been a great colleague. I admire his sense of humour and cheerfulness, and wish him the best of luck. Roel is just Roel, and I don't really like that guy, and I think I will have this animosity till the end of my 'lifespan'. As for Masha, I believe she will develop into a great motivated scientist one day and I am counting on seeing that in the future. I thank Antonio for his openness, guidance, and advice, and Petra for being a very mature, serious, yet fun person. As for Rachana, I think that we developed a great friendship without ever exchanging positive words. I admire her motivation and wish her the best in the future. I equally thank Julie for being the youngest, yet most mature member of our lab, and for being there for everyone. Finally, I would like to thank Magali, Véronique, and Tatiana for helping in administrative matters.

The years in Switzerland would not have been the same without the Lebanese mafia on- and off-campus. My sincere thanks go to Bilal for being the responsible person always helping me set my priorities straight. I really looked forward to the lunch-break discussions with Oussama (well, he is Egyptian). I hope that we will remain in touch. I thank Christopher, Sabine, and Wissam for being the most delightful group (especially Christopher for his spiritual guidance). I want to also thank Elie-Jacques, Bahaa, Momo, Pierre, Louis, Said, and Reem for being my Lebanese foster family. As for Joe, I prefer to express my love to him in person, since words do not describe the intensity of feelings I have for him. Last but not least, I would like to thank my childhood friends, a.k.a. "The Bad Boys", George, Sassine, Joe and Kamal for our long lasting and often immature relationship. The bond between us is impossible to sever, and I am grateful for having you as friends. I would like to thank Marwan, Chopin, and Wissam Baroud for their friendship and hope that we will once again be united in Mtein. As for Philip, I thank him for always believing in me.

A special thanks from the heart goes to my girlfriend, Elodie, who was always supportive and loving, especially towards the end of my PhD. Her presence at my side was reassuring. I hope our relationship flourishes in the future and that I can offer her as much love and support as she has to me.

Last but not least, I would like to thank my family. My parents have sacrificed a lot in war-torn Lebanon to make me the person I am. I cannot think of a way to repay them that except for succeeding in life. Their unconditional love is the buttress on which all my achievements were built. I hope I will make them proud. My older brother Sleiman has been the greatest positive influence in my life. He guided me through life with his unmatched love, generosity, and wisdom. Every brother should aspire to be like Sleiman. As for my lovely sister, Maria, I am lucky to have her support in every aspect of life. I love her from the bottom of my heart and wish her a long and happy family life with Rami.

Lausanne, June 24th 2016

Abstract

Genetic, physiological, and biochemical studies have successfully ascribed functions to genes in diverse processes. However, the majority of our knowledge in biology is qualitative in nature and is usually based on classical screens, where large effects on a qualitative phenotype are usually sought. While very essential to our mechanistic understanding, these methods can be inadequate when it comes to understanding inter-individual differences in complex quantitative traits. The intensive characterization of the *Drosophila* gut response to infection has led to the identification of many of its major players and canonical pathways. However, knowledge of what genes and pathways are relevant in determining inter-individual differences in a natural population is still lacking. This study addresses this question by using a systems genetics approach where the effects of natural genomic perturbations on the outcome of enteric infection are explored, often revealing unexpected determinants of infection resistance.

Keywords

Systems Genetics – Genetics – Quantitative Trait – Complex Trait - Natural Variation – Drosophila melanogaster – Gut – Enteric infection – Gene Expression – Alternative Splicing

Résumé

L'attribution de fonctions aux gènes, intervenant dans divers processus biologiques, a connu un large essor grâce aux nombreuses études en génétique, physiologie et biochimie. Cependant les connaissances actuelles en biologie sont principalement de nature qualitative et généralement basées sur des criblages basiques où seuls des effets au niveau de l'expression qualitative d'un phénotype sont recherchés. Bien qu'essentielles à la compréhension des mécanismes, ces méthodes peuvent être insuffisantes lorsqu'il s'agit de comprendre les différences inter-individuelles concernant des caractéristiques quantitatives plus complexes. La caractérisation intensive de la réponse immunitaire et physiologique à l'infection du tractus intestinal chez Drosophila a conduit à l'identification de nombreux de ses acteurs principaux et voies de signalisation majeures. Cependant, les connaissances concernant l'importance des gènes et des voies de signalisation dans la détermination des différences inter-individuelles, au sein d'une population naturelle, fait encore défaut. Cette étude tente de répondre à cette problématique en utilisant une approche de Génétique des systèmes, où les effets de perturbations génomiques naturelles sur la réponse à l'infection entérique sont explorés, révélant souvent des facteurs déterminants inattendus concernant la résistance à l'infection.

Mots-clés

Génétique des systèmes – Génétique - Caractéristique quantitative - Caractéristique complexe - Variation naturelle - Drosophila melanogaster – Tractus intestinal - Infection entérique - Expression de gène - Épissage alternatif

Contents

List of Figures

List of Supplementary Figures

List of Supplementary Tables

Chapter 1 Introduction

1.1 Heredity and Quantitative Traits

Organisms share very similar features, or a theme, yet each is a distinct variation on that theme. Even at the level of a single species, there exist infinite variations on the main theme. In his seminal book, *On the Origin of Species*, Charles Darwin proposed a theory that connected all life forms and provided a rational explanation for the presence of intra- and interspecific diversity (Darwin 1871). This paradigm shift laid the grounds for other scientists to uncover the determinants of this diversity, starting from the laws of genetic inheritance, to the discovery of the genetic code, to the current age of genomics (Mayr 1982).

It was clear that the mode of inheritance of some traits, like the those of Gregor Mendel's peas (Mendel 1965), were relatively easy to understand, since they are determined by a single genetic factor. However, most traits are quantitative in nature rather than discrete. The concept of Mendelian inheritance, which is based on the discrete inheritance of genetic loci, seemingly failed to explain continuous or quantitative traits and was initially shunned by a part of the scientific community (Franklin, Edwards et al. 2008). Biometricians, including Karl Pearson, believed that only statistics and mathematics could lead to answers, yet their approaches were only based on phenotypic observations and offered little biological explanation to the underlying mechanisms. Mendelians, on the other hand, claimed to have a better understanding of the laws of heredity, yet they were mathematically inept and incapable of performing rigorous statistical treatment of biological data. Each camp was actively looking for natural examples to support its theory. The Biometricians focused on small, continuous variations as proposed by Darwin's theory while the Mendelians believed large discontinuous variations are the major driving force in determining variation in a phenotype, and consequently in natural selection (Rushton 2000).

During this rather unproductive struggle between the two camps, the gap in our understanding of the genotype-phenotype relationship was not getting narrower. The breakthrough came when Sir Ronald Fisher effectively combined the two schools to help create the modern synthesis of evolution (Fisher 1918). Fisher proposed that what we perceive as quantitative variation could in fact be the result of the combined contribution of many factors with small effects, each having a Mendelian basis of inheritance. If a gene affects a trait, then the individuals carrying different alleles should deviate from a certain central value. Hence, the concepts of mean and variance could be directly applied to problems in genetics while respecting the laws that Mendel proposed and others confirmed. The next major milestone was the discovery of genetic linkage (Bateson, Mendel et al. 1902, Morgan 1911, Punnett 1911, Morgan 1915) and construction of the first linkage maps (Morgan 1911, Sturtevant 1913). By calculating frequencies of crossing over, the physical location of the so-called 'factors' leading to visible phenotypes could now be mapped to chromosomes, giving us the first coarse understanding of how hereditary information is organized. This allowed the analysis of quantitative traits from a new perspective that integrates the physical location of the genetic factors and their modes of inheritance. Segregating alleles that are physically close are expected to have higher linkage than those that are farther away, and alleles on different chromosomes are completely unlinked. Knowledge of the physical maps of alleles, as well as the relationship between individuals, could therefore inform geneticists about the possible effects of those alleles.

Subsequent experimental studies on quantitative traits (Castle and Little 1910, Altenburg and Muller 1920) led to the realization that multiple factors affect the levels of those traits, a concept termed the "polygene" (Thoday 1961). The polygene was defined as a set of loci underlying quantitative variation that share with Mendelian characters the same laws of inheritance. It was therefore crucial to identify the locations and contributions of those loci using Quantitative Trait Locus mapping, or QTL mapping. The first example of QTL mapping was performed by Sax in 1923 on the pigmentation and size of common beans, where he crossed beans of different sizes or colors and examined the F2 segregates and found that size, unlike pigmentation, is affected by the cumulative contributions of independently inherited linkage groups (Sax 1923).

Advances in molecular biology provided the first mechanistic understanding of biological processes and heredity, and led to the formulation of the 'central dogma of molecular biology' (Crick 1958). This was followed by rapid technological advances that nowadays give researchers unprecedented ability to acquire large-scale biological data. In what is now called systems genetics, extensive genomic as well as tran-

scriptomic and other data from many individuals is integrated to study the genetic architecture of quantitative traits as well as the basic biology underlying organismal function (Mackay, Stone et al. 2009, Civelek and Lusis 2014). In this thesis, I attempted to understand one complex trait where quantitative variation is very relevant: the inter-individual variation in resistance to enteric infection. Why and do different, immunocompetent, individuals succumb differently to a bacterial infection in the intestinal tract? Over the next chapters, I shall describe what I, along with colleagues, have uncovered.

1.2 Nature versus Nurture

A major question in genetics is the extent to which traits are determined at the genetic level, or in quantitative genetics terms, heritability. Fisher's pioneering work in analysis of variance (ANOVA) has been motivated by this question, and paved the way for the estimation of this central property (Fisher 1918, Wright 1921). According to Wright and Fisher, the observed variance in a certain trait can be partitioned into different sources of variation. In its most basic form, the total variance observed, or the phenotypic variance V_P, is the sum of the genotypic variance V_G and the environmental variance V_E. The ratio of genotypic variance to environmental variance reflects the degree of genetic determination, and is termed broad-sense heritability. Broad-sense heritability, however, does not estimate the extent to which genetic information passed from the parents affects the phenotype, since not all genetic effects are additive in nature. Therefore, V_G could be further subdivided into additive genetic variance V_A, the dominance variance V_D, and the interaction variance V_I, (Falconer and Mackay 1996). The narrow-sense heritability h^2 (referred to as heritability), is defined as the ratio of additive genetic variance to phenotypic variance. What h^2 essentially reflects is the proportion of observed variation that is attributable to genetic polymorphism in an additive manner and, more importantly, the degree of resemblance between relatives (Falconer and Mackay 1996).

Studies on many human and non-human complex traits have shown that many are highly heritable. Height, for instance, has a heritability ranging from 51-80% (Silventoinen, Magnusson et al. 2008, Zaitlen, Pasaniuc et al. 2014). Initial attempts to identify the genetic factors explaining this heritability were not successful. Genome wide association studies on human height have identified 54 genetic variants influencing the trait, yet they only explain 5% of the phenotypic variance (Visscher 2008). This observation and others have prompted scientists to call this problem "the missing heritability" (Manolio, Collins et al. 2009). Those attempts were mainly based on the premise that few loci in the genome should explain a large fraction of the observed phenotypic variation. While this can be true in some cases, it is in fact highly dependent on the

genetic architecture of the trait in question (Swami 2010, Zhang 2015). For instance, when considering all variants in the human genome simultaneously, up to 45% of the phenotypic variance can be explained (Yang, Benyamin et al. 2010). Interestingly, when the contribution to the heritability in height was broken down by chromosome, longer chromosomes were found to contribute more than shorter ones (Visscher, Macgregor et al. 2007). This supports the notion that height and possibly many other traits are highly polygenic and are affected by many loci with small effect sizes.

One should also be aware that the heritability estimate might be inflated by non-additive genetic interactions such as epistasis (Zuk, Hechter et al. 2012). Genetic interactions can exist between two alleles at the same locus, called dominance interactions, and at different loci, called epistatic interactions (Falconer and Mackay 1996). In other words, certain combinations of alleles affect the measured phenotype in a non-additive manner, which could lead to inaccurate estimates of the narrow-sense heritability (which is a top-down estimation from the phenotypes of the population), and therefore an inability to account for it using additive models with identified genetic loci. Also, as mentioned earlier, a proportion of the phenotypic variance can be attributed to variations imposed by the environment. For example, in human monozygotic twins, some traits are largely affected by either shared or unique environmental factors (Boomsma, Busjahn et al. 2002). It is important to acknowledge that some traits are more affected by environmental factors than others. Also, while a change in environmental conditions can cause the same phenotypic change in different genotypes, some genetic effects might be manifested differently in different conditions, a phenomenon called Genotype-by-Environment (GxE) interactions. These interactions are specifically important, for instance during disease and infection, and could be determining the prognosis (Baye, Abebe et al. 2011). Furthermore, some variations seem to be neutral in a range of conditions, and only manifest themselves when the environment is perturbed. These cryptic genetic variations can be especially pertinent in determining the penetrance of common diseases as well as the efficiency of plant and livestock breeding programs (Gibson and Dworkin 2004).

It is therefore imperative that any genetic or breeding study takes into account for or eliminates environmental factors (Falconer 1952). Ideally, by measuring the traits of individuals of the same genotype under different conditions, one could analyse these interactions and uncover cryptic variation. In most cases, like in human studies, controlling the environmental factors is impossible, and researchers must incorporate as many of those variables in their models as possible. Model systems could allow for a more precise control

of the environment, and thus prove to be more convenient to carefully dissect the relationship between genotype and phenotype (Lehner 2013).

1.3 The Genetic Architecture of Quantitative Traits

The number of loci affecting a trait, the distribution of their effect sizes, and their interactions with each other (dominance and epistasis) and with the environment constitute the mapping between the genotype and the phenotype. This mapping and its variational properties is referred to as the genetic architecture of a trait (Hansen 2006). Characterizing the genetic architecture is a major aim in biology for many important reasons: understanding and treating disease, animal and plant breeding, and understanding evolutionary processes such as speciation and adaptation.

Almost a century has passed since the concepts and approaches devised by Fisher and others have been proposed, yet they are still widely used today, and specifically in this thesis. What has changed is the scale at which we now perform our analyses. Advances in molecular biology, sequencing and computation have allowed us to acquire genetic information at an unprecedented pace, and now the main challenge is sifting through large datasets to identify the most relevant factors. A Genome Wide Association Study (GWAS) is one powerful strategy that attempts to understand the genotype-phenotype relationship. In its simplest form, a GWAS calculates the effect size of a certain polymorphism and the significance of its association with the phenotype (Bush and Moore 2012). It is worthy to note that an associated locus does not have to contain the causal variant, but might be linked to it. The result is a table with a number of rows equal to the number of tests, or variants, indicating their statistical significance and effect sizes. The distribution of those values is dependent on the genetic architecture of the trait. For instance, a trait with a simple architecture would have one or few very significant loci with large effect sizes. In contrast, a highly polygenic trait is expected to have very few significant loci with large effects, indicating that its levels are not imposed by a single or few major factors.

Due to its reliance on statistical testing at every segregating locus, the statistical power of each GWAS test is dependent on the allele frequencies of this locus. Consequently, it is impossible to perform associations for loci with rare variants. This is why GWAS is only well suited to uncover the effects of common variants,

which is also contributes to the inability to completely account for heritability. These effects can be estimated rather indirectly through region-based associations, where rare variants are aggregated by gene or region (Lee, Abecasis et al. 2014). Another limitation of GWAS is that it is limited to segregating variation in the selected population. Hence, conserved loci or loci that are not variable within the studied population cannot be interrogated. This undermines the reproducibility of many studies, especially when there is population stratification (Nebert, Zhang et al. 2008).

An intimate knowledge of genetic architecture at the highest resolution is of great scientific and practical interest. For instance, knowledge of heritability could be used to predict the response to selection in plant and animal breeding programs where a certain quantitative trait is sought to be optimized (Falconer and Mackay 1996, Lynch and Walsh 1998). The breeder's equation (*Response* = h^2 *x Selection differential*) predicts the change in the trait mean, or response, after one generation of selection based on the heritability and the selection differential, which is the difference in the mean value of the selected individuals and the population. The power of this equation is that there is no need to know all allelic effects and frequencies in the population. As selection leads to changes in allele frequencies in the subsequent generations, the genetic variance V_G would be changed, and hence heritability would be altered. This is why a good understanding the genetic architecture of a trait could help in streamlining breeding programs. By performing 'genomic selection' using high-resolution genotyping data, more informed breeding schemes could be devised (Goddard and Hayes 2007). For example, in dairy cattle, the rate of genetic improvement could be doubled by using information from 50,000 single nucleotide polymorphisms (Seidel 2009). In addition to that, the cost of the breeding programs in dairy cattle could be reduced by 92% (Schaeffer 2006).

1.4 Systems Genetics: moving from Genetic to Molecular Architecture of Complex Traits

One of the aims of quantitative genetics is to identify and estimate the effect sizes of polymorphic genomic loci (Quantitative Trait Loci or QTLs) that are possibly affecting a complex trait. This is achieved by testing for statistical association between alleles at a certain locus and a quantitative phenotype. However, it is neither able to single out the causal loci, nor the molecular mechanisms underlying their effects. In addition, most of the genome consists of non-protein-coding regions that include the majority of the genetic variation. This makes it very hard to generate hypotheses on the function of a putative causal locus. Therefore, any real molecular understanding of a complex biological trait requires the simultaneous and

comprehensive analysis of the organism at various levels. Systems genetics is the approach where different layers of information, or intermediate phenotypes, are integrated in populations whose individuals have a variable trait of interest.

Intermediate phenotypes such as gene expression and proteomics data could provide valuable insights into the biological networks affecting the measured phenotype (Ayroles, Carbone et al. 2009, Civelek and Lusis 2014). Intuitively speaking, genetic variation might have more direct and detectable effects on the intermediate phenotypes. For example, variations in transcription factor binding sites could lead to measurable differences in gene expression. We are only starting to appreciate the impact of variation in non-coding regions on tissue-, stage-, and environment-specific gene regulation (Cubillos, Coustham et al. 2012, Kilpinen, Waszak et al. 2013, Francesconi and Lehner 2014, Lee, Ye et al. 2014, Albert and Kruglyak 2015). In order to study this phenomenon, approaches very similar to GWAS can be used, but this time substituting the main phenotype with gene expression levels, or any other intermediate phenotype. When applied for gene expression data, *cis*- or *trans*-expression Quantitative Trait Loci (eQTLs) can be identified (Doss, Schadt et al. 2005, Nica and Dermitzakis 2013, Huang, Carbone et al. 2015). Indeed, eQTLs are highly enriched in transcription factor binding sites and gene promoters, indicating that natural variation that affects gene regulation is more likely to be in functional elements (Gaffney, Veyrieras et al. 2012, Massouras, Waszak et al. 2012).

The transcript or protein abundance, as well as that of any other intermediate phenotype, could itself be significantly associated with the quantitative trait. For example, a complementary approach would be to associate transcript data with the quantitative phenotype in order to identify Quantitative Trait Transcripts or QTTs (Passador-Gurgel, Hsieh et al. 2007). In this paradigm, the flow of information from genotype to intermediate phenotype to the main phenotype is modeled, and meaningful patterns can be deduced. In addition to the genetic architecture, the discipline of systems genetics could therefore provide unprecedented insights into the *molecular* architecture of complex traits and the flow of information in biological systems.

Variation is a ubiquitous and necessary concept in systems genetics. The phenotype, the genotype, and the intermediate phenotypes are all sources of variation. In addition to that, a system's response to stimulus or environmental perturbation constitutes another interesting source of variation. Studies on mammalian

primary immune cell lines show that hidden eQTLs are detected after stimulation with immunogens (Orozco, Bennett et al. , Fairfax, Humburg et al. 2014, Lee, Ye et al. 2014). Therefore, experiments in primary cell lines or model systems such as mice, flies, plants, and worms subjected to different environments could help understand which pathways and networks underly complex traits. Furthermore, overlapping eQTLs, as well as protein QTLs, and other intermediate-phenotype QTLs could help annotate the results of the growing body of GWAS results.

Another, often less studied source of variation is alternative splicing. Eukaryotic genes often produce a mixture of several isoforms to generate protein diversity as well as to fine tune the transcriptome. By producing multiple isoforms, different protein sequences could be generated from the same locus and/or distinct regulatory elements could be included in the transcript. This phenomenon is heritable and could be affected by genetic variation (Kwan, Benovoy et al. 2007). sQTL analysis is the systematic analysis of the effect of genetic variation on transcript isoform variation (The GTEx Consortium 2013, Monlong, Calvo et al. 2014, Zhang, Joehanes et al. 2015). In their study on whole blood from 5,257 Framingham Heart Study participants, Zhang and colleagues detect more than 500,000 cis-sQTLs corresponding to 2,650 genes. Interestingly, 395 sQTLs had a GWAS signal yet no eQTL signal. These findings further support the utility of intermediate phenotype associations such as sQTL analysis in providing mechanistic insight into GWAS results.

In some cases, the information gained in model systems could shed light on the basis of human common diseases (Flint and Mackay 2009). While the exact molecular variants might not be conserved across species, some commonalities in how the system functions could exist. Moreover, analytical tools that are tested and optimized in model systems could be applied on human data. One advantage of using model systems is that experiments can be done *in vivo* rather than on derived cell lines. In a following section, I shall introduce *Drosophila* as a powerful model system and the *Drosophila* Genetic Reference Panel, a very attractive resource that allows researchers to perform economically feasible *in vivo* quantitative and systems genetics studies.

1.5 *Drosophila melanogaster*: a brief background

Drosophila melanogaster, commonly referred to as the fruitfly, is an attractive choice for genetic studies (Jennings 2011, Hales, Korey et al. 2015). This genetically tractable species has been used for over a century by a large community of researchers who have amassed considerable information on its develop-

ment, physiology, and genome structure and sequence. Summarizing the vast knowledge on this organism is impossible, but I will outline the major motivations for its relevance in current biology and specifically quantitative genetics.

One very important reason why *Drosophila* is popular is purely technical in nature. Rearing flies is both inexpensive and easy (Greenspan 2004). Little space is required and few equipment are needed to manipulate it. It has a short generation time, around 10 days at room temperature, which means that performing crosses and studying multiple generations typically takes few weeks. Flies have high fecundity and females are able to lay up to 100 eggs a day. Another reason is the relative simplicity of the fly as a system compared to humans. At the level of the genome, it only has four chromosomes. Its complete genome was sequenced in 2000 (Adams, Celniker et al. 2000) and at the moment of writing this thesis, it contains 13,907 protein coding genes in contrast to humans, who have 20000-25000 genes. In addition, there is a high degree of conservation between basic developmental Drosophila and humangenes. 75% of human disease genes have related sequences in *D. melanogaster* (Reiter, Potocki et al. 2001). As a consequence, the fly is used as a genetically tractable disease model for many human disorders. The ease of inducing and tracking mutations in the fly have made large genetic screens possible, leading to the discovery of the function of a plethora of genes and pathways (St Johnston 2002). All these factors have contributed, over the last century, to the creation of genetic tools to manipulate the fly genome (Hales, Korey et al. 2015). The yeast GAL4/UAS system along with the GAL80 protein are of particular interest in this thesis (Ma and Ptashne 1987, Brand and Perrimon 1993). This system allows spatiotemporally-controlled transgene expression, allowing high-resolution manipulations of gene expression (Rodríguez, Didiano et al. 2011). Last but not least, the large *Drosophila* scientific community has generated several databases and online resources. The most widely used resource is Flybase, a very rich database that aims to integrate all knowledge accumulated in fly research (McQuilton, St Pierre et al. 2012, St Pierre, Ponting et al. 2014).

In addition to classical genetics, fruit flies have been extensively used in the fields of population and quantitative genetics (Flori and Mousseau 1987, Coyne and Orr 1989, Falconer and Mackay 1996, Lynch and Walsh 1998). Phenotypes, whether discrete or quantitative, can easily be measured for large samples, often collected from the wild (Klepsatel, Gáliková et al. 2013). Additionally, there are some classical phenotypes that have been studied for decades. For example, bristle number has been used extensively as a system to understand the genetic basis of quantitative variation as well as response to selection (Mackay and Lyman 2005). Finally, the availability of genome sequences of other related *Drosophila* species have al-

lowed in-depth genome-wide characterization of evolutionary forces shaping genomes (Drosophila 12 Genomes Consortium 2007).

1.6 The *Drosophila* Genetic Reference Panel

The *Drosophila* Genetic Reference Panel (DGRP) was conceived in order to reliably study quantitative traits in a model organism. It is a set of *Drosophila melanogaster* lines derived from an out-crossed population in Raleigh, North Carolina (Mackay, Richards et al. 2012, Huang, Massouras et al. 2014). Inseminated females were collected from the Raleigh State Farmer's Market, then their progeny were subjected to full-sib mating in order to approach full heterozygosity. Subsequently, the genomes of 205 lines were sequenced and made available for the scientific community. By comparing the lines' genome sequence to the *Drosophila* reference genome sequence, high quality genotype data is available which consists of 4,853,802 single nucleotide polymorphisms (SNPs) and 1,296,080 non-SNP variants including insertions and deletions (indels) and structural variants (Huang, Massouras et al. 2014). The panel therefore constitutes a living library of natural genetic variation that can be used to understand the genetic basis of multiple quantitative traits. Importantly, repeated phenotypic measurements can be performed on individuals with the same genetic makeup, thus granting researchers the ability to estimate the within-strain variability of a trait. By coupling the phenotypic information with the genotyping data through Genome Wide Association Studies (GWAS), researchers can identify quantitative trait loci (QTLs) to gain an understanding of the genetic basis of traits and the genes involved. The rapid decay in linkage disequilibrium and the lack of population structure in the DGRP make them suitable for GWAS (Mackay, Richards et al. 2012).

Several studies have already been published using the DGRP lines and a major theme emerged when common natural variants were investigated: most quantitative traits generally have complex genetic architectures with many genetic loci of small effect. Chill coma recovery, startle response, and starvation stress were among the first to be studied (Mackay, Richards et al. 2012). Those traits exhibited high broad-sense heritabilities, indicating that they have a large genetic component. However, and rather counterintuitively, very few QTLs passed genome-wide significance. In most of the cases, these QTLs were in non-coding regions or near genes that were not known to be involved in the quantitative trait of interest. The implications are two-fold. First, the results underline the possible importance of variation in non-coding regions and their possible impact on gene regulation. Second, the GWAS could identify novel players in a trait of interest.

On the other end of the spectrum, some traits in the DGRP appear to have a simpler genetic architecture. Resistance to viral infection is one example, with one common polymorphism explaining up to 47% of the heritability in susceptibility (Magwire, Fabian et al. 2012). Interestingly, many genetic loci identified in the DGRP lines were later identified in a multi-parent advanced intercross panel, the Drosophila Synthetic Population Resource (DSPR; (Long, Macdonald et al. 2014)), indicating that results from the DGRP are not specific to this panel (Cogni, Cao et al. 2016). The reason that some loci were not replicated is that the 8 founder lines of the DSPR were not polymorphic with respect to them. It is not clear whether this simple genetic architecture is specific to viral infection or whether it is a hallmark of host-pathogen interactions, and therefore one of the aims of this thesis is to describe the genetic basis of resistance to enteric bacterial infection.

1.7 The Immune System of *Drosophila*

The immune system is the compendium of mechanisms and structures that protect a host from the pathogenesis caused by other organisms. The first line of defence in many organisms is the physical barrier, whether it is the skin in humans, exoskeleton of insects, or the mucous membranes covering epithelial surfaces (Janeway, Travers et al. 1997). If and when these barriers are breached, a successful immune strategy hinges on the recognition of pathogens, the deployment of a controlled response to neutralize the threat, and the eventual restoration of homeostasis. Animal immune defence mechanisms against pathogens can be broadly classified into innate and adaptive responses. Innate immunity is the more ancient arm of the immune system (Bayne 2003, Litman, Rast et al. 2010). In the innate immune system, specific receptors, or pattern recognition receptors (PRRs) recognize specific pathogen-associated molecular patterns, or PAMPs (Akira, Uematsu et al. 2006, Takeuchi and Akira 2010). Following that, signalling cascades are triggered, leading to the induction of a diverse array of antimicrobial peptides (AMPs) that typically target microbial membranes (Zasloff 2002). The signal transduction pathways involved typically converge on transcription factors of the NF-κB family that share a common evolutionary origin (Kopp and Ghosh 1995, Huguet, Crepieux et al. 1997). Another aspect of the innate immune response is the production of bursts of reactive oxygen species (ROS), nitric oxide (NO), and NO derivatives that are microbicidal as well as components of diverse signal transduction pathways (Bogdan, Röllinghoff et al. 2000). The adaptive arm of the immune response is a more recent evolutionary innovation that produces tailored PRRs specific to an infecting pathogen and keeps a memory for subsequent infections (Janeway, Travers et al. 1997). While in innate immunity, all the PRRs are encoded by the germline-encoded genes, organisms with

adaptive immunity utilize somatic mutation and recombination or receptor gene segments to produce novel receptors (Schatz 2004). Not all animals have both arms of the immune system. Invertebrates, like *Drosophila*, only possess an adaptive immune system, while vertebrates rely on a combination of both. It is important to note that the two systems are not mutually exclusive. There is considerable bidirectional crosstalk between the two arms and immune cells could possess both functions (Getz 2005).

The *Drosophila* immune response has both cellular and cell-free (humoral) immune responses to pathogens, both of which have been extensively dissected and characterized (Lemaitre and Hoffmann 2007). The cellular immune response is mainly characterized by phagocytosis of small microbes and cellular encapsulation and melanization of larger parasites. The cell-free response is the expression of a battery of antimicrobial peptides (AMPs) and other effectors after recognition of specific microbial molecules. The production of AMPs is mainly dependent on the Imd and Toll pathways, both of which rely on NF-κB transcription factors. The Imd pathway is activated in response to infection with bacteria having *meso*-diaminopimelic acid (DAP) type peptidoglycan (mostly gram negative bacteria), whereas the Toll pathway is activated upon infection with bacteria having Lysine-type peptidoglycan (mostly gram positive bacteria), fungi, and yeast. The immune response could either be systemic or local, depending on the site of infection. The systemic immune response is mediated by the fat body, the fly's equivalent of a vertebrate liver, while the local response is mediated mainly by epithelial tissues that are in contact with the environment such as the gut. A potent local epithelial response is very important since flies feed on decaying material rich in potentially harmful microbes. This thesis is mainly concerned with the gut local response, which I will briefly introduce in the next section.

1.7.1 The adult *Drosophila* Gut in the Normal and Infected State

The gut is an early innovation that followed multicellularity (Stainier 2005). The presence of this body cavity allowed organisms to switch from intracellular to extracellular modes of digestion as well as for better control of the digestive process (Lemaitre and Miguel-Aliaga 2013). Guts of different animals have diversified to allow for different feeding habits and diets (Chapman, Simpson et al. 2013). Only recently have scientists started to explore the function and complexity of the long-neglected *Drosophila* gut. One important factor in sparking this interest is the discovery that the adult gut is maintained through the action of pluripotent stem cells (Micchelli and Perrimon 2006, Ohlstein and Spradling 2006). The last decade has seen a surge in studies relating to the *Drosophila* gut, making it impossible to comprehensively address

here. The major axes of research into this organ are the morphological and developmental aspects, the digestive function, and the interaction of the gut with pathogens and commensals.

In *Drosophila*, the gut is a simple epithelium that is surrounded by visceral muscles, trachea, and enteric nerves. Far from being a simple tube, the alimentary canal is a highly compartmentalized. At the highest anatomical level, it is organized into a foregut, midgut, and hindgut. The foregut and hindgut are composed of cells of ectodermal origin that are covered on the apical side by an impermeable cuticle. The midgut is of endodermal origin and is protected from the luminal environment by a peritrophic matrix, a physical barrier consisting of chitin and glycoproteins. The midgut is further regionalized into at least five regions with distinct cellular, chemical, and physiological characteristics (Buchon, Osman et al. 2013, Marianes and Spradling 2013).

The enterocytes (ECs) and the enteroendocrine cells (EECs) are the main differentiated cell types of the adult midgut. ECs are large polyploid cells that have secretory and absorptive functions and constitute the majority of the midgut cells. EECs are less common and are thought to control the intestinal physiology through the secretion of short peptides (Veenstra, Agricola et al. 2008). The adult gut cell population is constantly replenished by a small pool of interspersed intestinal stem cells (ISCs). These stem cells can undergo symmetric division, producing two identical ISCs, or asymmetric division to produce one ISC and a enteroblast. The enteroblast is a transient undifferentiated precursor of ECs and EECs. The choice between the two cellular identities is determined by Notch signaling activity (Ohlstein and Spradling 2007). Under normal physiological conditions, the midgut epithelium is renewed within one to two weeks (Micchelli and Perrimon 2006, Ohlstein and Spradling 2006). However, the mitotic activity of stem cells is affected by multiple cues such as nutritional status (O'Brien, Soliman et al. 2011), chemical agent-induced damage (Amcheslavsky, Jiang et al. 2009), and enteric infection (Buchon, Broderick et al. 2009, Chakrabarti, Liehl et al. 2012). A number of pathways are involved in the proliferation and differentiation of intestinal cells. Following biotic or abiotic stress, damaged cells release ligands to activate signaling pathways like the JAK-STAT pathway, epidermal growth factor (EGF) receptor, decapentaplegic (DPP) and Wingless (Buchon, Broderick et al. 2013, Lemaitre and Miguel-Aliaga 2013).

Given the constant exposure to potentially harmful pathogens, gut-bearing organisms developed an ensemble of molecular and cellular processes that together constitute “gut immunocompetence”

(Woolhouse, Webster et al. 2002, Obbard, Welch et al. 2009, Barreiro and Quintana-Murci 2010). Phylogenetically distant species share similarities in innate immune pathways (Kimbrell and Beutler 2001) and major structural and physiological gut features (Stainier 2005, Karasov, Martínez del Rio et al. 2011). The study of gut immunocompetence in one system can therefore shed light on general aspects throughout the phylogeny. In *Drosophila*, great strides have been made in elucidating the biological processes underlying gut immune defense. Notably, studies in the fly gut revealed that enteric infection leads to an intricate interplay between immunological, stress, and repair mechanisms (Lemaitre and Hoffmann 2007, Ryu, Kim et al. 2008, Buchon, Broderick et al. 2009, Apidianakis and Rahme 2011, Charroux and Royet 2012, Lemaitre and Miguel-Aliaga 2013). After oral ingestion, Gram-negative bacteria are mainly recognized by two members of the peptidoglycan recognition protein (PGRP) family: the surface receptor PGRP-LC and the intracellular PGRP-LE. As in humoral immunity, these receptors activate the Imd pathway that leads to the induction the of AMP genes. In order to avoid over-activation or activation by harmless endogenous bacteria, Imd pathway activity is fine-tuned at multiple levels by negative regulators. For instance, amidase PGRPs are secreted to scavenge peptidoglycan, PIRK/PIMS disrupts the interaction between PGRP-LC and its downstream adaptor, the transcription factor Caudal restricts the expression of Imd targets (Lhocine, Ribeiro et al. 2008, Ryu, Kim et al. 2008, Paredes, Welchman et al. 2011), and ubiquitination leads to proteasomal degradation of the members of the cascade (Khush, Cornwell et al. 2002, Thevenon, Engel et al. 2009, Yagi, Lim et al. 2013).

Another pillar of the gut immune response is the generation of Reactive Oxygen Species (ROS) by the enterocytes, which neutralizes the infectious microbes but also leads to cellular damage (Tzou, Ohresser et al. 2000, Ha, Oh et al. 2005). ROS is produced by the NADPH oxidase Duox, a surface receptor that activated by the Gαq-phospholipase C-ß-Ca^{2+} pathway, which is activated upon binding of bacterial-derived uracil to a yet unidentified G-protein coupled receptor (Ha, Lee et al. 2009, Lee, Kim et al. 2013). Unlike commensals, opportunistic bacteria produce uracil, an aspect that is exploited by the fly. *Duox* expression levels have been shown to be controlled by the p38 mitogen-activated protein (MAP) kinase through phosphorylation of Activating Transcription Factor ATF2 (Chakrabarti, Poidevin et al. 2014). ROS induction leads to host cell damage and inhibition of protein translation (Chakrabarti, Liehl et al. 2012). It is therefore important that ROS levels are controlled. This is achieved either through the secretion of IRC, an extracellular catalase, or through control by the Gαq-phospholipase C-ß-Ca^{2+} pathway (Ha, Oh et al. 2005, Ha, Lee et al. 2009). In fact, while Imd pathway mutants are very sensitive to infection with Gram-negative bacterial infection, they are more tolerant to enteric infection compared to flies manipulated genetically or chemically to have high ROS levels (Ha, Oh et al. 2005, Chakrabarti, Liehl et al. 2012).

1.8 Objectives and Overview of the Thesis

In this thesis, I use a panel of *Drosophila melanogaster* inbred lines (Huang, Massouras et al. 2014) as living snapshots of variations on the same theme (the "wild-type" fly) to understand the genetic basis of resistance to enteric infection. I employ a set of tools from classical to quantitative to systems genetics to dissect the gut response, specifically variations in that response. I first assess the extent of phenotypic variation and then try to explain it by examining at genetic, physiological, and molecular factors. The fly is subjected mainly to two environments through feeding on either sucrose or the highly entomopathogenic Gram-negative bacterium, *Pseudomonas entomophila*.

In **Chapter 2**, a systematic characterization of the phenotypic differences, namely survival to infection, between the different lines is performed then we perform a QTL analysis to find genetic variants associated with the trait. We characterize some lines from the phenotypic extremes at the transcriptional, physiological, and molecular levels to gain insights into some main determinants of variability in resistance.

In **Chapter 3**, a more systematic study of gene expression is performed to shed light at the interplay between resistance to infection, gene expression differences, and genetic variation. We attempt to predict the phenotype from gene expression signatures and catalogue possible regulatory variants affecting gene expression, and ultimately the organism's resistance to infection.

In **Chapter 4**, a special focus on splicing differences that occur in the *Drosophila* gut after infection is presented. Then we explore genetic and molecular factors mediating variation in these differences.

Chapter 5 summarizes the thesis and provides future directions and outlooks.

Chapter 2 Genetic, Molecular and Physiological Basis of Variation in *Drosophila* Gut Immunocompetence

This chapter is based on the published article "Bou Sleiman MS, Osman D, et al. (2015). Genetic, Molecular and Physiological Basis of Variation in Drosophila Gut Immunocompetence. Nature Communications 6(7829) doi:10.1038/ncomms8829" (Bou Sleiman, Osman et al. 2015) and was co-written by Dani Osman. The study explores how natural variation could lead to very different outcomes after enteric infection, with some individuals being inherently more susceptible or resistant than others. Using a wide gamut of approaches ranging from direct experimentation to bioinformatics, we attempt to identify genes and phenomena that contribute to the phenotypic differences.

Abstract

Gut immunocompetence involves immune, stress, and regenerative processes. To investigate the determinants underlying inter-individual variation in gut immunocompetence, we performed enteric infection of 140 Drosophila lines with the entomopathogenic bacterium *Pseudomonas entomophila* and observed extensive variation in survival. Using genome-wide association analysis, we identified several novel immune modulators. Transcriptional profiling further showed that the intestinal molecular states of resistant and susceptible lines differ, already pre-infection, with one transcriptional module involving genes linked to reactive oxygen species (ROS) metabolism contributing to this difference. We found that this genetic and molecular variation is physiologically manifested in lower ROS activity, lower susceptibility to ROS-inducing agent, faster pathogen clearance and higher stem cell activity in resistant versus susceptible lines. Together, this study provides novel insights into the determinants underlying population-level variability in gut immunocompetence, revealing how relatively minor, but systematic genetic and transcriptional variation can mediate overt physiological differences that determine enteric infection susceptibility.

Author Contributions and Acknowledgements

Maroun Bou Sleiman, Dani Osman and Bart Deplancke conceived the project, designed the experiments, interpreted the data, and wrote the manuscript. Bruno Lemaitre provided support for the experi-

ments and comments onto the manuscript. Maroun Bou Sleiman and Dani Osman carried out the experimental work. Maroun Bou Sleiman, Andreas Massouras, and Ary Hoffmann performed computational analyses. This work was supported by SNF (fellowship for Maroun Bou Sleiman and Bruno Lemaitre), FEBS (Long-Term fellowship for Dani Osman), funds from the EPFL, AgingX (SystemsX.ch), and the SNSF (CRSI33_127485) for Bart Deplancke. We acknowledge Julien Ayroles, Antonio Meireles Filho, and François Leulier for critical reading of the manuscript. We thank Julien Dow, Bloomington and DGRC stock centers for the fly stocks, and Guennaelle Dieppois for providing the PA14 bacterial strain.

2.1 Introduction

Gut immunocompetence is the repertoire of molecular and cellular processes that an organism employs in order to fight off harmful pathogens (Woolhouse, Webster et al. 2002, Obbard, Welch et al. 2009, Barreiro and Quintana-Murci 2010). How host genetic variation impacts these processes and how this is specifically encoded at the molecular and cellular levels is however still poorly understood, even though there are multiple examples where genetic variation affects an organism's susceptibility to infectious agents, including intestinal pathogens (Barreiro and Quintana-Murci 2010). This may have far-reaching implications beyond acute disease. Indeed, the inability to effectively clear pathogens, to restrain the mounted immune response, or to repair the damaged intestinal region may lead to chronic gut pathologies (Mann and Saeed 2012). Elucidating the genetic and molecular determinants that mediate variation in gut immunocompetence is therefore of critical importance.

To address this, we used *Drosophila* not only because it is quickly gaining importance as a useful model to study the etiology of inflammatory bowel diseases (Amcheslavsky, Jiang et al. 2009, Bonnay, Cohen-Berros et al. 2013), but also since it allows the analysis of molecular and organismal traits in a physiologically relevant and highly accessible system. The use of inbred fly lines allows assessment of the impact of infection on distinct, but constant genetic backgrounds to tease out the effect of the genotype from environmental effects (Lazzaro, Sceurman et al. 2004, Tinsley, Blanford et al. 2006, Mackay, Stone et al. 2009, King, Merkes et al. 2012, Magwire, Fabian et al. 2012, Massouras, Waszak et al. 2012, Huang, Massouras et al. 2014). This ability has been previously exploited to examine naturally occurring variation in pathogen susceptibility at a systemic level (Lazzaro, Sceurman et al. 2004, Tinsley, Blanford et al. 2006, Magwire, Fabian et al. 2012), albeit to our knowledge not yet in the gut. Specifically, we used the Drosophila Genetic Reference Panel (DGRP) (Mackay, Richards et al. 2012, Huang, Massouras et al. 2014) to explore variability in gut immunocompetence-related parameters and aimed to decipher the molecular and physiological determinants

driving them. We found striking variation in survival to enteric bacterial infection and identified key underlying genetic variants, transcriptional modules, and physiological signals.

2.2 Results

2.2.1 Genetic variation in susceptibility to enteric infection

To assess the extent of gut immunocompetence variation in genetically-distinct individuals, we measured fly survival following enteric infection with the entomopathogenic bacterium *Pseudomonas entomophila* (*P.e.*) (Vodovar, Vinals et al. 2005) in 140 DGRP lines whose genomes have been comprehensively characterized for single nucleotide polymorphisms (SNPs) and non-SNP variants (Massouras, Waszak et al. 2012, Zichner, Garfield et al. 2013, Huang, Massouras et al. 2014). We found striking and reproducible variation in the DGRP lines' survival (**Fig. 2:1a, Supplementary Fig. 2:1a, Supplementary Table 2:1**), comparable to previous observations regarding natural variation in systemic immunity in *Drosophila* (Lazzaro, Sceurman et al. 2004). While around 50% of the tested lines harbour the natural endosymbiont *Wolbachia* (Massouras, Waszak et al. 2012), this had no effect on susceptibility (**Supplementary Fig. 2:1b**). To eliminate the possibility that the differential susceptibility of the lines is due to differences in commensal bacteria (Buchon, Broderick et al. 2013), we infected five lines randomly chosen from each phenotypic class (resistant or susceptible), in germ-free conditions. The loss of commensals did not alter their relative susceptibility, indicating that the endogenous microbiota do not majorly impact on susceptibility class (**Supplementary Fig. 2:1c**). We also evaluated whether our results could be biased by differences in feeding behaviour between DGRP lines but found no consistent difference in food uptake between resistant and susceptible lines (**Supplementary Fig. 2:1d**). To determine if this variability in survival is specific to enteric infection, we assessed susceptibility of DGRP lines to systemic infection with *Erwinia carotovora carotovora* 15 (*Ecc15*) (**Fig. 2:1b**). We did not use *P.e.* since it leads to very fast lethality in this condition, which renders the scoring of a meaningful phenotype difficult. We found little correlation between the two infection conditions and pathogens (r=0.23, n=78, p=0.0395). This observation suggests that the determinants of gut immunocompetence are distinct from those that govern systemic immunity (Martins, Faria et al. 2013). However, one line, *#25745*, was highly susceptible in both infection conditions (**Fig. 2:1b)**. We found that this fly line contains a null mutation in the *dredd* gene, a component of the immune deficiency (Imd) pathway required to resist Gram-negative bacterial infection (Leulier, Rodriguez et al. 2000, Lemaitre and Hoffmann 2007) (**Supplementary Fig. 2:2a-d**). Mutations with such a strong loss-of-function phenotype tend to be rare in a natural population and do not capture most of the underlying natural variation in gut immunocompetence (Mackay, Stone et al. 2009). For instance, the mutation we identified in *dredd* was found in only one of 205 genotyped DGRP lines (Huang, Massouras et al. 2014). Moreover, in a natural population, such a rare reces-

sive allele would be mostly found in heterozygous form, which could explain why it has not been eliminated by purifying selection. We next examined whether the observed differences in survival is specific to *P.e.* by orally infecting DGRP lines with a clinical isolate of *Pseudomonas aeruginosa* (*PA14*). Specifically, using a similar infection protocol as for *P.e.* (**Methods**), we infected four randomly selected lines from the lower 10% in terms of survival to *P.e.* infection (*i.e.* resistant) and four randomly from the upper 90% (*i.e.* susceptible, excluding the *dredd* mutant line discussed above) and compared survival after three days (**Supplementary Fig. 2:3**). DGRP lines that were resistant to oral infection by *P.e.* were also resistant to *PA14*, while three of the four tested lines that were susceptible to *P.e.* were also susceptible to *PA14*. These results suggest that the DGRP phenotypes observed for *P.e.* infection may reflect a more general pattern in that they may be due to a common, likely bacterium-independent genetic and molecular mechanism that mediates oral infection susceptibility.

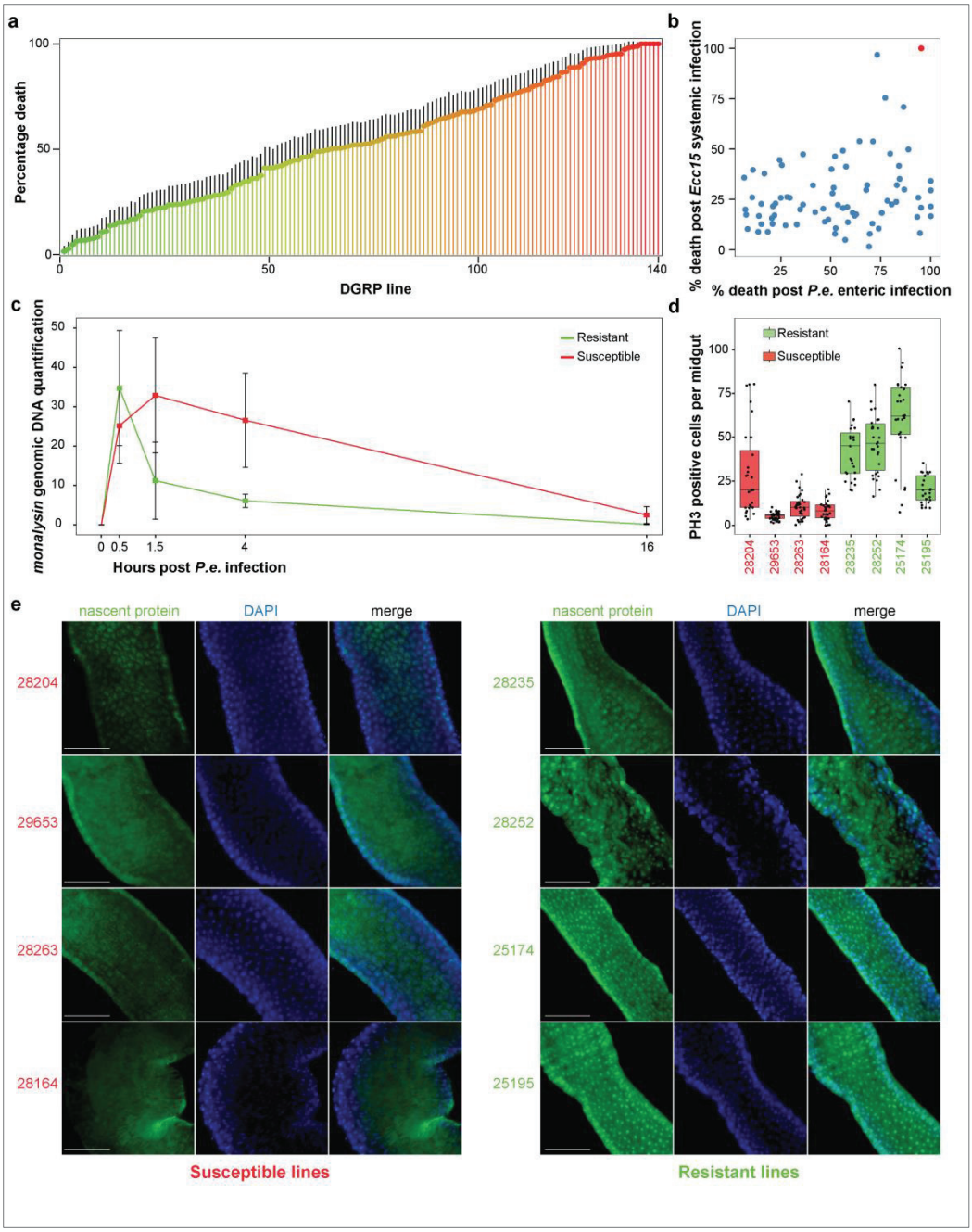


Figure 2:1 **Susceptibility to infection is highly variable among DGRP lines and multifactorial.**

(a) Bar graph showing for each of the 140 DGRP lines (x-axis) the percentage of dead female flies (y-axis) 3 days post-enteric infection with *P.e.* (OD 100). Data shown are averages from three biological replicates (± SE of the proportion; n > 60 females/line). (**b)** A scatter plot of 78 DGRP lines revealing an absence of correlation in proportion death between enteric (by 3 days post *P.e.* ingestion) and systemic (by 10 days post septic injury with *Ecc15*) infection. DGRP line *#25745* (red) is highly susceptible in both conditions and features a rare mutation in the *dredd* gene. (**c)** Quantification of *P.e.*-specific *monalysin* genomic DNA by qPCR reveals differences in *P.e.* clearance between four susceptible

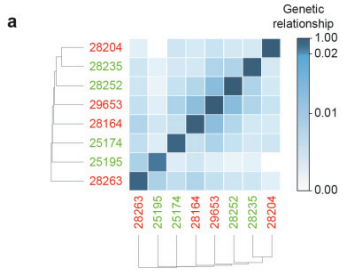
and four resistant DGRP lines over time (ANOVA p=0.00343 for the effect of susceptibility class; see **Methods** for details on statistics). (**d)** Quantification of PH3-positive cells per female midgut dissected 8 hours post enteric infection with *P.e.* reveals that infected resistant lines have more mitotically active stem cells than those of susceptible lines (n>30 guts/line; ANOVA p<0.00001 for difference between susceptibility classes). (**e**) Measurement of the incorporation of a methionine analog, L-azidohomoalanine (green staining), in the R2 region(Buchon, Osman et al. 2013) of the anterior midgut shows that susceptible lines are not able to synthetize proteins after infection in contrast to resistant lines. Note that while the same midgut region was sampled, no gross morphological differences in the shape or regionalization of the gut can be observed between resistant and susceptible flies after infection. However, this does not rule out subtle differences at the cellular level.

2.2.2 Characterization of lines from the phenotypic extremes

We then assessed the dynamics of intestinal pathogen colonization and clearance in the same eight DGRP lines as used for the *PA14* infection experiment. Here, we quantified *P.e.* genomic DNA in the guts of the flies at different time points post-infection (**Fig. 2:1c**), providing new insights into the colonization behaviour of *P.e.* in the fly gut. Resistant and susceptible lines exhibited no significant difference in intestinal *P.e.* loads 30 minutes post infection, corroborating the results of the feeding assay. In addition, both classes of lines were able to clear *P.e.* from the gut after approximately 16h (**Fig. 2:1c**), suggesting that the impact of enteric infection with *P.e.* on survival is determined by the initial pathogen exposure and not persistence. Importantly, the rate of clearance was different between the two phenotypic classes with resistant lines reducing intestinal *P.e.* levels much faster than susceptible lines (ANOVA p=0.0033 for susceptibility class). This indicates that rapid eradication of *P.e.* as an immediate defence response could play a role in the final outcome of the infection. In *Drosophila* laboratory strains, *P.e.* infection causes severe irreversible intestinal epithelial damage in comparison to other pathogens (Jiang, Patel et al. 2009, Chakrabarti, Liehl et al. 2012). Specifically, *P.e.*-induced inhibition of protein synthesis in the gut impairs both immune and repair programs leading to low epithelial renewal (Chakrabarti, Liehl et al. 2012). We examined whether the two DGRP phenotypic classes exhibit differences in protein synthesis and, as a consequence, variations in gut regenerative capability by measuring intestinal stem cell division, a quantitative readout of epithelial renewal. We found that guts of resistant lines are still able to translate proteins and induce a greater number of mitotic stem cells than those of susceptible lines (**Fig. 2:1d and fig. 2:1e**). Collectively, our findings indicate that *P.e.* infection does not always lead to lethality caused by translation inhibition as previously suggested (Chakrabarti, Liehl et al. 2012), re-emphasizing the importance of host genetic background in determining the response to as well as outcome of infection.

2.2.3 Genetic architecture of susceptibility to enteric infection

It is conceivable that physiological and survival differences between resistant and susceptible lines are a mere consequence of high genetic relatedness among lines from each phenotypic class. To explore this possibility, we used the available genetic relationship matrix for the eight DGRP lines (http://dgrp2.gnets.ncsu.edu/), but did not observe genetic clustering of phenotypic classes, as expected (Huang, Massouras et al. 2014) (**Fig. 2:2a**). However, a significant part of the observed variation in survival is due to genetic factors as the heritable component estimate is 0.61 (**Methods**). To gain insights into the genetic architecture of survival, we performed a complete diallel cross, where we generated all possible hybrid combinations by crossing the eight lines to each other. We then measured their susceptibility to *P.e.* infection. The F1 progeny from crosses between different resistant lines were resistant (**Fig. 2:2b**) and the F1 progeny from crosses between different susceptible lines were mainly susceptible, thus there was no evidence of consistent heterosis. The lack of resistance appearing in crosses between susceptible lines implies that susceptibility is not a mere consequence of inbreeding depression. Moreover, F1 progeny from crosses between resistant and susceptible lines tended to exhibit an intermediate susceptibility phenotype as expected when there are additive effects. Indeed, an analysis of the diallel cross data (**Supplementary Table 2:2)** revealed both additive effects reflected in general combining ability (p=0.00001) and dominance effects reflected in specific combining ability (p<0.00001)(Griffing 1956). There were also various interactions between strains due to male and female parental combinations (**Supplementary Table 2:2)**, suggesting that the extent of susceptibility depends on the specific combination of strains tested. In general, these patterns indicate that natural variation in survival to infection is partly additive, but also depends on the combination of strains being crossed, suggesting a complex genetic architecture.



b

		Female							
		Resistant				Susceptible			
	line #	28235	28252	25174	25195	28164	28263	29653	28204
Male	28235	**9%**	6%	3%	0%	20%	14%	13%	22%
	28252	5%	**6%**	0%	0%	13%	5%	0%	25%
	25174	0%	5%	**4%**	2%	24%	24%	4%	10%
	25195	0%	0%	0%	**0%**	13%	2%	2%	3%
	28164	28%	17%	44%	8%	**74%**	64%	58%	47%
	28263	28%	8%	37%	2%	67%	**53%**	71%	62%
	29653	29%	2%	9%	8%	63%	76%	**82%**	25%
	28204	19%	9%	14%	3%	63%	64%	33%	**79%**

Figure 2:2 **Gut immunocompetence is a largely additive, complex trait.**

(**a)** The genomic relationship matrix shows an absence of genetic relatedness among either resistant or susceptible lines respectively. (**b)** Percentage death for F1 flies in a full diallel cross between four susceptible and four resistant DGRP lines (by 3 days post enteric infection with *P.e.* (OD 100)).

2.2.4 Genome-wide association study for survival to infection

To uncover genetic determinants underlying immunocompetence, we performed a genome-wide association study (GWAS) on survival using both a non-parametric (**Fig. 2:3a**) and parametric test (**Supplementary Fig. 2:4a**). Unlike a previous study dealing with survival to viral infection in DGRP lines in which one quantitative trait locus (QTL) with large effect was identified (Magwire, Fabian et al. 2012), we obtained 27 QTLs at an arbitrary *p*-value of 10^{-5}, even though there was no clear point of departure from expectations in the Q-Q plot (**Supplementary Fig. 2:4b**). The results were largely consistent between both GWAS analysis procedures and a maximum of 19% of the phenotypic variance could be explained by a single QTL (**Supplementary Table 2:3**). The small sample size and the truncated distribution from which QTLs are chosen to estimate effect sizes can result in an overestimation of the proportion of variance explained, a phenomenon known as the 'Beavis effect' (Beavis 1998). This could be further exacerbated by linkage between SNPs (**Supplementary Fig. 2:4a**). To account for redundancy between linked SNPs, we also performed an iterative multiple-SNP regression (Harbison, McCoy et al. 2013). Interestingly, as few as four SNPs can explain ~50% of the phenotypic variance (**Supplementary Table 2:4**). Moreover, we performed a permutation analysis to evaluate the Beavis effect. In short, we sampled groups of lines of different sizes, ranging from 70 to 140, and performed multi-SNP regression. For each sample size, we performed 100 permutations with random resampling (**Supplementary Fig. 2:5**). We found that the proportion of variance explained, R^2, decreases as the sample size increases, as expected, yet starts levelling-off at larger sample sizes, suggesting that the correct proportion of variance accounted by the SNPs is being approached at the larger sample sizes.

The most significant QTLs were located in the *Neurospecific receptor kinase* (*Nrk*) gene, which belongs to an evolutionarily conserved stress-response network from *Drosophila* to mammals (Kirienko and Fay 2010). One of the three linked *Nrk* QTLs (**Supplementary Table 2:3**), which explains 14% of the phenotypic variance, is a non-synonymous polymorphism (p=3.6×10^{-06}) in residue 306 of the protein (G or V). The minor allele (15% frequency) appears to be the ancestral allele since it is found in the four closest sequenced *Drosophila* species. Interestingly, lines harbouring this minor allele were mainly susceptible (**Fig. 2:3b**). To test if *Nrk* affects the antibacterial immune response, we measured the activity of the Imd pathway reporter

Diptericin-lacZ (*Dpt-lacZ*) (Tzou, Ohresser et al. 2000) in wild-type and *Nrk* knock-down flies. In contrast to infected control guts, where *Dpt-LacZ* reporter was induced in the cardia and anterior midgut, *Nrk* knock-down flies have markedly reduced *Dpt-lacZ* activity (**Fig. 2:3c**). We also investigated the knockdown effect of several other genes that harboured strong QTLs with *Gyc76C* producing the most robust and greatest reduction in *Dpt-lacZ* activity (**Supplementary Fig. 2:6c**). *Gyc76C* contains a QTL (p=1.86×10^{-05}) that explains 15% of the variance (**Supplementary Table 2:3**), and has recently been described as a modulator of the Imd pathway in response to salt stress in the Malpighian tubules(Overend, Cabrero et al. 2012). Susceptible DGRP lines carrying the G-allele of the QTL expressed *Gyc76C* at higher levels than resistant lines (A-allele) post infection (**Fig. 2:3d**). Remarkably, endogenous *Dpt* transcript induction followed a similar trend (**Fig. 2:3e**). Knocking down *Gyc76C* expression specifically in enterocytes of adults also showed that *Gyc76C* diminishes *Dpt* induction (**Fig. 2:3e**) and reduces fly survival after enteric infection (**Fig. 2:3f**). Since Gyc76C is a membrane receptor capable of the activation and nuclear translocation of the Imd transcription factor Relish(Overend, Cabrero et al. 2012), it may activate the Imd pathway in the gut independent of PGRP-LC, the canonical Imd pathway receptor. Taken together, these results suggest that our GWAS identified at least two novel genes that are capable of modulating gut immunocompetence and that were not previously implicated in canonical gut immune response pathways.

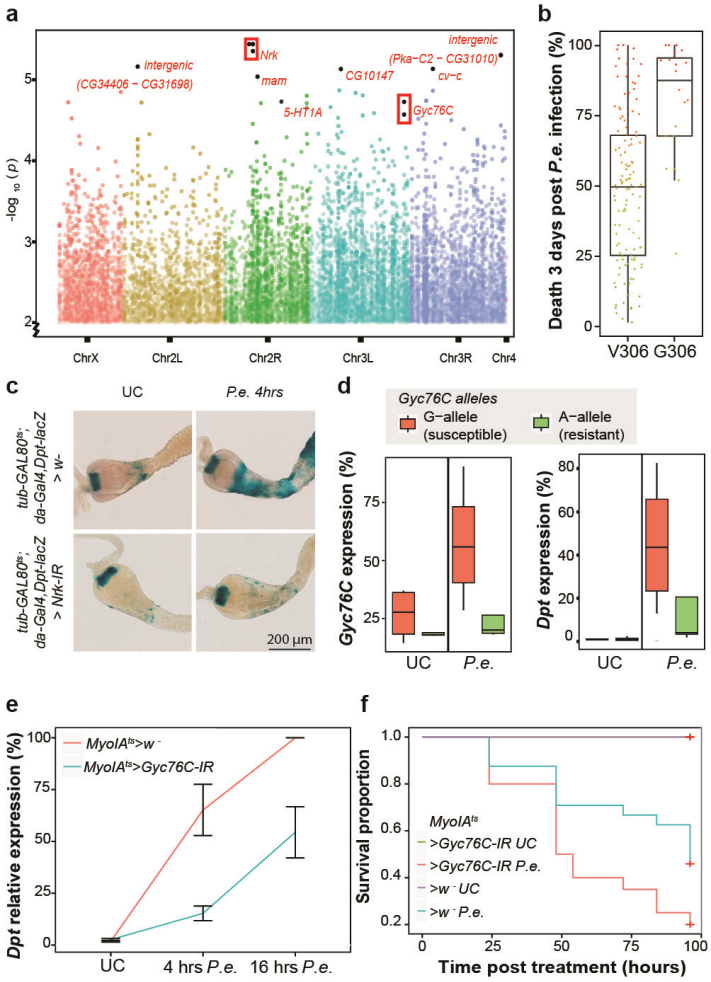


Figure 2:3 **GWAS reveals genetic loci underlying susceptibility to infection.**

(a) Manhattan plot of the *p*-values (y-axis) for the association between genomic variants in DGRP lines and *P.e.* susceptibility. A non-parametric Kruskal-Wallis test was performed using proportion death at day 3 as phenotype. The x-axis represents the genomic location. Multiple variants in a single gene are bounded by a box. (**b)** Susceptibility of DGRP lines grouped by the *Nrk* allele (GWAS *p*=3.6e-6) that changes the coding sequence at position 306 of the protein (at chr2R:9048897). Note that *D. simulans*, *D. sechelia*, *D. yakuba*, and *D. erecta* all have the variant G allele. (**c)** Knock-down of the top GWAS hit, *Nrk*, using a ubiquitous driver (*da-gal4*) highly reduces the activity of the immune activation reporter *Dpt-lacZ* in the gut as revealed with X-Gal staining (*P.e.* OD 50 was used to avoid the anticipated inhibition of translation effect of *P.e.* at OD 100(Chakrabarti, Liehl et al. 2012)). UC = unchallenged flies. (**d)** RT-qPCR experiments on gut total RNA from females show that four susceptible DGRP lines harbouring the G-allele at the *Gyc76C* locus (chr3L:19769316) express *Gyc76C* at higher levels after *P.e.* infection, in comparison to resistant lines carrying the A-allele. *Dpt* transcript induction is higher in susceptible DGRP lines carrying the G-allele in *Gyc76C* (ANOVA *p* for allele effect in the challenged condition for *Gyc76C* and *Dpt* is 0.00205 and 0.0344, respectively). (**e**) *Gyc76C*

knockdown in enterocytes using the thermosensitive *MyoIA-gal4* driver shows that *Gyc76C* regulates the induction of *Dpt* transcript in the gut 4 and 16 hours post *P.e.* infection (ANOVA p=0.00741 for line effect; error bars represent standard deviation around the mean of three replicates). (**f)** Survival analysis of females that are orally infected with *P.e.* shows a lower survival rate of *MyoIA^{ts}>Gyc76C-IR* flies compared to wild-type (Log-Rank test p=0.0351 for comparison between *Gyc76C* knockdown and wild-type in challenged condition). For (**d**-**f**) data is based on at least three independent biological replicates.

2.2.5 Transcriptomic analysis of phenotypic extremes

Variability in survival and physiology among DGRP lines could in part be explained by system-specific transcriptional differences. We therefore performed RNA-seq on 16 gut samples comprising the same four susceptible and four resistant lines as introduced above in the unchallenged condition and 4h after *P.e.* infection (**Supplementary Fig. 2:7a**). 1287 genes were differentially expressed 4h post-infection compared to the unchallenged condition when all eight lines were treated as replicates (FDR adjusted p-value<0.05 and two-fold change, **Supplementary Table 2:7**). This set of genes overlaps with what we have previously shown when characterizing the gut transcriptional response to *P.e.* infection, even though that analysis was carried out using microarrays and on a different genetic background (*Oregon^R*)(Chakrabarti, Liehl et al. 2012). However, when we looked for differences in gene expression between the four resistant and four susceptible lines by pooling the samples of each susceptibility class, very few genes exhibited significant differential gene expression. Specifically, the expression of only 5 and 34 genes were changed in the unchallenged and challenged guts respectively when comparing phenotypic classes (**Fig. 4a, Supplementary Table 2:8**). This may reflect reduced statistical power given the large number of genes that are compared. In addition, it is possible that small but systematic differences in gene expression collectively differentiate resistant from susceptible profiles. We therefore performed principal component analysis (PCA) on the 2000 genes with the highest expression variance in the 16 transcriptomes. Since infection status has a large impact on the transcriptome, expression profiles derived from infected samples were separated from those of unchallenged samples on the first principal component (PC), which explains 53% of the variance (**Fig. 2:4b**). Strikingly, even prior to infection, profiles of resistant lines were separated from those of susceptible lines based on the second PC, which explains 7.3% of the variance (**Fig. 2:4b**). This separation implies that the basal intestinal transcriptional state of resistant lines is distinct from that of susceptible lines, which may either define or reflect a molecular pre-disposition to enteric infection susceptibility. To dissect the molecular signatures that underlie this transcriptional stratification of the two phenotypic classes, we performed modulated modularity clustering (Ayroles, Carbone et al. 2009) on the same 2000 genes. We identified 24 transcriptional modules including more than 15 correlated genes (**Fig. 2:4c, Supplemen-**

tary Table 2:9). Based on Gene Ontology analysis and manual annotation(Huang da, Sherman et al. 2009), we assigned the genes within the modules to six functional groups (**Fig. 2:4d**). To identify those modules whose gene levels clearly separate the lines according to treatment and phenotypic class, we systematically performed PCA on each module by taking the expression levels of its genes (**Fig. 2:4e**). We found that in module #96, samples are clearly separated on the first PC, even though the probability for such a separation to spuriously occur is less than 3 in 10,000 (**Fig. 2:4e, Supplementary Fig. 2:7b,c**). This module contains 20 genes, of which 9 are related to stress response and most notably to reactive oxygen species (ROS) metabolism (**Fig. 2:4e,f**) and collectively explains 29% of the observed phenotypic variation (**Supplementary Table 2:5**). Other modules such as #102 (16 genes) also separated the samples on the first two PCs (**Supplementary Fig. 2:8**). Interestingly, module #102 likewise contains several ROS-related genes such as *Cyp6a9* and *Thioredoxin-2* (*Trx-2*)(Tsuda, Ootaka et al. 2010). ROS are essential signalling molecules and immune effectors that are induced by the infected gut to neutralize pathogens (Ha, Oh et al. 2005) and promote intestinal renewal (Amcheslavsky, Jiang et al. 2009). However, a high ROS load can also cause inhibition of protein translation and consequently severe intestinal damage (Chakrabarti, Liehl et al. 2012), necessitating a finely tuned regulation of ROS production and metabolism (Schieber and Chandel).

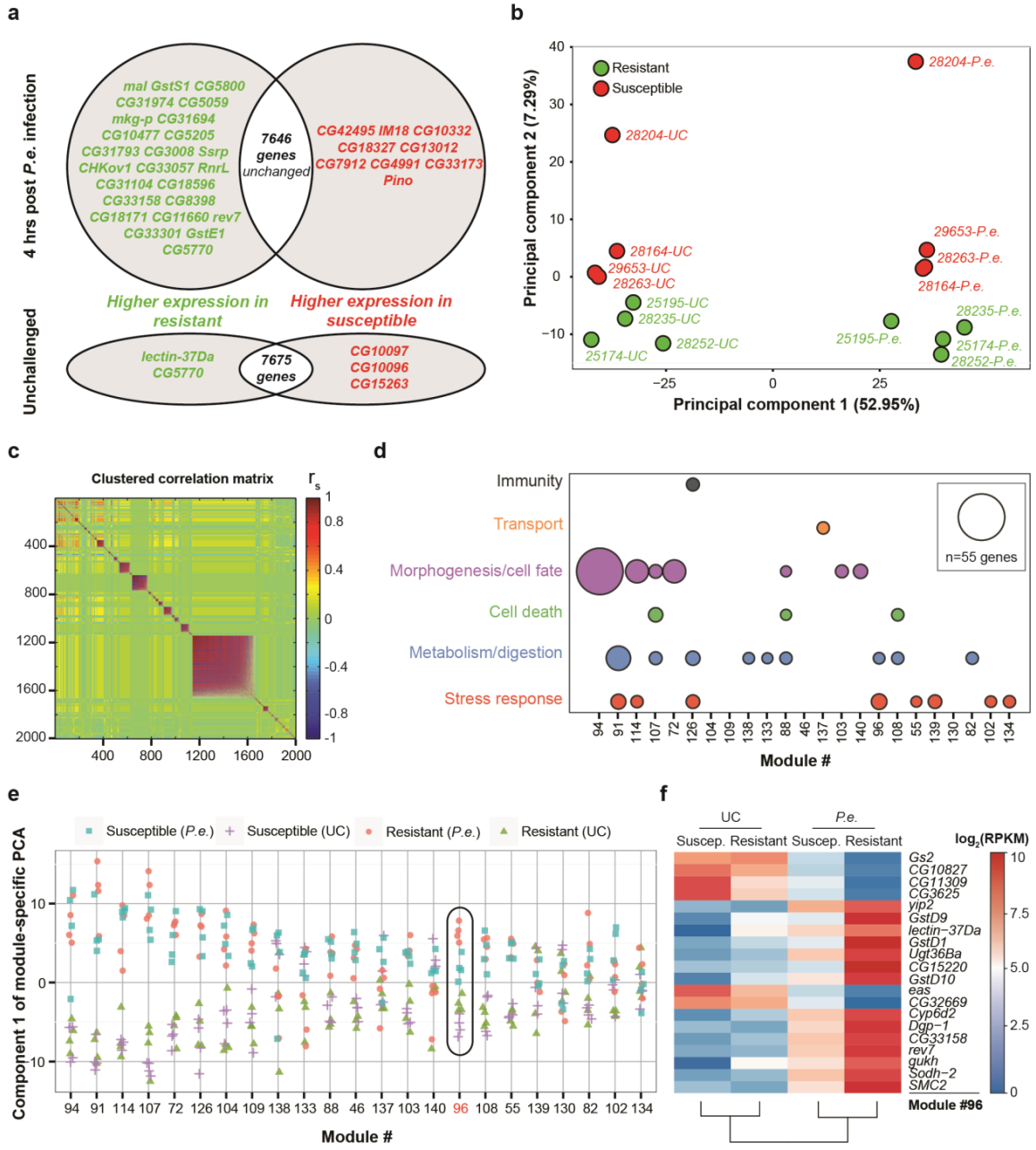


Figure 2:4 **Specific gene expression signatures define susceptibility to bacterial enteric infection.**

(**a**) Venn diagram showing differentially expressed genes (as revealed by RNA-seq experiments) between four resistant and four susceptible DGRP lines, in the unchallenged condition and 4 hours post *P.e.* infection (*q*-value<0.2, two-fold change). Genes in red and green have higher levels in susceptible and resistant lines respectively. The number of genes (black) indicated in the intersections represents the total number of non-differentially expressed genes. (**b)** Principal Component Analysis (PCA) on the top 2000 varying genes between the 16 samples reveals that resistant lines

cluster separately from susceptible lines, prior (*UC*) to and post *P.e.* infection (*P.e.*). PC1 separates samples based on treatment whereas PC2 separates them based on susceptibility class. (**c)** Modulated modularity clustering analysis on the top 2000 varying genes identifies 24 correlated transcriptional modules (n≥15 genes). Each coloured point represents the spearman correlation (r_s) between two genes. (**d)** A selection of functional categories identified by GO analysis of genes belonging to the different modules identified in **c** (excluding the largest module with n=523, **Supplementary Table 9**). For the GO analysis, we used the Database for Annotation, Visualization, and Integrated Discovery (DAVID). (**e)** PCA using the expression levels of genes within each of the 24 modules identifies module #96 as the only module for which the lines are clearly separated on the first principal component according to treatment and susceptibility. (**f)** Heat map of gene expression levels in module #96 reveals important differences across susceptibility classes and treatment conditions.

2.2.6 A role for ROS in variation in susceptibility

To investigate the physiological relevance of ROS in mediating inter-individual differences in gut immunocompetence, we compared ROS levels in resistant versus susceptible lines (**Fig. 2:5a,b**). Importantly, ROS levels were significantly lower in resistant lines in both conditions (ANOVA p=2.98x10^{-7} for susceptibility class in unchallenged condition, and p=1.43x10^{-11} in challenged condition). This may reflect a more efficient ROS metabolism in resistant lines, possibly mediated by the higher expression levels of the majority of genes in the focal module #96 compared to susceptible lines (**Fig. 2:4f**). Since too much ROS inhibits translation and epithelial renewal resulting in lethality (Chakrabarti, Liehl et al. 2012), it appears that resistant lines utilize ROS in a more effective and less noxious manner than susceptible lines (**Fig. 2:1c,e**). To investigate this hypothesis, we evaluated the survival of the same lines to ingestion of paraquat, a ROS-catalyzing chemical reagent. Most susceptible lines showed higher lethality compared to resistant lines (**Fig. 2:5c**), supporting the role of ROS as one of the principal components underlying variation in gut immunocompetence.

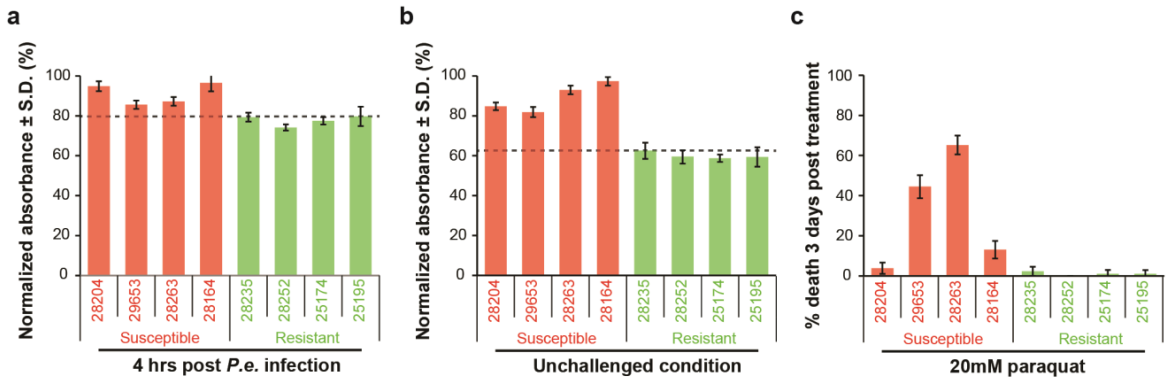


Figure 2:5 **Diversity in ROS metabolism is a feature of variable susceptibility.**

(a-b) Measurement of ROS activity in flies before and after *P.e.* infection reveals lower ROS levels in resistant compared to susceptible DGRP lines (mean normalized absorbance ± SD, n=5 females per line and replicate, 3 replicates, ANOVA $p<0.0001$ for difference between susceptibility classes in both conditions). The dashed line marks the maximum level in resistant lines. **(c)** Percentage of dead female flies three days after Paraquat treatment. Percentages are averages from three experiments (± SD, n>60 females/line, ANOVA $p<0.0001$ for difference between susceptibility classes).

2.3 Discussion

Direct exposure to environmental insults such as pathogens has driven the alimentary canal to establish numerous protective and homeostatic mechanisms (Buchon, Broderick et al. 2013). Considerable efforts have been invested in characterizing mechanisms underlying intestinal immunity using model organisms like *Drosophila*. However, most of these studies identified genes with large effects involved in canonical immune pathways (Lemaitre and Hoffmann 2007). The aim of our study was to go beyond these classical analyses to uncover first of all the extent of inter-individual variation in gut immunocompetence and in a subsequent step the underlying genetic and molecular determinants. We found striking differences in the overall susceptibility to enteric infection, not only in survival, but also in related physiological aspects including bacterial load, stem cell activity, and infection-induced inhibition of translation. A first important implication of these findings is that the outcome of classical *Drosophila* genetics experiments involving standard laboratory strains may not always be generalizable to all wild-type strains. Indeed, while the use of such standard strains is valuable to increase reproducibility, a downside is that it may lead to conclusions

that are only true in specific genetic backgrounds (Linder 2001, Wolfer, Crusio et al. 2002) as we demonstrate here for pathogen-induced inhibition of translation (or lack thereof) in DGRP lines.

This phenomenon likely reflects the inherently complex nature of traits like gut immunocompetence since they are the result of the interplay of many biological processes, each of which could be affected by many genomic loci with small to medium effects. The results from our GWAS analysis are consistent with this hypothesis as they suggest that relatively common alleles located in various parts of the genome drive gut immunocompetence in additive manner. If rare variants resulted in reduced survival to infection in susceptible lines, then crossing two susceptible lines should have resulted in a resistant hybrid. Moreover, deleterious mutations affecting gut immunocompetence could be under strong purifying selection, further reinforcing a genetic architecture of multiple loci with relatively small effects (Houle, Morikawa et al. 1996, Merila and Sheldon 1999).

A consequence of such a genetic architecture is that it renders the prediction of a trait from genotypic information difficult. An attractive approach to improve phenotypic predictions is the complementation of genetic data with *in vivo* measurements of molecular parameters since the latter may yield mechanistic insights that may not be immediately obvious from GWAS analyses, which, similar to our study, are often performed on rather coarse-grained phenotypic read-outs (such as survival here)(Lehner 2013). Our finding that the transcriptomes of resistant and susceptible extremes can be separated by PCA even before infection is interesting in this regard, as it suggests that there are systematic molecular differences underlying susceptibility to enteric infection. This observation also implies that with a large-enough sample size, signatures of susceptibility could be mined from the data for both a better biological understanding and prediction of gut immunocompetence. In this study, we provide a proof of concept by clustering correlated transcripts into modules and identifying a candidate module linked to ROS metabolism. While the involvement of ROS in intestinal infection and homeostasis has been previously established (Ha, Oh et al. 2005, Buchon, Broderick et al. 2009, Chakrabarti, Liehl et al. 2012, Buchon, Broderick et al. 2013, Lee, Kim et al. 2013), it is particularly intriguing that it may also be one of the important factors that either mediate (or reflect) naturally occurring variation in gut immunocompetence, since lines from the phenotypic extremes contained significantly different intestinal ROS levels even before infection and reacted distinctly after exposure to the ROS-inducing chemical paraquat. As such, ROS levels, which are an indirect measure of stress, may have phenotype-predictive value, irrespective of whether differential ROS levels are a cause or a consequence of differences in gut immunity. Better utilization of ROS by the resistant lines may then constitute a tolerance

rather than an active resistance mechanism (Schneider and Ayres 2008). But clearly, alleles for low tolerance have persisted in the population and we speculate that the underlying mechanisms could be conceptually similar to variation in immunity, where environmental heterogeneity and fitness trade-offs limit the effect of natural selection(Lazzaro and Little 2009).

Since enteric infection has a major impact on human and animal health, resolving the genetic and physiological contributions underlying continuous variation is of great importance. This is particularly the case in the developing world where almost 20% of child deaths can be linked to a pathogenic invasion of the intestine(Flores and Okhuysen 2009). In many cases, this invasion is by opportunistic pathogens on immunocompromised individuals, who might have a functioning innate immune system like AIDS patients (Amancio, Japiassu et al. 2013). In addition, enteric infections by opportunistic *Pseudomonas* species have been reported in hospitalized patients (Driscoll, Brody et al. 2007, Markou and Apidianakis 2013). Understanding the role of genetic variation in innate immunity could therefore shed more general light on susceptibility to opportunistic pathogens (Muszynski, Nofziger et al. 2014) including members of the *Pseudomonas* genus(Driscoll, Brody et al. 2007). Our study now reveals that identifying causal factors may present a substantial challenge in that the observed, overt physiological differences between resistant and susceptible lines appear to be driven by multiple genetic effects. We therefore postulate that a promising strategy could be the identification of transcriptional modules as informative biomarkers of disease susceptibility given their inherent dependence on the interaction between a genome and its environment. Alternatively, since transcriptome analyses are expensive diagnostic tools, knowledge gained from the study of transcriptional modules could be used in the discovery of novel biomarkers. Such insights into the molecular determinants of gut immunocompetence may help in developing control programs in invertebrate disease vectors as well in better understanding the mechanisms underlying variability in susceptibility to enteric infections in human populations.

2.4 Materials and Methods

2.4.1 Fly stocks

DGRP lines were obtained from the Bloomington stock center and reared at room temperature on a standard fly medium. The fly medium recipe that we used is the following (for 1L water): 6.2g Agar powder (ACROS N. 400400050), 58.8g Farigel wheat (Westhove N. FMZH1), 58.8g yeast (Springaline BA10), 100ml grape juice; 4.9ml Propionic acid (Sigma N. P1386), 26.5 ml of Methyl 4-hydroxybenzoate (VWR N. ALFAA14289.0) solution (400g/l) in 95% ethanol, 1L Water. For RNAi (IR) studies, F1 progeny carrying one

copy of the *da-Gal4* or *MyoIA-Gal4* with *tub-Gal80^{ts}* transgenes (and *Diptericin-lacZ* reporter in the case of *da-Gal4*) as well as one copy of *UAS-IR* (all in the *w^{1118}* background) were kept at 18°C for three days post-eclosion, and then moved to 29°C for 8 days to activate the *UAS-IR*. The *UAS-Gyc76C-IR* line is a gift from Julien Dow, the *UAS-Nrk-IR* (CG4007 R2 and R3) fly lines were obtained from the DGRC stock center. Imd pathway mutants used are *Dredd^{B118}* (Leulier, Rodriguez et al. 2000) and *Relish^{E20}* (Hedengren, BengtÅsling et al. 1999).

2.4.2 Infection, Paraquat treatment, and Survival experiments

Pseudomonas entomophila (P.e.) and *Erwinia carrotovora carrotovora 15 (Ecc15)* pathogens were cultured in LB medium at 29°C overnight. *Pseudomonas aeruginosa* clinical isolate *PA14* was cultured in Brain Heart Infusion broth at 37°C overnight. For enteric infection, 3-5 day old females were first starved 2-3h at 29°C, and then transferred into vials with fly medium covered with filter disks soaked in a mix of bacterial pellet at $OD_{600\ nm}$ of 100 and 1.5% sucrose. For survival analysis, flies were transferred onto a fresh fly medium 24 hours post-infection, and maintained on a fresh and healthy medium during the survival assay. For Paraquat treatment, the same procedure as oral infection was followed except for the addition of a solution of 20mM Paraquat dichloride hydrate (FLUKA Analytical #36541) in 1.5% sucrose instead of the bacterial pellet. For systemic *Ecc15* infection, adult flies were pricked in the thorax with a tungsten needle that had been dipped into a concentrated bacterial pellet with an $OD_{600\ nm}$ of 200.

2.4.3 RT-qPCR

Total RNA was extracted from 20 guts including the crop, the cardia and the midgut using TRIzol reagent (Invitrogen). Malpighian tubules were removed from the samples. cDNA was then synthesized from 1 ug total RNA using *SuperScript II* enzyme (Invitrogen). qPCR experiments were performed with a Light-Cycler® 480 machine and the SYBR Green I kit (Roche). Relative gene expression was calculated after normalization to the control *RpL32* mRNA. Given the polymorphic nature of the DGRP lines, we assured that the primers did not target sites with polymorphisms. The primer sequences are available in Supplementary Table 6.

2.4.4 Bacterial load measurement

Flies were orally infected with *P.e.* and then transferred to a fresh medium 30 minutes post-infection. The DNA fractions were then isolated at indicated time points using the TRIzol manufacturer's protocol (Invitrogen). The bacterial load quantification was then assessed by qPCR with *P.e. monalysin*-specific primers (Opota, Vallet-Gely et al. 2011) (Supplementary Table 6). Normalization has been performed on the host *RpL32* DNA.

2.4.5 Assessment of nascent protein synthesis

To assess the levels of protein translation in susceptible and resistant guts, we used the Click-iT AHA for Nascent Protein Synthesis commercial kit (Invitrogen). Flies were orally infected for 16 hours as described above, but by adding AHA reagent at 50 μM as final concentration to the infection mix. Guts were then dissected in 1X PBS Triton 0.3%, fixed for a minimum of 30 min in PBS 4% paraformaldehyde, and finally washed with PBS Triton 0.3%. DAPI reagent (Sigma) was used to stain DNA. The R2 region (Buchon, Osman et al. 2013) of the gut was visualized with an Axioplot imager (Zeiss).

2.4.6 PH3 staining

Guts were dissected in Grace's insect medium (life technologies) and fixed for 15-20 minutes in PBS 4% paraformaldehyde. They were subsequently washed in PBS 0,1 triton (PBT), blocked in PBT 0,1% BSA (PBTA) for 1 hour, and then incubated 2 hours at 4°C with primary and secondary antibodies in PBTA. Antibody used was 1/500 rabbit anti-PH3 (Millipore), 1/500 Alexa-594 anti-rabbit (life technologies).

2.4.7 ROS measurement

To assess homeostatic ROS level as well as *P.e.*-induced ROS, we used the Amplex Red reagent (Invitrogen # A12222) as described previously(Lee, Kim et al. 2013), by incubating 6 flies of each genotype with 100 µl of reaction buffer (pH 7,4) and 0,25 Unit/ml of horseradish peroxidase (Sigma) for 1 hour at 37°C. The fluorescence was measured in a microplate reader at 550 nm.

2.4.8 Genome wide association analysis

We performed two genome wide association studies. The first was performed on angle transformed proportion death at day 3 using PLINK v1.07 (Purcell, Neale et al. 2007). Specifically, means of three repeats per line were taken as phenotype, and only biallelic SNP markers were considered. We calculated empirical p-values by using default adaptive permutation settings. The other GWAS was performed directly on the proportion data using a non-parametric Kruskal-Wallis one-way analysis of variance by ranks test. In this pipeline, all variants can be considered, including non-SNPs, even if they are not biallelic. Specifically, we grouped overlapping variants for each line, creating a list of loci with two or more alleles in the population with a minimum allele count of 10. We then grouped the phenotypic measurement according to the allele of its line and performed a Kruskal-Wallis test. For each variant, 1000 permutations of the phenotype data were performed to estimate the false discovery rate. Since our GWAS hits are of marginal significance, the false discovery rate within this range of *p*-values is high (for example, at *p*-value ≤ 2e-05, the FDR is 0.66). Nevertheless, the two approaches yielded very similar candidate lists. For the multiple-SNP GWAS, please refer to the legend in Supplementary table 4.

2.4.9 RNAseq analysis

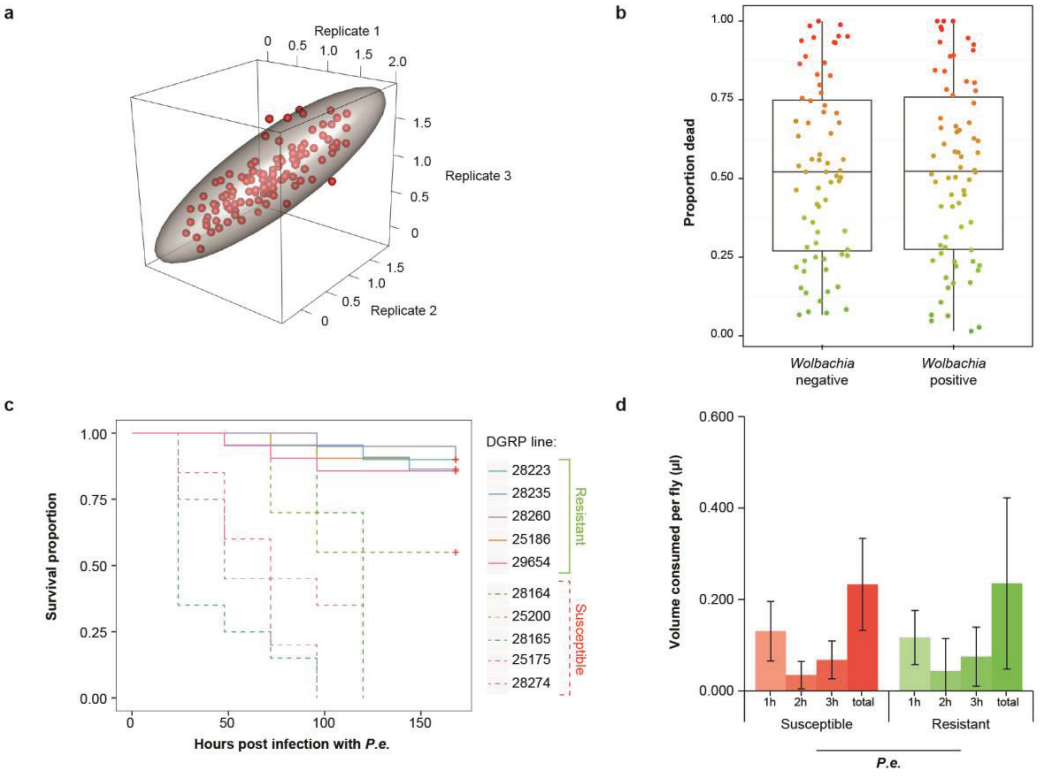
Four resistant (Bloomington #*28235*, #*28252*, #*25174*, #*25195*) and four susceptible DGRP lines (Bloomington #*28164*, #*28263*, #*29653*, #*28204*) were selected for RNA-seq experiments. These eight lines were infected 4 hours with *P.e.* as indicated above, in parallel, the same eight lines were kept on 1.5% sucrose as controls. 25 guts for each of the 16 samples were dissected and subsequent TRIzol RNA extraction was performed. We chose the 4h post-infection time point for multiple reasons. First, we have previously shown that major changes occur in the transcriptome as early as 4h post infection. Importantly, these changes are not restricted to immediate immune responses, but extend to the homeostatic mechanisms like intestinal stem cell-induced regeneration and repair. So we reasoned that differences between resistant and susceptible lines could be resolved by that time. Another motivation for this choice stems from the fact that *P.e.* does not persist in the gut, and therefore, resistant lines could return to an uninfected state relatively quickly. In addition, fly mortality is still low to non-existent at 4h post-infection in susceptible lines. Libraries were prepared using the Illumina Truseq RNA kit and sequenced for 100 cycles on the Illumina HiSeq 2000 in the University of Lausanne Genomic Technologies Facility. Post processing was performed using Casava 1.82. There was an average of 25 million reads per sample. Reads were mapped to individual DGRP-predicted transcriptomes (Massouras, Waszak et al. 2012). Count data was normalized using the Voom package in R. Each gene's RPKM value was calculated by averaging the RPKM values of its

associated transcripts. Analysis of differential expression was performed using limma (Smyth 2005). Gene RPKM values were used to perform principal component analysis using the FactoMineR package. Modulated modularity clustering was performed as in (Mackay, Stone et al. 2009) on the RPKM values of the 2000 genes with the largest variance. We used the R built-in heatmap function with default settings for mean gene expression levels by phenotypic class in module #96. The raw and analysed expression data is available on GEO through this accession: GSE59411

2.4.10 Quantitative genetic and statistical analyses

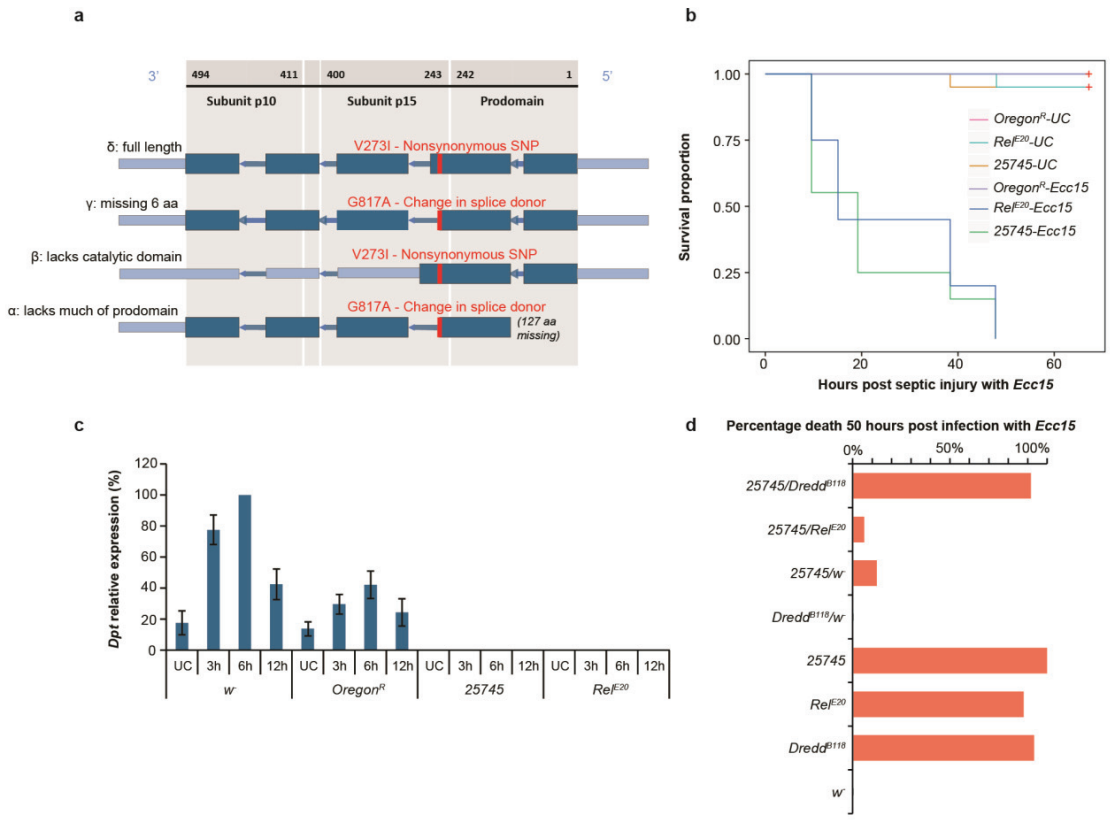
All statistical analyses were performed in R version 3.0.2 unless otherwise noted. We used angular transformation on percentage death data in all parametric analyses. For calculating the heritable component, we treated the transformed percentage death at day 3 as a Gaussian response in a random effects model of the form $Y = \mu + L + R + \varepsilon$ where μ is the mean proportion death of all lines, L is a random variable representing deviation of each line from the mean, R is a random variable representing the deviation of each line's biological replicate from the line mean, and ε is the residual error. We assumed that all variation is additive and that there is no epistasis and estimated the heritable component as V_A/V_A+V_E, where V_A is the additive genetic variance and is equal to half the between-line variance, V_L, since the lines are almost entirely homozygous and V_E is the environmental variance such that $V_E= V_R+V_\varepsilon$. To estimate the proportion of variance accounted for by a certain QTL, we calculated R^2 by performing linear regression taking the SNPs as factors. Pearson's product moment correlation between oral infection and septic injury was performed on the angular transformed line means between oral infection at day 3 and septic injury at day 10. For the bacterial load experiment, we analysed $\log_2$ relative ratios to *Rpl32* values using ANOVA where the line was nested in susceptibility class and treated as a fixed effect, time post infection was treated as a fixed effect, and experimental replicate was treated as a random effect. Nested ANOVA, where line is nested within susceptibility class, was used to compare the $\log_2$ transformed PH3 counts of the susceptibility classes. For the analysis of the effect of RNAi knockdown of *Gyc76C* on *Dpt* induction, ANOVA was used with genotype and time post infection as fixed effects and experimental replicate as a random effect. Separate nested ANOVA by condition was used to determine the effect of susceptibility class on ROS levels (normalized absorbance) where line was nested in susceptibility class and treated as a fixed effect and experimental replicate was treated as a random effect. We used the R built-in heatmap function with default settings to plot the genetic relationship matrix data.

2.5 Supplementary Materials



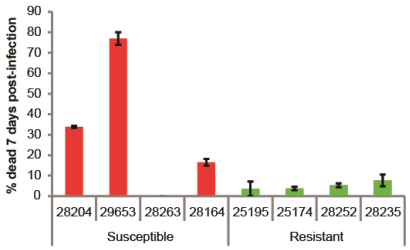
Supplementary Figure 2:1 **Feeding behaviour, *Wolbachia*, and microbiota do not have a major influence on susceptibility to enteric infection.**

(a) The three survival experiment repeats represented in a three-dimensional scatter plot showing proportion deaths (after angular transformation) three days post infection. Each red point is a DGRP line and the confidence ellipsoid is in grey. (**b)** *Wolbachia* infection status does not correlate with susceptibility (Nested ANOVA p=0.51 for *Wolbachia* status effect on survival). 68 lines and 70 lines are *Wolbachia* negative and positive, respectively. (**c)** Flies that were either resistant or susceptible to enteric infection in non-axenic conditions were infected with *P.e.* under axenic conditions. Absence of the endogenous intestinal microbiota does not alter the relative susceptibility of the DGRP flies. (**d)** A Capillary Feeder (CAFE) assay shows that susceptible and resistant DGRP flies ingest a comparable volume of bacteria during the first three hours post *P.e.* infection.



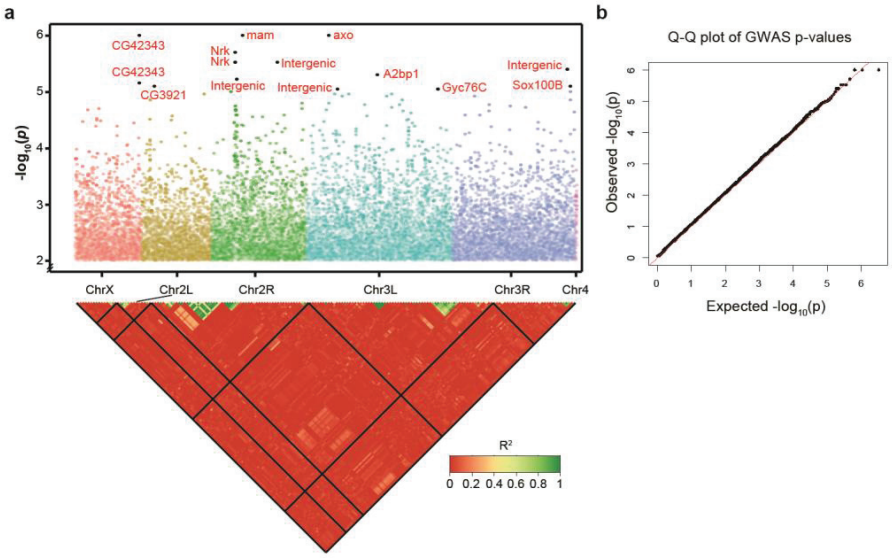
Supplementary Figure 2:2 **Identification of a loss of function mutation in the *dredd* locus in one DGRP line.**

(a) Four isoforms related to the *dredd* gene have been previously described in (Di Fruscio, Styhler et al. 2003). γ and δ isoforms differ only by six amino acids. The α isoform lacks much of its prodomain and the β isoform lacks its catalytic domain. One SNP has been identified in the *dredd* locus of the DGRP line *#25745*, causing a change in the splicing donor site (G817A) in the α and γ mRNA, or an amino acid change (V273I) in δ and β isoforms. The light blue colours represent non-coding regions, the dark blue ones depict exons. **(b)** Survival analysis of females systemically infected with *Ecc15* shows a lower survival rate of the *#25745* line and *relish* mutant (Rel^{E20}) compared to controls (Log-Rank test $p<0.05$). **(c)** RT-qPCR experiments show that, similar to *relish* mutants, the *#25745* line systemically infected with *Ecc15* has no detectable *diptericin* (*Dpt*) expression as shown in w^- and $Oregon^R$ control flies. Data is normalized to 100% ± S.D. *w-* flies consistently had the highest level of *Dpt* induction (100%), hence the missing error bar. **(d)** Percentage of dead female flies 50 hours post *Ecc15* systemic infection is monitored. Only complementation of *#25745* line with a *dredd* mutant line fails to restore the wild-type survival, revealing that the identified SNP in the *dredd* gene is the causal locus of susceptibility to bacterial infection. Data presented in **b** and **c** are derived from three independent replicates.



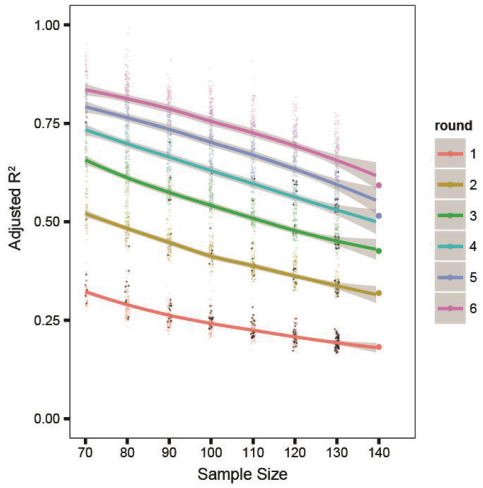
Supplementary Figure 2:3 **Lines resistant to *P. entomophila* are also resistant to a clinical isolate of *Pseudomonas aeruginosa*.**

Bar chart showing the proportion of dead flies after 7 days post-infection (± s.d.; three biological replicates). The lines in the susceptible and resistant classes were identified based on their susceptibility to *P. entomophila* oral infection.



Supplementary Figure 2:4 **Different statistical approaches yield highly similar GWAS top hits.**

(a) Above: Manhattan plot of the *p*-values (y-axis) for the association between genomic variants in DGRP lines and *P.e.* susceptibility. The x-axis represents the genomic location. A linear model was implemented in PLINK using angular-transformed proportion death at day 3 as phenotype. **Below:** heatmap of pairwise LD between all SNPs with a p-value $< 10^{-4}$ (n=188). **(b)** Q-Q plot of the linear association.



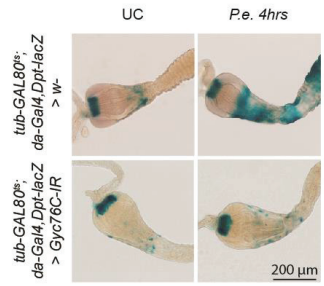
Supplementary Figure 2:5 **Illustration of the Beavis effect.**

A plot of the adjusted R^2 values obtained through random sampling of lines with different sample sizes (100 random samples per size group) and multi-SNP association (six rounds of association). The curves are loess fits with 95% confidence interval, and black points correspond to SNPs that have been identified in the full population.

a

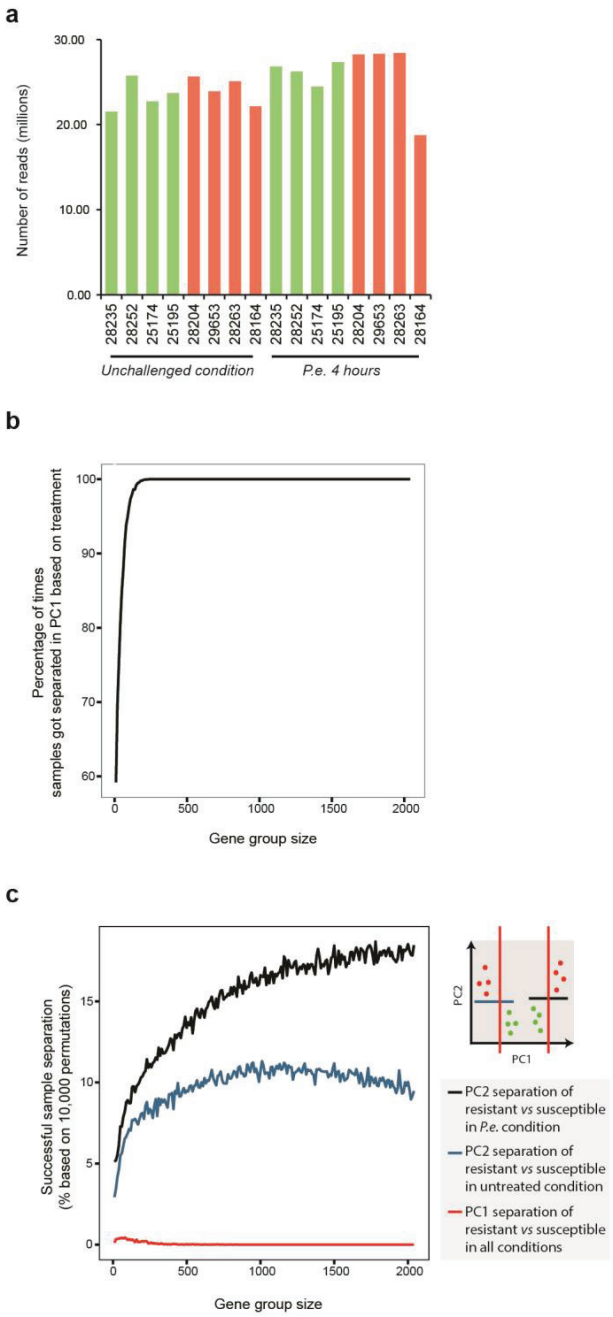
tub-GAL80ts;da-Gal4,Dpt-lacZ	UAS-RNAi	GWAS *p*-value	Dpt-LacZ induction
	Nrk	3.60E-06	-
	cv-c	7.28E-06	++
	CG10147	7.32E-06	+
	mam	9.10E-06	-
	5-HT1A	1.85E-05	++
	Gyc76C	1.86E-05	-
	control	-	+

b



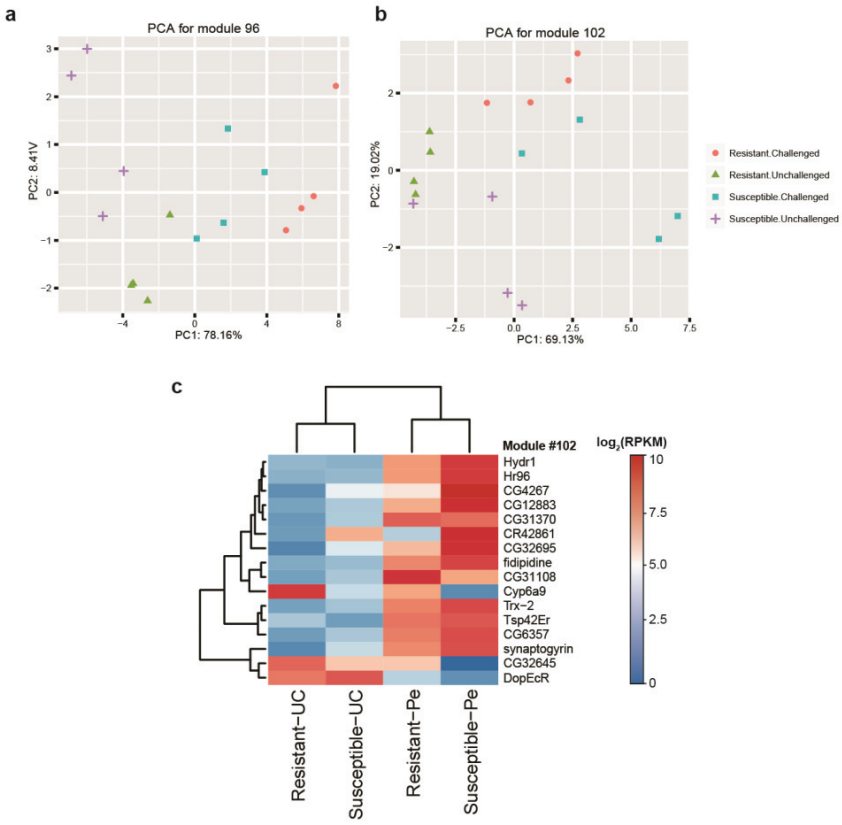
Supplementary Figure 2:6 **Validation of candidate genes.**

(a) UAS-RNAi lines screened for an effect of *dipt-LacZ* reporter induction under a ubiquitous driver (*da-gal4*). "+" and "-" indicate higher and lower induction than control (w^{1118}), respectively, and the number of +'s scales with the extent of induction. (**b**) Knock-down of the top GWAS candidate gene, *Gyc76C*, using *da-gal4* highly reduces the induction of the immune activation reporter *Dpt-lacZ* in the gut as revealed with X-Gal staining.



Supplementary Figure 2:7 **Permutations of random sampling followed by PCA of the RNA-seq data.**

(a) RNA-seq library sizes of the 16 samples used in the study. **(b)** Random sampling of gene groups with sizes ranging from 10 to 2000 (10,000 permutations per group size), followed by PCA analysis on their gene expression levels revealed that treated and untreated samples are always separated by the first PC for groups greater than 250. **(c)** The same random sampling and PCA as in **b**, but with different separation criteria (see legend).



Supplementary Figure 2:8 **Principal component analysis of modules**

(a)#96 **and (b)**#102**. (c)** Heatmap of average expression levels of genes in module #102 by susceptibility/treatment (unchallenged = UC or infected = *Pe*) class.

2.6 Supplementary Tables

Supplementary Table 2:1 **Percentage death of tested DGRP lines 3 days post-infection with *Pseudomonas entomophila***

DGRP#	Bloomington stock number	Percentage dead at day 3
DGRP-897	28260	0.5%
DGRP-802	28235	1.8%
DGRP-320	29654	4.3%
DGRP-738	28223	4.8%
DGRP-208	25174	7.6%
DGRP-857	28252	4.5%
DGRP-486	25195	10.0%
DGRP-129	28141	7.2%
DGRP-313	25180	7.2%
DGRP-360	25186	2.4%
DGRP-303	25176	10.5%
DGRP-142	28144	9.7%
DGRP-907	28262	15.0%
DGRP-217	28154	12.7%
DGRP-801	28234	16.3%
DGRP-379	25189	20.8%
DGRP-158	28147	24.4%
DGRP-237	28160	16.6%
DGRP-441	28198	16.9%
DGRP-440	28197	22.0%
DGRP-426	28196	11.7%
DGRP-399	25192	24.7%
DGRP-321	29655	23.1%
DGRP-894	28259	23.1%
DGRP-45	28128	20.3%
DGRP-335	25183	20.7%
DGRP-307	25179	25.2%
DGRP-837	28246	27.2%
DGRP-91	28136	21.5%
DGRP-161	28148	25.8%
DGRP-705	25744	35.5%
DGRP-377	28186	29.3%
DGRP-822	28244	22.0%
DGRP-804	28236	30.9%
DGRP-861	28253	21.8%
DGRP-799	25207	32.7%
DGRP-812	28240	37.4%
DGRP-356	28178	23.1%
DGRP-370	28182	27.9%
DGRP-373	28184	40.8%
DGRP-437	25194	33.5%
DGRP-195	28153	26.3%
DGRP-406	29657	24.1%
DGRP-318	28168	45.0%
DGRP-136	28142	34.3%
DGRP-41	28126	42.7%
DGRP-461	28200	27.7%
DGRP-805	28237	25.7%
DGRP-517	25197	46.7%
DGRP-563	28211	46.1%
DGRP-352	28177	53.6%
DGRP-75	28132	51.4%
DGRP-315	25181	50.4%
DGRP-642	28216	49.3%
DGRP-737	28222	50.1%
DGRP-371	28183	56.3%
DGRP-391	25191	48.8%

DGRP-859	25210	41.9%
DGRP-256	28162	44.9%
DGRP-42	28127	51.0%
DGRP-855	28251	48.9%
DGRP-362	25187	44.5%
DGRP-884	28256	52.0%
DGRP-350	28176	52.2%
DGRP-513	29659	41.0%
DGRP-808	28238	33.3%
DGRP-177	28150	52.7%
DGRP-786	25206	57.3%
DGRP-783	28230	49.7%
DGRP-375	25188	52.6%
DGRP-374	28185	63.0%
DGRP-381	28188	55.0%
DGRP-508	28205	54.1%
DGRP-820	25208	63.3%
DGRP-832	28245	52.6%
DGRP-57	29652	55.9%
DGRP-83	28134	68.3%
DGRP-492	28203	41.6%
DGRP-589	28213	40.9%
DGRP-239	28161	57.7%
DGRP-309	28166	68.2%
DGRP-796	28233	65.0%
DGRP-427	25193	71.1%
DGRP-304	25177	70.4%
DGRP-555	25198	72.1%
DGRP-26	28123	72.9%
DGRP-324	25182	57.3%
DGRP-491	28202	77.3%
DGRP-310	28276	50.4%
DGRP-712	25201	64.9%
DGRP-892	28258	57.4%
DGRP-380	25190	77.7%
DGRP-332	28171	15.2%
DGRP-409	28278	59.5%
DGRP-595	28215	82.8%
DGRP-776	28229	68.4%
DGRP-338	28173	59.2%
DGRP-392	28194	77.6%
DGRP-181	28151	58.6%
DGRP-509	28206	63.9%
DGRP-730	25202	80.2%
DGRP-732	25203	82.9%
DGRP-233	28159	77.7%
DGRP-109	28140	87.0%
DGRP-176	28149	85.1%
DGRP-911	28264	68.8%
DGRP-358	25185	88.1%
DGRP-365	25445	80.7%
DGRP-879	28254	79.2%
DGRP-28	28124	86.8%
DGRP-359	28179	95.3%
DGRP-531	28207	67.8%
DGRP-790	28232	77.9%
DGRP-502	28204	94.9%
DGRP-228	28157	88.8%
DGRP-405	29656	93.8%
DGRP-153	28146	84.4%
DGRP-639	25199	96.6%
DGRP-818	28241	93.0%
DGRP-882	28255	95.7%
DGRP-714	25745	98.3%
DGRP-535	28208	98.2%

DGRP-38	28125	89.4%
DGRP-386	28192	96.0%
DGRP-890	28257	93.4%
DGRP-761	28227	66.8%
DGRP-138	28143	80.5%
DGRP-721	28220	93.6%
DGRP-101	28138	96.8%
DGRP-40	29651	99.4%
DGRP-229	29653	99.5%
DGRP-908	28263	99.6%
DGRP-280	28164	100.0%
DGRP-287	28165	100.0%
DGRP-301	25175	100.0%
DGRP-85	28274	100.0%
DGRP-227	28156	65.8%
DGRP-707	25200	100.0%
DGRP-765	25204	64.4%
DGRP-774	25205	93.5%

Supplementary Table 2:2 **Analyses of variance for diallel survival data (after angular transformation).**

Effect	df	Mean Square	F	P
ANOVA on male/female effects due to resistant/susceptibility category				
Male resistance category	1	5.477	271.819	<0.001
Female resistance category	1	2.266	38.159	0.001
Male strain (nested within category)	6	.020	.275	0.946
Female strain (nested within category)	6	.060	.807	0.570
Male category x Female category	1	.814	11.025	0.002
Male strain x Female strain	46	.075	2.813	<0.001
Replication	63	1.715	64.183	<0.001
Diallel ANOVA testing for general and specific combining ability				
General combining ability	7	0.264	6.682	<0.001
Specific combining ability	28	0.215	5.444	<0.001
Reciprocal	28	0.166	4.184	<0.001
Maternal	7	0.13	0.735	0.64287
Maternal interaction	21	0.177	4.481	<0.001
Error	63	0.04		

The first ANOVA tests for effects due to male/female strain and susceptibility class (susceptible or resistant) and their interactions on survival. Strain was nested within the resistant or susceptible categories and treated as a random variable. The second ANOVA represents the diallel analysis according to Griffing (1956)(Griffing 1956) testing for general combining ability (additive effects and their interactions) and specific combining ability (dominance effects and their interactions) as well as effects due to reciprocal differences in the crosses, maternal contributions, and their interactions.

Model for ANOVA: $Y_{ijklm} = \mu + m_i + f_j + s_{k(i)} + t_{l(j)} + m_i f_j + s_{k(i)} t_{l(j)} + e_{ijklm}$ where μ is the population mean, m_i is the ith male category, f_j is the jth female category, $s_{k(i)}$ is the kth male strain within the male category, $t_{l(j)}$ is the lth female strain within the female category and e_{ijklm} is the residual. Strain within categories are random, other terms apart from replication are fixed.

Model for diallel analysis: $Y_{ijklm} = \mu + g_i(g_j) + s_{ij} + r_{ij} + m_i + n_{ij} + e_{ijk}$ where μ is the population mean, $g_i(g_j)$ is the general combining ability for the ith (jth) parents, sij is the special combining ability for the cross between the ith and jth parents, r_{ij} is the reciprocal effect, m_i is the maternal effect, n_{ij} is the interaction of the ith maternal effect with the jth parent, and e_{ijk} is the error term. The analysis follows Method 1 (parents and reciprocal F1s measured) under Model 1 of Griffing (1956)(Griffing 1956) with maternal terms added (Cockerham and Weir 1977, Kaushik and Puri 1984).

Supplementary Table 2:3 **Summary of top QTLs obtained in common between parametric and non-parametric association studies.**

Genomic location	Variant annotation	Kruskal-Wallis p valuea	PLINK empirical p-valueb	Number of per-mutationsc	R2 d
Chr2R:9048826	Nrk (intron)	3.60E-06	3.00E-06	1000000	0.14
Chr2R:9048897	Nrk (exon V306G)	3.60E-06	3.00E-06	1000000	0.14
Chr2R:9048840	Nrk (intron)	4.40E-06	2.00E-06	1000000	0.14
Chr3R:26527712	Intergenic - Pka-C2(dist=4852),CG31010(dist=2770)	4.93E-06	4.00E-06	1000000	0.15
Chr3R:26527703	Intergenic - Pka-C2(dist=4843),CG31010(dist=2779)	4.93E-06	4.00E-06	1000000	0.15
Chr2L:3172873	Intergenic - CG34406(dist=123);CG31698(dist=411)	6.83E-06	3.10E-05	1000000	0.12
Chr3R:10229978	cv-c (intron)	7.28E-06	1.70E-05	1000000	0.13
Chr3L:6480167	CG10147 (exon, synonymous)	7.32E-06	1.20E-05	1000000	0.13
Chr2R:9892328	mam (intron), CG30482 (exon)	9.10E-06	1.00E-06	1000000	0.16
Chr3L:6076155	Intergenic - CG6619(dist=1520),CG13293(dist=4214)	1.35E-05	5.85E-05	752247	0.11
Chr3R:10227723	cv-c (intron)	1.36E-05	2.20E-05	1000000	0.14
ChrX:21324090	CG42343 (intron)	1.41E-05	1.00E-06	1000000	0.19
Chr3L:9361423	CG4452 (intron)	1.45E-05	1.70E-05	1000000	0.12
Chr3L:10570926	A2bp1 (intron)	1.55E-05	5.00E-06	1000000	0.17
Chr2R:19991068	enok (exon, synonymous)	1.57E-05	1.60E-05	1000000	0.10
Chr2R:14967476	5-HT1A (intron)	1.85E-05	5.11E-05	861138	0.10
Chr3L:19769316	CG42637,Gyc76C (intron)	1.86E-05	9.00E-06	1000000	0.15
ChrX:4208879	mei-9 (3' UTR)	1.89E-05	3.40E-05	1000000	0.10
Chr2L:3794426	CG3921 (exon, synonymous)	1.90E-05	8.00E-06	1000000	0.15
Chr2R:10603181	Intergenic - mspo(dist=2055),CG12865(dist=23043)	1.94E-05	8.09E-05	544000	0.11
Chr2R:8613576	CG42663 (intron)	2.76E-05	1.00E-05	1000000	0.16
Chr2R:8613586	CG42663 (intron)	4.34E-05	1.00E-05	1000000	0.16
Chr2R:16288827	Intergenic - CG11192(dist=46270),CG12484(dist=23014)	5.18E-05	3.00E-06	1000000	0.15
Chr3R:5045687	pum (intron)	7.31E-05	5.67E-05	776000	0.11
Chr2R:12715416	CG34459(dist=1264), CG34460(dist=1013)	8.12E-05	2.80E-05	1000000	0.08
ChrX:12947763	CG12715 (exon, synonymous)	9.30E-05	1.48E-04	298402	0.10
Chr2L:8635001	Sema-1a (intron)	2.25E-04	2.70E-05	1000000	0.11

[a] Non-parametric association p-value

[b] Empirical p-value after adaptive permutation as implemented in PLINK(Purcell, Neale et al. 2007)

[c] Number of permutations performed for each SNP

[d] Linear model R^2 for single SNPs

Supplementary Table 2:4 **Additive multiple-SNP model results**

GWAS Round	Top SNP	Coefficient	p-value	Adjusted R2
1	Chr3L:4668479	-0.3251	1.46E-07	0.18
2	Chr2R:9892328	-0.2683	5.64E-07	0.32
3	Chr2L:3355610	0.4201	1.54E-06	0.43
4	Chr3L:13828661	-0.3401	1.90E-06	0.51
5	Chr2L:3355661	-0.4183	7.30E-07	0.52
6	Chr2L:2836880	0.3875	1.84E-06	0.59
7	Chr2L:2836903	-0.3888	2.00E-06	0.58
8	Chr3L:15759197	-0.1970	5.79E-06	0.64
9	Chr3R:15278253	0.1810	5.08E-06	0.69
10	Chr3R:15278255	-0.1810	5.08E-06	0.69
11	Chr3L:9600645	0.1600	5.35E-06	0.74
12	Chr2L:12809795	0.1499	1.25E-05	0.78
13	Chr3L:9680631	-0.1815	3.66E-06	0.83
14	Chr3R:9554355	-0.1739	2.10E-06	0.87
15	Chr3R:9554381	-0.1739	2.10E-06	0.87
16	Chr2L:18589931	0.1971	3.59E-05	0.90
17	Chr2R:10000342	0.1574	0.000117	0.91
18	Chr3L:3312435	-0.1575	0.000171	0.93
19	Chr2R:16922817	-0.1419	5.25E-05	0.94
20	ChrX:20010029	-0.1835	4.19E-05	0.95

Successive iterations of the GWAS were performed using a linear model of the form $Y = \mu + SNP_1 + SNP_2 + SNP_3 + \ldots SNP_N + \epsilon$, where SNP_1, SNP_2, SNP_3, …, SNP_N , are the most significant SNPs fitted in succession as in Harbison et al., 2013(Harbison, McCoy et al. 2013). In short, for each round a GWAS is performed and the SNP with the most significant QTL is recorded, which is then incorporated in the linear model of the next round.

Supplementary Table 2:5 **Multiple-SNP regression for SNPs in module #96**

geneID	GWAS p-value	snpID	Estimate	Std. Error	t	value
-	-	(Intercept)	0.33911	0.23743	1.428	0.15611
eas	1.40E-03	ChrX:16175381	0.13048	0.05661	2.305	0.02309
rev7	6.54E-01	Chr3R:1414703	0.13655	0.11685	1.169	0.24513
CG33158	5.54E-05	Chr3L:16415271	-0.22319	0.06629	-3.367	0.00105
Cyp6d2	1.11E-02	Chr2R:18540150	-0.10239	0.08927	-1.147	0.25393
CG10827	5.38E-03	Chr3R:16832600	-0.10546	0.07278	-1.449	0.15022
CG32669	7.07E-02	ChrX:10737211	0.06322	0.05821	1.086	0.27986
Gs2	4.12E-02	ChrX:11322919	-0.0274	0.06418	-0.427	0.67023
CG3625	8.59E-03	Chr2L:284365	-0.23906	0.0887	-2.695	0.00816
GstD10	8.17E-02	Chr3R:8191081	-0.02762	0.06186	-0.446	0.65618
yip2	4.24E-02	Chr2L:9915438	0.16849	0.12498	1.348	0.18044
SMC2	9.89E-02	Chr2R:10736815	-0.13882	0.10309	-1.347	0.18095
lectin-37Da	1.71E-02	Chr2L:19418365	-0.14842	0.1091	-1.36	0.17654
Dgp-1	5.59E-02	Chr2R:14057889	0.02383	0.10814	0.22	0.82603
GstD9	1.25E-01	Chr3R:8192383	0.20098	0.10987	1.829	0.07014
Ugt36Ba	1.01E-01	Chr2L:16794249	0.05927	0.06907	0.858	0.39268
CG11309	4.37E-02	Chr3L:21297350	0.08747	0.08353	1.047	0.29735
GstD1	1.42E-01	Chr3R:8194750	0.01066	0.1343	0.079	0.93691
gukh	3.36E-03	Chr3R:14827525	0.19141	0.09755	1.962	0.05233
Sodh-2	3.97E-02	Chr3R:6702928	0.06843	0.11717	0.584	0.56044
RPA3	1.28E-01	ChrX:11615178	0.06898	0.06521	1.058	0.29256

Residual standard error: 0.3029 on 108 degrees of freedom

(11 observations deleted due to missingness)

Adjusted R-squared: 0.2961

F-statistic: 3.693 on 20 and 108 DF, p-value: 5.569e-06

One SNP with the lowest GWAS p-value in the GWAS was chosen for each of the 20 genes in the module. The 20 SNPs were fitted simultaneously in a linear model of the form $Y = \mu + SNP_1 + SNP_2 + SNP_3 + \ldots SNP_{20} + \epsilon$.

Supplementary Table 2:6 **List of primer sequences used in the study**

Target	Forward primer	Reverse primer
diptericin	ACCGCAGTACCCACTCAATC	CACACCTTCTGGTGACCCTG
RpL32	GACGCTTCAAGGGACAGTATCTG	AAACGCGGTTCTGCATGAG
Gyc76C	AAACATCGGATGAGCAGGCA	GTGTAGTCGCAGCCACAGAT
monalysin	CTGGGTAATGGCCGACAAGT	ACAGAATGTGACGACCACCC

Please refer to the online publication for the following tables:

Supplementary Table 2:7 **Differential expression analysis between all challenged and all unchallenged samples**

Supplementary Table 2:8 **Analysis of genes differentially expressed in resistant versus susceptible lines.**

Supplementary Table 2:9 **Modulated modularity clustering modules.**

Chapter 3 The impact of gene expression *cis*-regulatory variation on the outcome of enteric infection in *Drosophila*

This chapter goes deeper into the dissection of systematic differences in gene expression between genetically distinct individuals before and after infection. Then it explores the effect of cis-regulatory variation, that is, genetic variations surrounding a gene, on gene expression levels. A special emphasis is placed on the prediction of an individual's susceptibility to infection based on gene expression levels, cis-variation, or a combination of the two. At the moment of writing this chapter, the project is still not complete, and there are several planned follow-ups with colleagues in the Deplancke lab.

Abstract

In a polymorphic population, complex traits like resistance to enteric infection can be affected by genetic variations affecting multiple genes and many pathways. Coming to grips with the molecular underpinnings of such traits in a variable world requires systems genetics approaches. Here, we use a *Drosophila* enteric infection model to study differences in gene expression between susceptible and resistant inbred lines in the naïve and infected state. With the exception of the gene *Nutcracker* (*ntc*), we find no consistent differences in gene expression levels between the two classes, indicating that the membership to a certain phenotypic class is mediated by small differences in many genes. By using statistical learning approaches, we identify gene signatures that reliably predict resistance class with 75% and 100% success in the naïve and infected state respectively. For each condition, we detect an equivalent number of expression quantitative trait loci (eQTLs), with 40% of genes being shared. Finally, we show that *ntc* has infection-specific eQTLs that not only correlate with its expression level, but also to susceptibility of other DGRP lines for which we do not have expression data. The eQTLs overlap with putative transcription factor binding sites around the ntc locus, which could mechanistically explain their effect.

Author Contributions and Acknowledgements

Maroun Bou Sleiman, Dani Osman, and Bart Deplancke designed the study. Maroun Bou Sleiman and Dani Osman prepared the RNA-sequencing samples. Maroun Bou Sleiman performed the statistical and

computational analyses with assistance from Tommaso Andreani. Michael Frochaux and Maroun Bou Sleiman performed infection and RT-qPCR experiments. The ntc^{ms771} stock was a kind gift from Professor Hermann Steller. Sequencing was performed in the the University of Lausanne Genomic Technologies Facility. The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics.

3.1 Introduction

Deciphering the relationship between genomic and phenotypic variability is one central question in genetics. Genome Wide Association Studies (GWAS) have been extensively used to address this question by looking for variations that could explain a certain fraction of the genetic variance of phenotypes (Manolio 2010). More often than not, those variants lie in non-coding regions of the genome, rendering the inference of their putative function very hard. Therefore, the study of intermediate traits, such as gene expression levels, and how they are affected by genomic variation is a powerful complementary tool to link genotype to phenotype (Nica and Dermitzakis 2013).

Ever since the first expression quantitative trait locus (eQTL) report on yeast in 2002 (Brem, Yvert et al. 2002), it was clear that eQTLs could explain variability in gene expression. eQTLs are more likely to be present in open chromatin regions and in transcription factor binding sites, and in cell-type specific regulatory elements (Gerrits, Li et al. 2009, Fairfax, Makino et al. 2012, Gaffney, Veyrieras et al. 2012). These observations collectively point to the importance of genetic variation in regulatory regions. Moreover, eQTLs can mediate different responses to external stimuli. For instance, studies in monocytes and dendritic cells that have been subjected to different stimuli have been successful in determining genetic variants that mediate the differential responses to stimulation (Lee, Ye et al. 2014). This has important implications in understanding the genetic basis of disease susceptibility since it could help pinpoint factors that mediate differences between individuals at a very high resolution, thus paving the way for personalized medical interventions.

In this study, we go beyond studies on cell lines and explore the effect of genetic variation on gene expression and the organismal phenotype in the context of enteric infection. We have previously shown that gut immunocompetence is highly variable and heritable in a set of 140 DGRP lines and characterized gene expression differences between 4 lines from each phenotypic extreme (Bou Sleiman, Osman et al. 2015). Here, we generated a larger set of gut transcriptomes in order to systematically investigate the link be-

tween gut expression levels and genetic variation. We show that the genotype is a major determinant of gene expression levels and that the resistance class can be predicted based on specific gene signatures. Then we catalogue the eQTLs that are in *cis* with expressed genes and identify *nutcracker (ntc*) as a gene that is differentially expressed between the resistance classes, probably due to *cis*-regulatory variation in transcription factor binding sites. The genetic tractability of the fruitfly, the ability to replicate experiments on the same genetic backgrounds, and the study at the whole organism level render our approach very powerful to understand enteric infection variability.

3.2 Results

3.2.1 Few or no genes are significantly different between resistance classes

We selected 38 DGRP lines, 20 of which are susceptible and 18 resistant to *P.e.* enteric infection (**Fig. 3:1a**), infected adult female flies, and performed RNA-sequencing on their dissected guts 4 hours post infection. For each line, we also sequenced guts of unchallenged, sucrose-fed flies. In total, we sequenced the poly-A enriched transcriptome of 76 samples. Since the lines have been shown to be highly polymorphic, we opted for analyses on individualized genomes. For that, we used the available genotype data(Huang, Massouras et al. 2014), including single nucleotide as well as indels and structural variations, to generate individualized genomes and gene annotations (see **Methods**) which we used throughout the analyses.

7 of the lines were already included in one of our previous study (Bou Sleiman, Osman et al. 2015), which allowed us to assess the biological reproducibility of the RNA-sequencing experiment. For that, we combined the expression count data from the two experiments, then performed normalization while accounting for the batch, and performed hierarchical clustering (**Supplementary Fig. 3:1a**). The samples from the same line and condition always cluster together, indicating that genotypic differences mediate expression-level differences and that batch effects are weaker than the infection or genotype effect. Principal Component Analysis (PCA) on the same data also supports this observation (**Supplementary Fig. 3:1b-c**).

Using standard gene-based differential expression analysis, we identified around 2400 genes that are either up- or down-regulated 4 hours post *Pe* infection (FDR<0.05, log fold change > 2, **Fig. 3:1b**). This is consistent with previous findings using microarray data (Chakrabarti, Liehl et al. 2012), as well as our previous RNA-

sequencing results (Bou Sleiman, Osman et al. 2015). In our previous study, however, we found very little differences between the resistance classes, and we had to relax the significance thresholds for exploratory purposes. This could have been due to either the small number of lines tested or to the possibility that there are few consistent differences between the classes at the single gene level. When comparing resistance classes of the 38 lines, we find no differentially expressed genes in the naïve state, and only one gene, *nutcracker* (*ntc*), in the treated state (**Fig. 3:1c**). This observation supports the hypothesis that the differences between the classes, while being very clear at the physiological and organismal level, cannot be explained at the single gene level.

To gain an unbiased overall insight into the retaledness of the transcriptomes, we performed PCA on the levels of expressed genes (**Fig. 3:1d**). While the infection effect is obvious and recapitulated by the first principal component (PC), lines from different resistance classes do not show any clear separation on the first two PCs. This is in contrast to our previous study, where we were able to see such separation on the second PC. Furthermore, performing PCA on the expression levels within conditions yields a similar result, with no obvious separation of the resistance classes on the first two principal components. One explanation to why we no longer see such a clear separation is the fact that we expanded the number of lines, therefore reducing the phenotypic spread. Another possibility is that the selected lines in the previous study show this separation due to genotypic effects and not specifically resistance class, and since there were no other samples to compare with at the time, a biological interpretation was performed. Taken together, our findings suggest that while the effect of infection is similar among all the tested lines and the phenotypic differences are striking between the two resistance classes, the underlying transcriptomic differences are neither evident at the single-gene nor the transcriptome-wide level. This is in line with our previous findings that higher-level modules could explain differences between resistance classes.

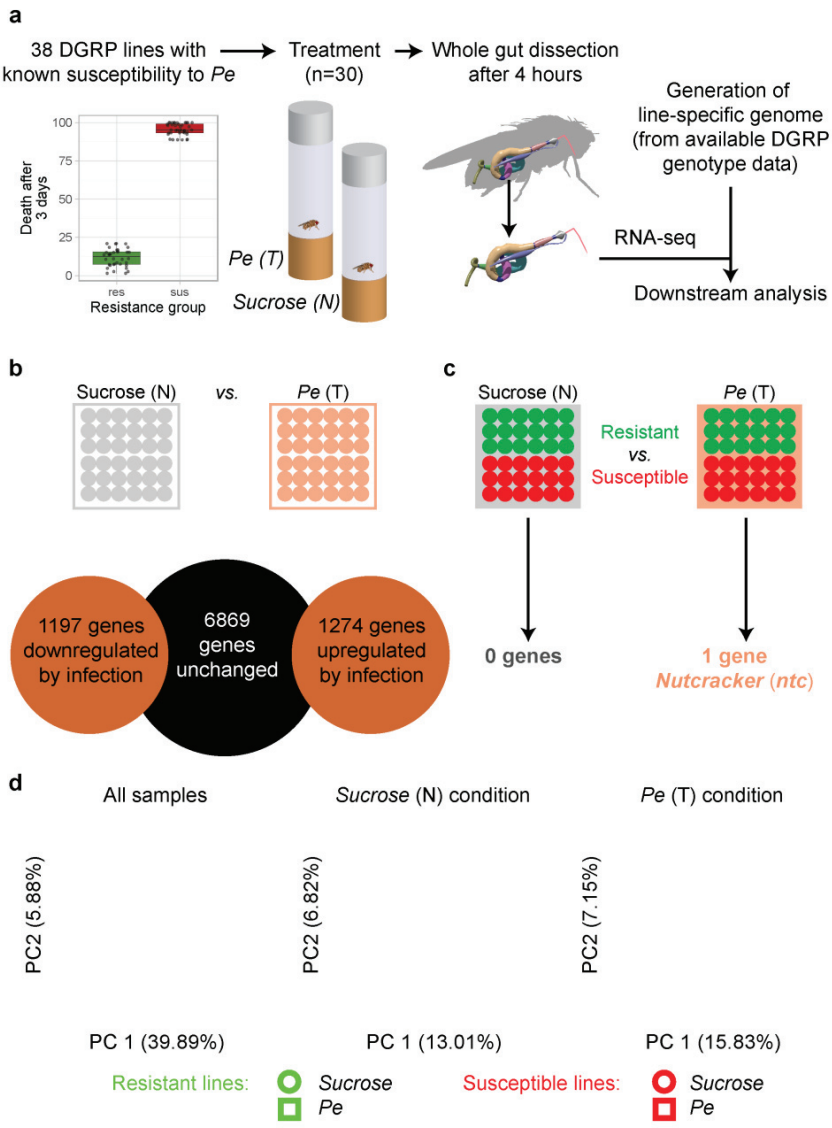


Figure 3:1 **Few or no genes are consistently different between fly resistance classes**

(a) Study design: Adult female flies from from two phenotypic extremes (18 resistant and 20 susceptible) of the DGRP were infected orally with *Pe*, or fed sucrose. Whole guts of ~30 flies were dissected per condition and line, then RNA-sequencing was performed. Sequencing reads were mapped to individualized genomes, and the number of reads was counted per gene. **(b)** Infection leads to the differential expression of around 2400 genes (BH-corrected p-value < 0.05, fold change > 2). **(c)** When lines of the two resistance classes are compared within condition, no genes are significantly differentially expressed in the naïve condition, and only one gene in the treated condition. **(d)** Principal component analysis plots of all the samples (left), the naïve condition (middle), and the treated condition alone (right). The R package FactomineR was used to obtain the coordinates of each sample in the first two components, as well as the variance explained by each component (in parentheses).

3.2.2 Feature selection and prediction of treatment condition and susceptibility class from the gut transcriptome.

Since differential gene expression analysis was unable to resolve clear differences between the susceptibility groups, we opted for a machine learning approach that capitalizes on the size of our dataset to look for differences across groups of genes. One basic limitation is that we do not have any prior knowledge of the potential number of features that could reliably lines from the phenotypic extremes. For that, we relied on elastic net regularization with cross validation on a training set of 31 lines from this dataset (7 lines were removed since they are also in the previous study), as well as prediction on the 8 lines from the previous study (test data). Briefly, we scanned possible values of the mixing parameter *alpha* (with *alpha* = 0 and *alpha* = 1 corresponding to ridge and lasso regression respectively) for the sparsest model that maximizes the prediction on a randomly-drawn validation set (size = 8 or 4) from the 31 lines. The best model for each *alpha* value was obtained through cross validation to select the model that minimizes *lambda*, the elastic net shrinkage parameter (**Fig. 3:2a**, **Supplementary Fig. 3:1**). To account for sampling variability in the selection of folds, we performed the analysis 100 times for each *alpha* value.

We first tested our approach with the easiest scenario: predicting treatment condition. Only three genes were sufficient to obtain 100% prediction of the sample's infection condition (**Fig.3:2b**). Interestingly, the three selected genes *Kayak/Fos-related antigen*, *Relish/NFkB*, and *Supressor of cytikine signaling at 36E* (*Socs36E*) are transcription factors involved in the Jun kinase (JNK), the Immunodeficiency (Imd), and the JAK/STAT pathways respectively. All three pathways have been previously implicated in the gut defense response, whether at the level of antimicrobial peptide induction(Lemaitre and Hoffmann 2007, Buchon, Broderick et al. 2013) (for the JNK and Imd pathways) or at the level of damage-induced stem cell proliferation(Jiang, Patel et al. 2009) (for the JAK/STAT pathway). The three genes are highly induced following infection and are therefore very good predictors of infection status. One should note that given the dramatic differences caused by *P.e.* infection, other combinations of genes could also have the same predictive power.

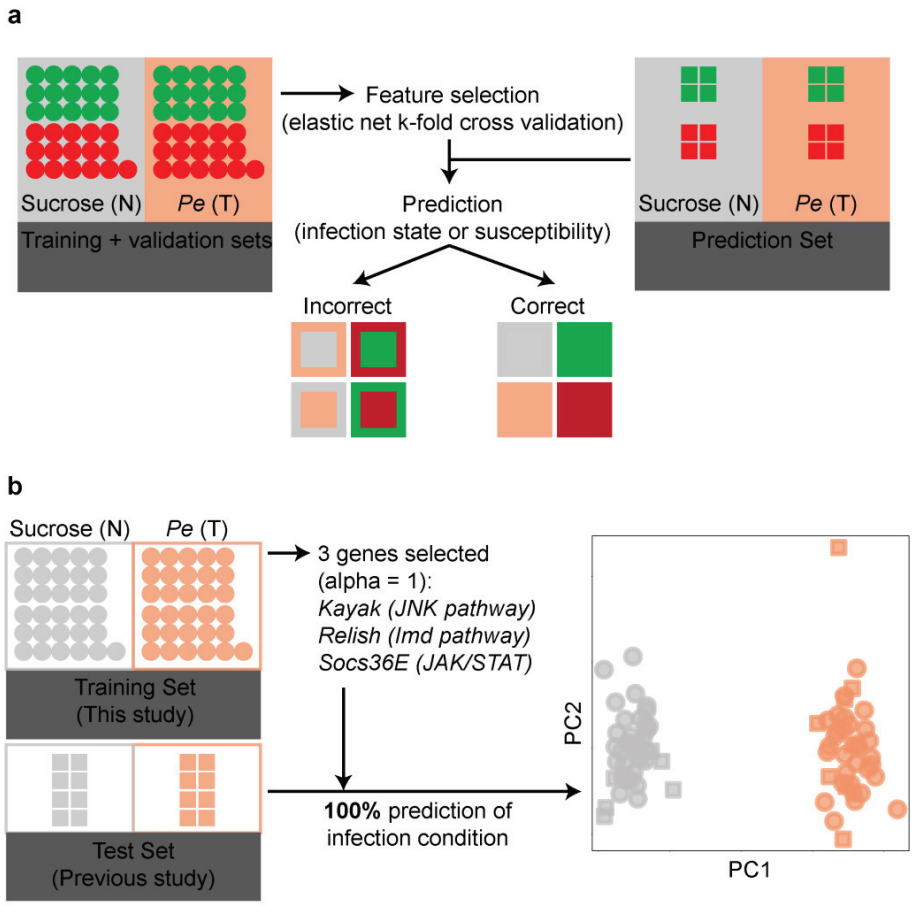


Figure 3:2 **Feature selection and prediction of treatment condition from the gut transcriptome.**

(a) The transcriptome data in this study was used as a training set, and the data from 8 lines from a previous study(Bou Sleiman, Osman et al. 2015) was used as a test set (see chapter 2). Samples from DGRP lines that are replicated in the two studieswere removed from the training set. For each scenario, elastic net regularization was used to select the most predictive set of features. Specifically, values of the mixing parameter α, ranging from 0 (*i.e.* ridge-regression) to 1 (*i.e.* lasso) were tested, and for each of those values, we used cross-validation within the training set to select the value of the elastic-net mixing parameter λ that minimizes the misclassification error in a logistic regression model. Then we selected the simplest model that could yield the best prediction on a validation set (chosen randomly from the training set). We used the obtained α value in a new round of cross-validation containing all the training set, and performed prediction on the test set. **(b)** Predicting treatment condition, which could be done at 100% rate with as few as 3 genes. A PCA that is weighted by the absolute value of coefficients of the model is drawn to show relationships between the lines or samples.

3.2.3 Resistance class can be fully predicted based on specific gene signatures.

Having established that our method could select a limited set of predictive genes from a full transcriptome, we next applied it to predict susceptibility in three different scenarios: in the naïve condition, the treated condition, and on the fold changes in gene expression after infection. Our ability to predict resistance class in the naïve state was modest, with 75% success based on 785 selected genes (**Fig. 3:3a**). The fact that so many genes were selected reflects that the resistance class signature in the naïve state is rather weak. Nevertheless, to gain insights into the possible function of the selected genes, we performed gene ontology (GO) enrichment analysis, and found that two of the most highly-enriched GO terms are proteasome-mediated ubiquitin-dependent protein catabolic process and regulation of organ growth.

Predicting resistance class in the treated condition was 100% successful based on a set of 409 genes enriched for ATP hydrolysis-coupled proton transport (**Fig. 3:3b**). The increased success could be due to slightly different responses of resistant lines compared to the susceptible lines which uncover condition-specific differences that are not discernible in the uninfected state. Moreover, the enrichment for ATP-hydrolysis-coupled proton transport is indicative of possible differences in Reactive Oxygen Species (ROS) metabolism. However, understanding the mechanistic basis of the contributions of each of those genes is a challenge, especially since they are numerous and subtle.

Finally, we performed prediction based on the fold changes of each gene in each DGRP line. Our method was also 100% successful in classifying the test set, this time based on a set of 457 genes. Two of the most enriched GO categories are SCF-dependent proteasomal ubiquitin-dependent protein catabolic process, to which the gene *ntc* belongs, and the regulation of mitotic cell cycle. Taken together, our analyses show that reliable predictive transcriptional signatures could be identified to predict susceptibility class yet the dissection of the relative roles of each gene is still a challenge.

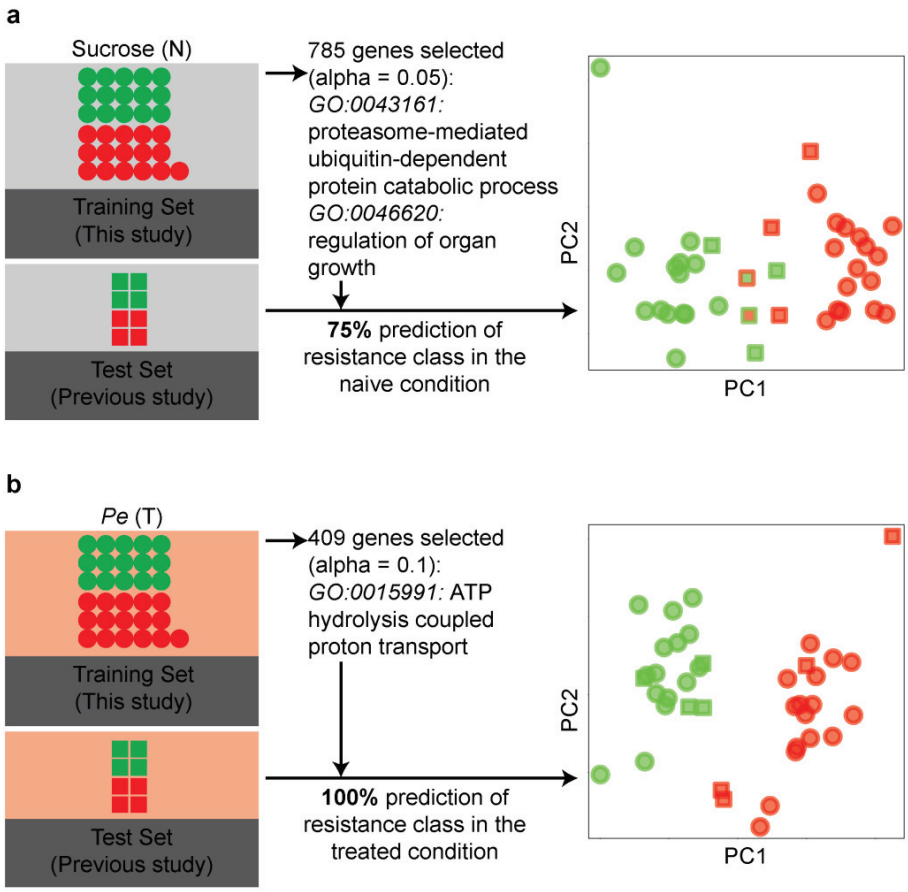


Figure 3:3 **Resistance class can be fully predicted based on specific gene signatures.**

(a) Resistance class of 6/8 samples can be predicted correctly based on the gene expression levels of 785 genes in the naïve state. **(b)** Prediction of resistance based on the gene expression level after infection is at 100%. For each scenario, a PCA that is weighted by the absolute value coefficients of the model is drawn to show relationships between the lines or samples.

3.2.4 *cis*-eQTL analysis links natural variation to gene expression levels.

After establishing that expression-level signatures could be predictive of resistance class, we sought to catalogue the effect of genetic variation on gene expression levels. For that, we used Matrix-eQTL (Shabalin 2012) to identify *cis*-Quantitative Trait Loci (QTLs) whose alleles correlate with the expression levels of nearby genes. We performed the analysis separately for the naïve and treated states and identified around 6700 *cis*-eQTLs for 1450 genes in each state (**Fig. 3:3a**). Interestingly, around 40% of the genes

with *cis*-eQTLs are shared between the two conditions, strongly supporting the notion that genomic variations indeed mediate transcriptomic differences.

It has already been shown that variant density in 39 of the DGRP lines is lower in genes than the overall background, and that this density increases to background levels upstream of, and then drops sharply at the TSS site(Massouras, Waszak et al. 2012). On the other hand, the p-values of the cis-eQTLs based on whole-adult are, on average, higher at the TSS, with the highest being immediately downstream. Indeed, a metaplot of the density of the gut eQTL distances from their genes' respective transcription start sites (TSS) shows that they are more likely to be present around the TSS, with a peak immediately downstream of the TSS (**Fig. 3:3b**). These observations are consistent with studies both on DGRPs and in other systems (Doss, Schadt et al. 2005, Stranger, Forrest et al. 2007, Massouras, Waszak et al. 2012).

We found at least one eQTL in around 25% of the genes expressed in the gut. In order to explore whether genes involved in specific biological processes are more affected by natural variation than others, we performed Gene Ontology enrichment analysis on three sets of genes: the naïve-only, treated-only, and the shared genes (**Fig. 3:3c**). The most enriched term in the shared genes is chitin metabolic process. We observe some degree of similarity in the terms enriched in the condition-specific genes, with a tendency of the naïve condition to have genes involved in the establishment of polarity. Furthermore, some infection-specific terms emerge, including response to endogenous stimulus and the regulation of the ERK1/ERK2 cascade. Taken together, our analyses catalogue possible genomic loci that could be affecting the expression levels of genes, some of which could explain susceptibility to infection. An assessment of the effect of those variations on possible transcription factor binding sites could help isolate causal variants and give a better molecular understanding of their implication in the normal and diseased physiology of the gut.

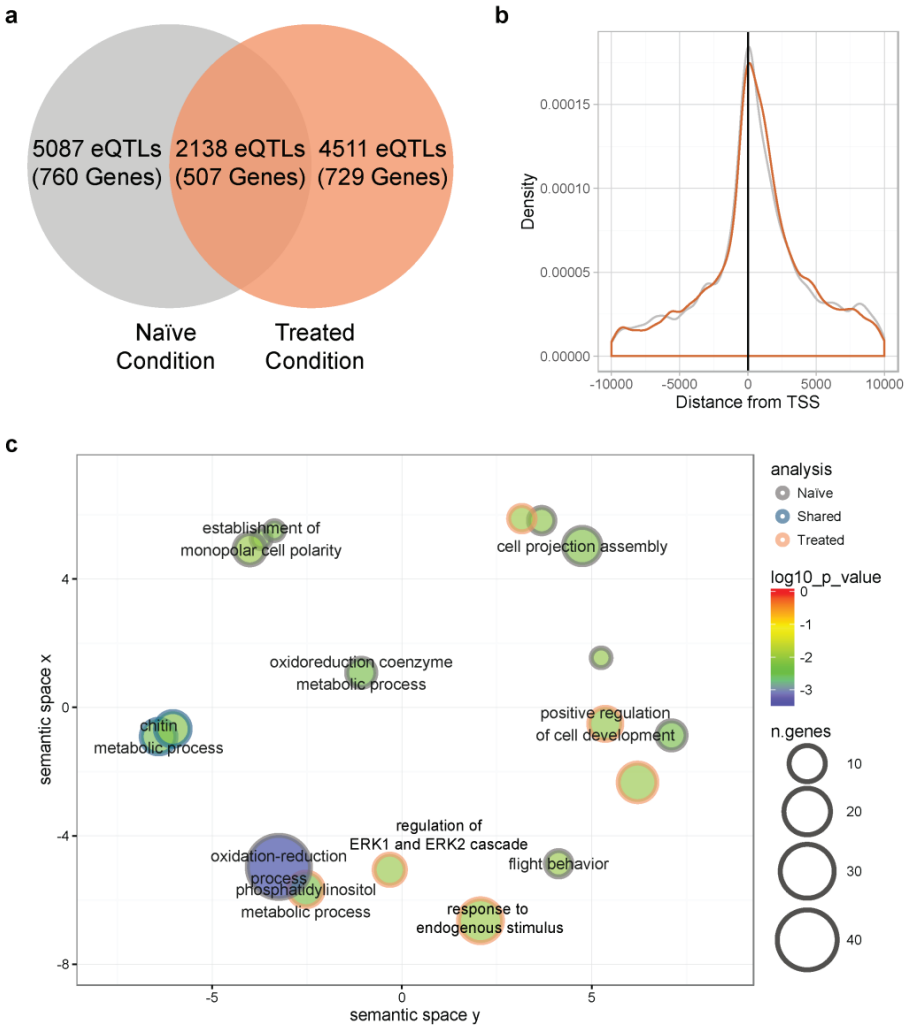


Figure 3:4 ***cis*-eQTL analysis links natural variation to gene expression levels.**

(a) Variants with a minor allele frequency greater than 5 in the 38 lines and that are within a 10kb window of each expressed gene were tested for their association with gene expression levels. Results of two *cis*-eQTL analyses (one for each infection condition) using Matrix eQTL (Shabalin 2012) are presented in a Venn diagram (FDR < 0.05). The number of genes with significant associations is indicated in parentheses. (b) Metaplot of locations of *cis*-eQTLs with respect to their associated genes' transcription start sites (TSS). Solid grey line and dashed orange line are for the naïve and treated states respectively. (c) Graphical representation of enriched biological process gene ontology terms based on the lists of genes with significant *cis*-eQTL associations. The GO analysis was performed using the GOstats(Falcon and Gentleman 2007) R package (Hypergeometric test p-value < 0.005), and REVIGO (Supek, Bo?njak et al. 2011) was used to reduce redundancy in the ontology groups and plot them by semantic similarity (allowed similarity = 0.7). The size of the circle indicates the number of genes belonging to a certain GO category, and the color indicates enrichment significance.

3.2.5 The gene *nutcracker* is induced in resistant lines, has cis-eQTLs, and is involved in the gut response.

We have previously seen that *nutcracker* (*ntc*) is the only differentially expressed gene between the resistant and susceptible lines (**Fig. 3:1b**). The gene is mainly induced after infection, but more so in some resistant lines (**Fig. 3:5a**). In some susceptible lines, its expression level even decreases. This observation prompted us to investigate its possible involvement in the gut response. For that, we obtained lines that have P-element-induced mutations, ntc^{f03797} and ntc^{f07259}, in or around the *ntc* locus, and tested their susceptibility to *P.e.* infection compared to a control line from the same genetic background, *w1118*. Both lines show increased susceptibility, with ntc^{f03797} having the more severe phenotype (**Fig. 3:5b**, log-rank test p-value < 0.05 when compared to w^{1118}). Furthermore, we performed RT-qPCR on dissected guts and saw that *ntc* induction is reduced in those lines compared to control. Interestingly, *diptericin* induction is almost completely abolished in these lines, suggesting that the Imd pathway activation upon *P.e.* infection is compromised in those mutants (**Supplementary Fig. 3:4a**). We believe that this is not the case in the DGRP lines, where the difference in *ntc* activation between susceptible and resistant lines are not as severe as those in the P-element mutants. Furthermore, we also infected a line that is heterozygous for a point mutation in the F-box domain of *ntc*, ntc^{ms771}, and also found that it is more susceptible than the control (**Supplementary Fig. 3:4b**, log-rank test p-value < 0.05 when compared to w^{1118}). Flies homozygous for this mutation are fragile and have a short lifespan.

Interestingly, we also identified 5 infection-specific eQTLs belonging to two linkage groups in *ntc*, two 7.6kb upstream and three 4.5kb downstream of its TSS (**Fig. 3:5c**). This raises the possibility that these variations affect *ntc cis*-regulatory elements that could partly explain differences between the resistance classes. For that, we predicted transcription factor binding sites (TFBS) in and around the *ntc* locus, and looked for overlaps with the eQTLs. Indeed, we find overlapping TFBS for the Broad Complex and Daughterless transcription factors in the upstream eQTLs, and a TFBS for Relish/NFkB in one of the downstream eQTLs. The alleles in both sites show good correlation with the *ntc* expression on 38 lines, but when associated with resistace of 140 DGRP lines, the allele at the Broad/Daughterless site had a lower p-value ($6.1*10^{-5}$ vs. 0.0215). It is worthy to note that this allele was previously tested in our genome-wide association study, yet it did not pass the $1*10^{-5}$ p-value threshold that we used for reporting the results. In addition, since *ntc* is not the closest gene to it, there was no way of linking it to *ntc*. This illustrates how eQTL analysis could help explain and prioritize GWAS hits that would otherwise be ignored. Taken together, our data suggest that *ntc* could be a previously uncharacterized player in the gut immune response, and that differences in its induction could be due to a single variation affecting an upstream *cis*-regulatory element that could partly explain differences between susceptible and resistant individuals in a population.

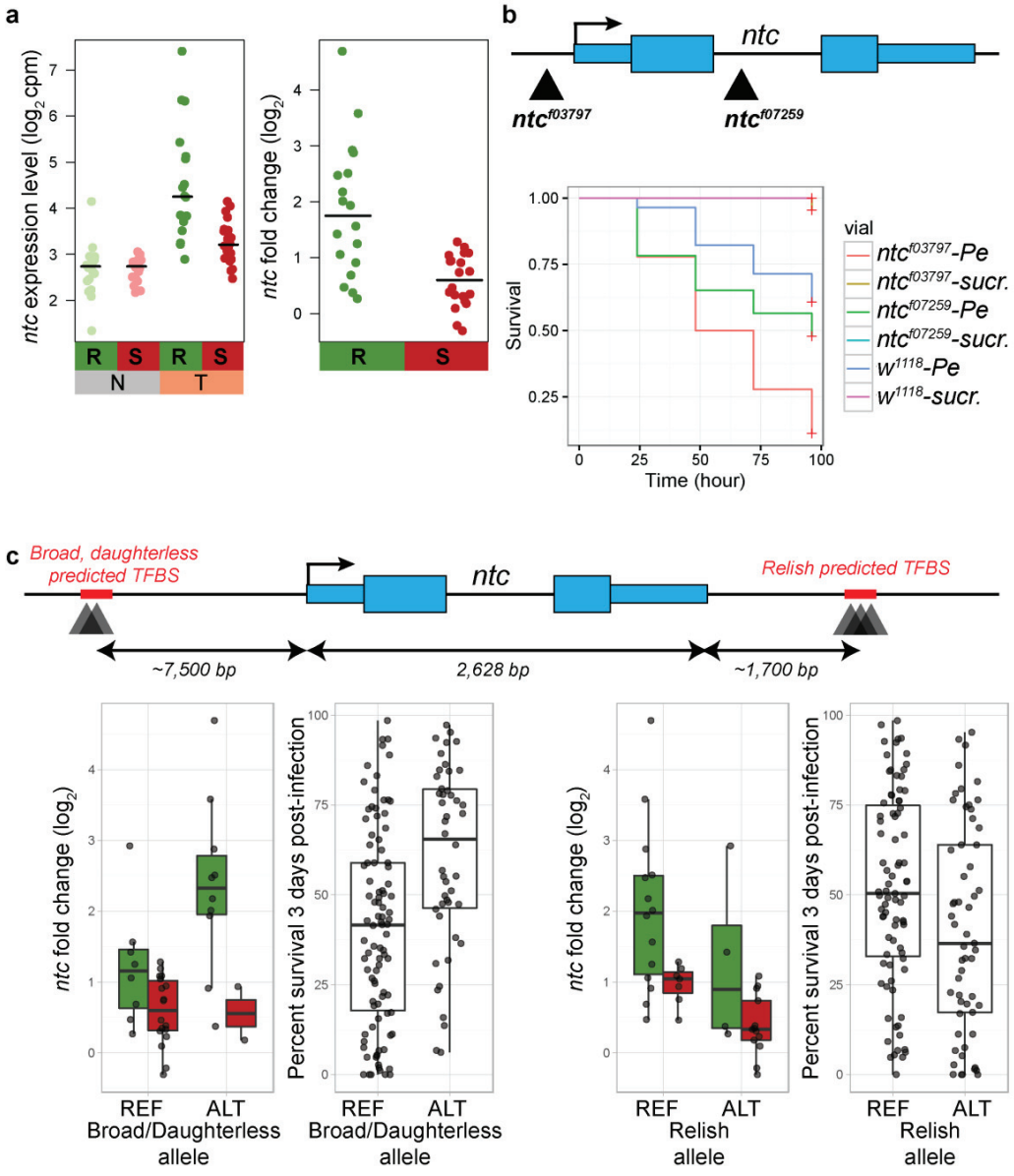


Figure 3:5 **The gene *nutcracker* is induced in resistant lines, has *cis*-eQTLs, and is involved in the gut response.**

(a) Left panel: Levels of expression (in $\log_2$(cpm)) of the ntc gene by resistane class and infection condition. **Right panel:** Fold change of ntc by resistance class. **(b)** Survival of two P-element insertion lines to *Pe* infection compared to a w^{1118} control. **(c)** *cis*-eQTLs around the ntc locus, and their overlap with predicted transctiption factor binding sites (TFBS). TFBS prediction was done using FIMO(Bailey, Johnson et al. 2015) and motifs from the Fly Factor Survey (Enuameh, Asriyan et al. 2013) and OnTheFly (Shazman, Lee et al. 2013) databases. The expression fold change by resistance class and two of those alleles (termed the broad/daughterless allele, and the relish allele) is plotted, as well as the percentage death of 140 DGRP lines(Bou Sleiman, Osman et al. 2015).

3.3 Discussion

Some of the most interesting findings in this study are the ones that we typically perceive as negative. It is surprising how DGRP lines with diametrically opposite resistance to infection all have a similar response after ingestion of a pathogenic bacterium (**Fig. 3:1**). We show that this is not due to our inability to detect genotype-specific differences, since lines of the same genotype cluster together at the transcriptional level (**Supplementary Fig. 3:1**). It is therefore clear that genomic variation imparts line-specific systemic differences on the transcriptome, yet only a subset of those differences is relevant in determining resistance. Furthermore, as are measuring a relatively early timepoint after infection, differences between resistance classes might increase afterwards. We used machine learning to identify sets of genes whose expression levels collectively define resistance class and validated them by predicting susceptibility of samples from a previous study with up to 100% success. The sizes of those sets are in the order of 400 or more genes, a number that is too large to be experimentally tractable. Nevertheless, considering that the guts were profiled only four hours post infection, and that the flies start dying well after that timepoint, this approach could be useful in predicting the prognosis of an ongoing infection. In a medical setting, such information, if obtained in a timely manner, could be very helpful in administering personalized treatments for patients.

To directly assess the effect of genomic variation on gene expression levels, we catalogued the possible *cis*-eQTLs around all expressed genes. eQTL analysis is a useful method to make sense of GWAS QTLs, prioritize candidates, and study GxE interactions (Gibson, Powell et al. 2015). Our study design is a very powerful one, since we have an organismal phenotype from inbred lines, gut-specific transcriptomes under two controlled conditions, and the whole-genome information. In both the treated and naïve state, around a third of all associations are unchanged, confirming that genotypic differences indeed drive gene expression differences. For instance, the eQTLs around the *ntc* locus are only associated with *ntc* levels in the treated state, which could be an example of cryptic variation contributing to infection resistance (Gibson and Dworkin 2004, Gibson, Powell et al. 2015). Variants in *ntc* are not only associated with its expression level, but with the resistance level of the whole fly. Moreover, these variants overlap with predicted TFBS, implying that a causal role of their effect could be assigned through further experimentation. Allele-specific expression of F1 hybrids carrying the two alleles could show whether the two copies of *ntc* are being induced differently. Enhancer-trap lines for the different regions spanning the eQTLs could help identifying the enhancer involved in the induction. Finally, the effect of polymorphism on transcription factor-DNA binding would also serve as an *in vitro* validation. If causality is established, it could constitute a rare example of an

eQTL that modifies an ecologically-relevant complex trait through its effect on binding of a transcription factor in a specific environmental condition.

Nutcracker was initially in a screen for mutants that fail to undergo sperm individualization due to inability to activate caspases (Bader, Arama et al. 2010). Through its F-box domain, ntc interacts with other partners to form an SCF (Skp, Cullin, F-box) ubiquitin ligase (E3) complex that controls caspase activity in *Drosophila* (Bader, Benjamin et al. 2011). Caspases play important roles in insect immunity and homeostasis through both apoptotic and non-apoptotic pathways. For instance, Dredd, the homolog of human Caspase-8 is required for Relish cleavage and activation (Leulier, Rodriguez et al. 2000). Futhermore, activation of the IKK complex is dependent on ubiquitination (Zhou, Silverman et al. 2005). In addition, studies in mammals have shown that commensal bacteria could affect ROS levels, leading to modification of the activity of the SCF complex, thus affecting NF-κB signaling (Kumar, Wu et al. 2007). Given all the possible mechanisms, the exact function of *ntc* in the gut and enteric infection is not clear and should be the subject of a more mechanistic study.

The gut is a highly regionalized organ (Buchon, Osman et al. 2013, Marianes and Spradling 2013) that consists of multiple cell types (Dutta, Dobson et al.). One limitation in our approach is that we profiled whole gut transcriptomes, without taking regional or cell-type differences into consideration. Future studies could address how different eQTLs mediate gene expression at a finer level, revealing conditional eQTLs whose function is restricted to a certain cell-type or environment.

3.4 Materials and Methods

3.4.1 Fly Stocks and infection experiments

For fly medium composition and oral Infection procedures, see methods in Chapter 2. The ntc^{f03797} and ntc^{f07259} stocks were obtained from the Bloomington Stock Center. The ntc^{ms771} stock was a kind gift from the Hermann Steller lab.

3.4.2 RNAseq

RNA extraction: RNA extraction was performed using Trizol Reagent (Invitrogen) using the standard protocol.

Library preparation and sequencing: Standard Illumina Truseq libraries were prepared from 1ng total RNA as measured by a Nanodrop 1000 device (Thermo Scientific) by the Lausanne Genomic Technologies Facility. Single end sequencing was performed for 100 cycles. Initially, 80 samples from 40 lines were sequenced but we excluded 4 samples from two lines. One of the lines was contaminated, as its reads came from two genotypes and another DGRP line had a smaller library size in one condition, with led to its elimination from the analysis.

Mapping to invidualized genome: For each DGRP line, we generated an individualized fasta genome sequence based homozygous variants in the published Freeze 2 DGRP genotypes and the Release 5 reference genome. We also generated individualized gene annotations by applying the offsetGTF tool included in the mmseq package (Turro, Su et al. 2011) on the Ensembl BDGP5.25 . For each sample, reads were mapped to the respective genome usint STAR aligner. Reads for each gene were counted using HTseq-count.

Normalization and differential expression: We used the edgeR package to perform TMM normalization, followed by conversion to Counts Per Million Voom with quantile normalization. When we combined samples from this study and the previous study, we used the same approach, starting from combined gene counts, with the addiction of the removeBatchEffect function in the limma package. Differential expression was performed in limma using the weights obtained by voom while adjusting for intra-line correlations using the duplicatecorrelation function with the DGRP lines as the blocking factor. The following model was used: y = treatment + class + treatment:class. For each predictor variable, genes having a fold change of 2 and a Benjamini-Hochberg corrected adjusted p-value of 0.05 were deemed differentially expressed.

Principal component analyses: The FactoMineR package was used to perform the principal component analyses with scaling and centering.

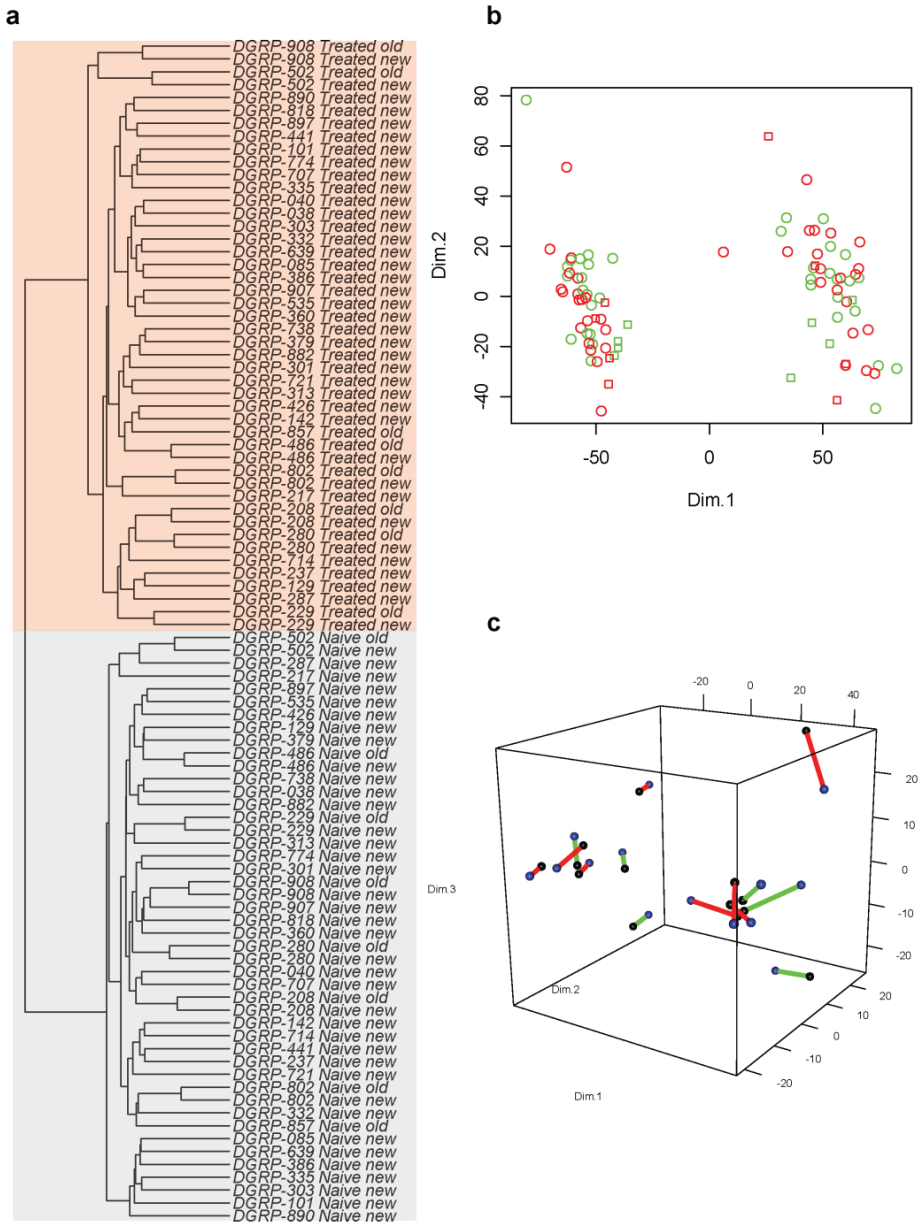
3.4.3 Machine learning and prediction

Gene expression CPM values of the combined experiments were used throughout the analyses. For the condition-specific predictions, the filtering and normalization of genes was performed separately. For the Fold Change and treatment effect analyses, the whole dataset was used. We used the GLMnet package for feature selection and prediction. For the selection of the best value of alpha, we tested all values of alpha from 0.05 to 1 with 0.05 increments. For each value of alpha, we randomly split the samples from the new study (that are not replicated) into a cross-validation set for lambda estimation and a validation set for prediction accuracy estimation. We then peformed cross-validation, selected the lambda that minimizes misclassification error, and used the corresponding model to predict the resistance class of the validation set. To circumvent any problems that could arise due to the random sampling, we repeated the process 100 times. We defined the best alpha value, which ultimately determines the number of selected features, as the one that yields the maximum mean prediction success in the validation set. We then used cross-validation again on the whole set of lines from the new study to re-estimate the minimum lambda value. Then we selected the features that will constitute the final prediction set. Finally, we used the resulting model to predict the resistance classes of the samples from the previous study. We performed PCA on the gene expression levels of the selected features, weighed by their coefficients to represent the relationships between the samples. The GO analysis was performed using the GOstats (Falcon and Gentleman 2007) R package (Hypergeometric test p-value < 0.005).

3.4.4 *cis*-eQTL analysis

We performed separate analyses for each infection condition using Matrix-eQTL. Variants that are within 10kb of an expressed gene and whose minor allele frequency is greater than 5 in the 38 tested lines were used. Cis-eQTL associations with an FDR corrected p-value that is less than 0.05 were considered significant. Metaplots were plotted in R. The GO analysis was performed using the GOstats (Falcon and Gentleman 2007) R package (Hypergeometric test p-value < 0.005), and REVIGO (Supek, Bo?njak et al. 2011) was used to reduce redundancy in the ontology groups and plot them by semantic similarity (allowed similarity = 0.7)

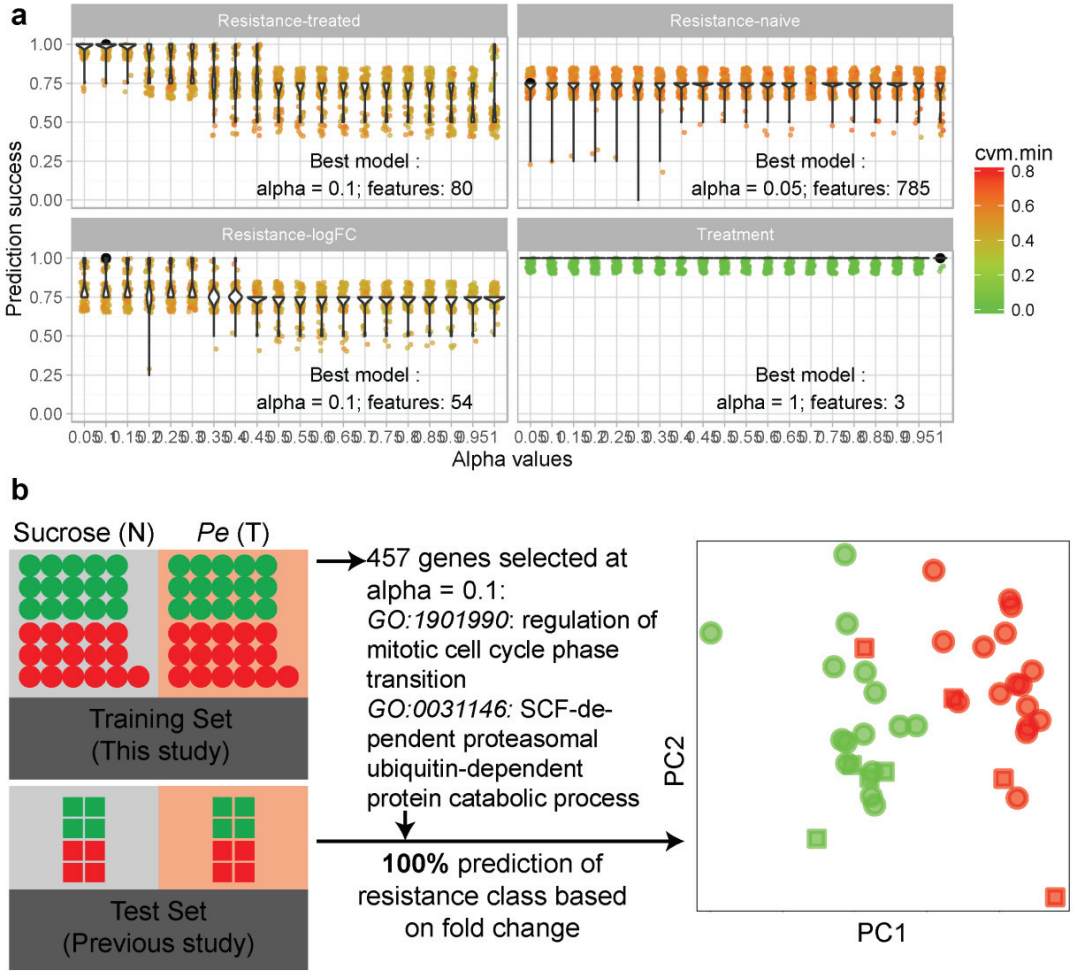
3.5 Supplementary Materials



Supplementary Figure 3:1 **Reproducibility of line-specific transcriptomes**

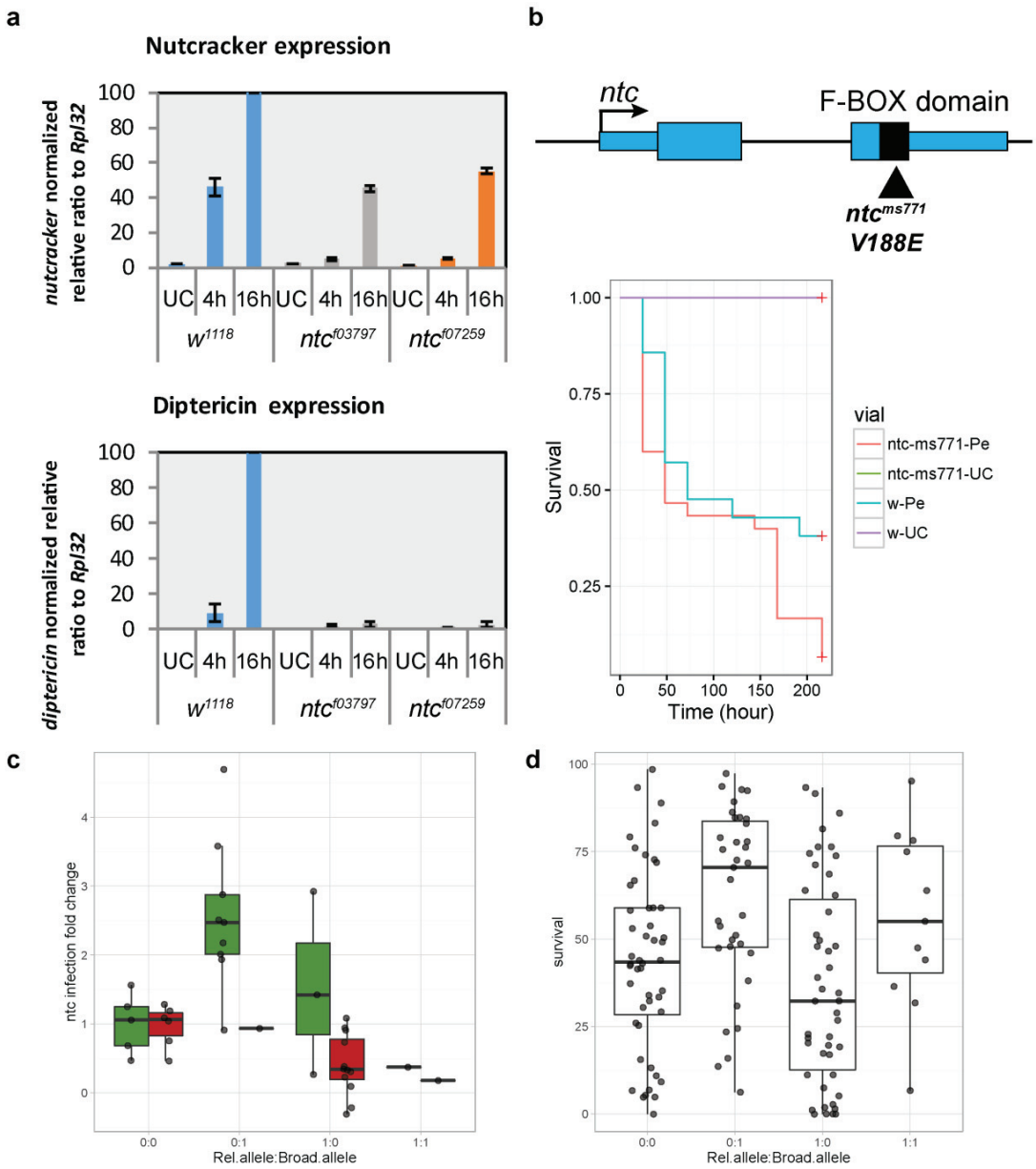
(a) Hierarchical clustering of the combined samples from this study and that of Bou Sleiman and Osman, 2015. Hclust was used on the Euclidean distance matrix in R. **(b)** Principal component analysis based on the gene expression profiles of the combined samples. Samples from the new and old study are represented as circles and squares, respectively. **(c)** Three dimensional representation of the first three principal components based only on the samples that belong to lines replicated between the two studies. Corresponding samples

are connected by a segment that is colored based on susceptibility group. The sphere color indicates the batch (blue is new, black is old).



Supplementary Figure 3:2 **Feature selection and prediction of resistance class**

(a) Results of 100 rounds of cross-validation with different levels of alpha and prediction on a validation set from the learning set. For each value of alpha, a subset of the training set is taken as validation set, then a cross-validation is performed to select the best value of lambda. Then the prediction success of each model is assess using the validation set. The best model is the one that has the highest mean prediction susccess. Cvm.min is the minimum mean cross validated error obtained after cross-validation, based on which lambda is selected. **(b)** Prediction of resistance class based on gene fold changes after infection.



Supplementary Figure 3:3 **The gene *nutcracker* is induced in resistant lines, has cis-eQTLs, and is in-volved in the gut response.**

(a) Top: Expression levels of the *nutcracker* gene in adult female guts as measured by RT-qPCR in the unchallenged state (UC) and after 4hrs and 16hrs of *P.e.* oral infection. **Bottom:** Expression levels of the diptericin gene in the same samples. Error bars represent standard deviations. In both panels, the levels relative ratios to *Rpl32*, normalized to 100%. **(b) Top:** Gene diagram of ntc showing the F-box domain in black, and the location of the substitution in the ntc^{ms771} mutants. **Bottom:** Survival curve of the ntc^{ms771} mutant compared to a w^{1118} wild-type after infection with *P.e.* **(c)** Plot of ntc fold changes by Rel and Broad alleles. 0 indicates reference, 1 indicates variant. **(d)** Similar plot to **(c)**, but this time plotting the survival percentage after three days of *P.e.* infection among 140 DGRP lines.

Chapter 4 The alternative splicing landscape of the *Drosophila* gut upon enteric infection

This chapter explores the alternative splicing landscape of the Drosophila gut when it is exposed to an oral pathogen. This often-neglected aspect of gene expression not only generates protein diversity, but also fine-tunes how the transcriptome is translated. A collaboration with the laboratory of Prof. Roderic Guigo through Tommaso Andreani, who was doing an internship there, has sparked the beginning of this project.

Abstract

RNA Splicing is a key mechanism that not only generates protein diversity, but contributes to the fine tuning of the transcriptome. This ability to diversify and control the transcriptional output of the genome may facilitate how the organism adapts to a changing environment. We employ a systems approach in the study of isoform ratios in the infected and uninfected guts of females from 38 inbred lines of Drosophila melanogaster. We find that infection leads to extensive and consistent differences in isoform ratios, which result in a more diverse transcriptome, that is skewed toward longer transcripts, due to longer 5'UTRs. Additionally, we establish a role for genetic variation in mediating inter-individual differences, with splicing Quantitative Trait Loci being more numerous in the infected state and preferentially located in the 5' end of transcripts and directly upstream of the splice donor sites. Moreover, we find a general increase in intron retention events concentrated in 5' ends of transcripts. The length, CG content, and RNA Polymerase II occupancy of the retained introns suggest that they have exon-like characteristics and are possibly being translated. Finally, we show that the sequences of retained introns are enriched with the Lark/RBM4 RNA-binding motif, and establish a critical role of Lark in mediating the gut defense response. For the first time, we describe a link between splicing and the gut's response to enteric infection, which could have general implications on gene regulation and protein translation.

Author Contributions and Acknowledgements

Maroun Bou Sleiman, Tommi Andreani, Bart Deplancke, and Roderic Guigo conceived the project. Maroun Bou Sleiman and Dani Osman prepared the samples. Maroun Bou Sleiman and Tommi Andreani performed the analyses. Maroun Bou Sleiman performed experiments and wrote the chapter. Michael Frochaux performed the ChIP-sequencing experiments and mapping.

4.1 Introduction

The eukaryotic genome is expressed and regulated by diverse mechanisms that ensure robustness and flexibility to adapt to different conditions. RNA splicing is one major mechanism that contributes in achieving this complex task. One of its obvious functions is the increase in the repertoire of protein-coding genes through the production of multiple isoforms(Leoni, Le Pera et al. 2011). Moreover, alternative splicing often generates transcript isoforms that have the same coding potential but diverse untranslated regions, which could have implications on stability and translation efficiency of the transcripts(Hughes 2006). There has been a surge in available RNA sequencing data in the last years, yet the study of alternative splicing has often been ignored. Estimating alternative transcript abundances is still challenging since most methods rely on annotations, many assumptions, and short sequencing reads (Ozsolak and Milos 2011). Perhaps more importantly, drawing biological conclusions from transcript-level data is conceptually more challenging since it adds another layer of complexity.

There is a growing body of evidence that splicing is not an isolated process and that it interacts with transcription and RNA export (Reed 2003). Specifically, it has been shown that the RNA polymerase II recruits splicing factors through its cytoplasmic tail domain (CTD) to promote splicing (David, Boyne et al. 2011). Moreover, it seems that there is a reciprocal link between co-transcriptional splicing and RNA-pol II kinetics (Nojima, Gomes et al. , Kornblihtt, De La Mata et al. 2004). Alternative splicing has been shown to be affected by external stressors, notably heat shock (Biamonti and Caceres , Lin, Hsu et al. 2007, Dutertre, Sanchez et al. 2011, Shalgi, Hurt et al. 2014). The first report of splicing being affected by heat shock was in the fly, where pre-mRNAs of *Hsp83* and *Adh* accumulated at severe temperatures (Yost and Lindquist 1986). On a genome-wide scale, posttranscriptional splicing seems to be inhibited after heat shock leading to widespread intron retention (Shalgi, Hurt et al. 2014). To our knowledge, however, no study has assessed the effect of bacterial infection on splicing.

Here, we systematically assess splicing in the context of enteric infection in *Drosophila melanogaster*. While the *Drosophila* gut transcriptome under different physiological conditions has been studied by us and others, to our knowledge, the extent and role of splicing has never been addressed. In addition to classical laboratory strains, we make use of a large RNA sequencing study of 38 lines from the Drosophila Genetic Reference Panel (DGRP) inbred lines to study this phenomenon under different environmental (infection) and genomic perturbations. This is the first study of its kind in the *Drosophila* gut, which has lately attracted a lot of attention in the scientific community, especially as a system to understand enteric infection (Lemaitre and Miguel-Aliaga 2013).

4.2 Results

4.2.1 Enteric infection with different pathogens leads to widespread changes in intron retention

Initially, we sought to characterize potential infection-induced differences in splicing patterns at the single intron level. For that, we used high quality paired-end RNA-sequencing data of adult female guts of the widely used w^{1118} strain. Adult female flies were either fed sucrose (1.5X), *Pseudomonas entomophila* (OD_{600} = 100 and 1.5X sucrose), or *Erwinia carotovora carotovora 15* (OD_{600} = 100 and 1.5X sucrose), then their whole guts were dissected, followed by RNA extraction and sequencing. We then mapped the resulting reads to the reference genome. Using an available annotation that is specific to intron retention events (McManus, Coolon et al. 2014), we estimated the percent spliced in (PSI or Ψ) value for each of the 32895 introns, which is the number of retention reads (spanning the exon-intron boundary as well as the reads in the intron) divided by the sum of the number of retention and splicing reads (spanning the exon-exon boundary as well as in the flanking exons) using MISO (Katz, Wang et al. 2010) (**Fig. 4:1a**). When we compared the two infection conditions to the uninfected state, we found that both conditions lead to differences in intron retention events (**Fig. 4:1b-c**, bayes factor > 10, delta psi > 0.2). *Ecc15* infection, which is less pathogenic than *Pe*, leads to fewer differences overall, with around 40% being shared with the *Pe* condition (**Supplementary Fig. 4:1a-b**). Interestingly, the number of intron retention events with a positive delta PSI value in both infection conditions was around double that of those with a negative value, indicating an overall increase in retention post infection. This significant impact of infection on splicing prompted us to investigate the phenomenon and its possible consequences in more detail, both at the intron and the transcript level. Moreover, since it was not clear whether this effect is unique to the tested laboratory strain,

we decided to perform all our analyses on RNA-sequencing data from 38 lines from the Drosophila genetic reference panel (DGRP), a set of inbred lines derived from a natural population.

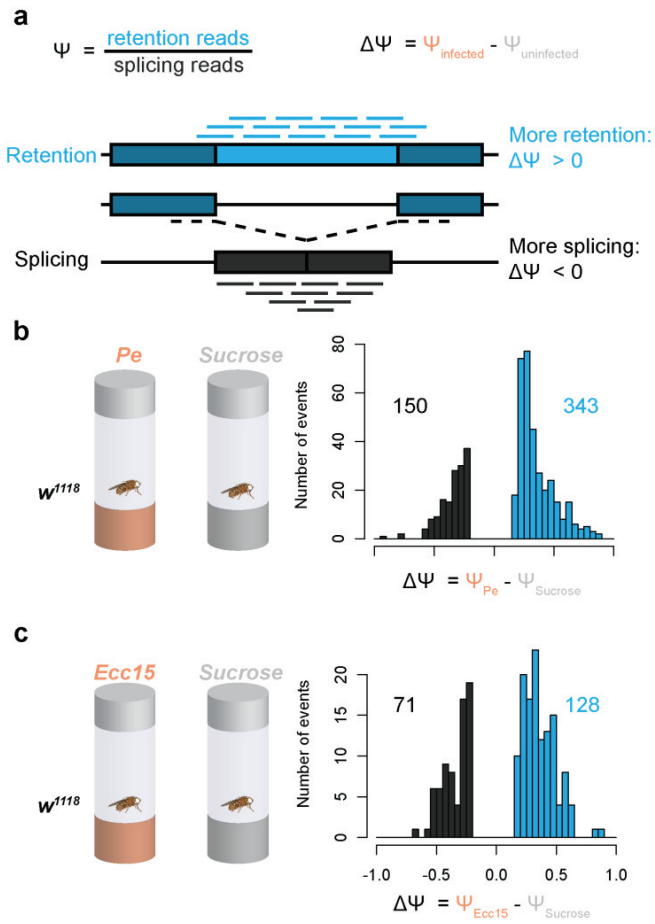


Figure 4:1 **Enteric infection with different pathogens leads to widespread changes in intron retention**

(a) Diagram depicting how intron retention changes are computed. For each sample, delta PSI values for different splicing events(McManus, Coolon et al. 2014) were calculated by subtracting the PSI value of the uninfected sample from that of the infected one. **(b-c)** Histogram of delta PSI values of intron retention events whose PSI values are significantly different (Bayes factor > 10, delta PSI > 0.2) from the unchallenged (sucrose fed) state four hours after infection with **(b)** *Pe* and **(c)** *Ecc 15*

4.2.2 Enteric infection leads to extensive changes in transcript isoform ratios

We have previously measured the resistance of 140 DGRP lines to enteric infection with *Pseudomonas entomophila* (*P.e.*) (Bou Sleiman, Osman et al. 2015). In this study, we selected 38 DGRP lines, 20 of

which are susceptible and 18 resistant to P.e. enteric infection (**Fig. 4:2a**), infected adult female flies, and performed RNA-sequencing on their dissected guts 4 hours post infection. For each line, we also sequenced guts of unchallenged, sucrose-fed flies. In total, we sequenced the poly-A enriched transcriptome of 76 samples. Since the lines have been shown to be highly polymorphic, we opted for analyses on individualized genomes. For that, we used the available genotyping data (Huang, Massouras et al. 2014), including single nucleotide as well as indels and structural variations, to generate individualized genomes and gene annotations (see **Methods**) which we used throughout the analyses.

To gain insight into the changes in the isoform composition of each gene after infection, we used a multivatiate distance-based approach decribed in Gonzalez-Porta et al. (2012) (Gonzàlez-Porta, Calvo et al. 2012). Briefly, we estimated the isoform composition of each gene using MISO(Katz, Wang et al. 2010) software. Then we used a non-parametric test as described in Anderson (2012)(Anderson 2001) with the Hellinger distance as a dissimilarity measure to identify genes that have condition-specific isoform ratios. Of the 1877 genes that passed filtering (see **Methods**), 40% were significantly changed after infection (**Fig. 4:2b**, p-value of homogeneity > 0.05, BH-corrected p-value < 0.05, effect size > 0.2). Interestingly, only 25% of the significant genes based on splicing ratios are known to be differentially expressed after infection (**see Chapter 2 and 3)**, suggesting that gene-level differential expression could overlook important aspects of the gut transcriptional response to enteric infection. We were not able to find significantly different ratios between the resistance classes, yet some genes showed weak trends of such differences (**Supplementary Fig. 4:2**). A gene ontology analysis shows that genes associated with RNA-metabolism, organelle organization and biogenesis, and epithelial tissue development are enriched within this set (**Fig. 4:2c**). Interestingly, the set of genes we obtained is not enriched with immunity gene ontology terms. This could possibly be due to different regulatory restraints imposed on genes involved in the immediate immune response (i.e. in the resistance mechanisms (Schneider and Ayres 2008)), many of which are typically switched on and massively produced after infection, versus genes involved in homeostasis (i.e. the tolerance mechanisms (Schneider and Ayres 2008)), which are required to function in both conditions, albeit with different dynamics.

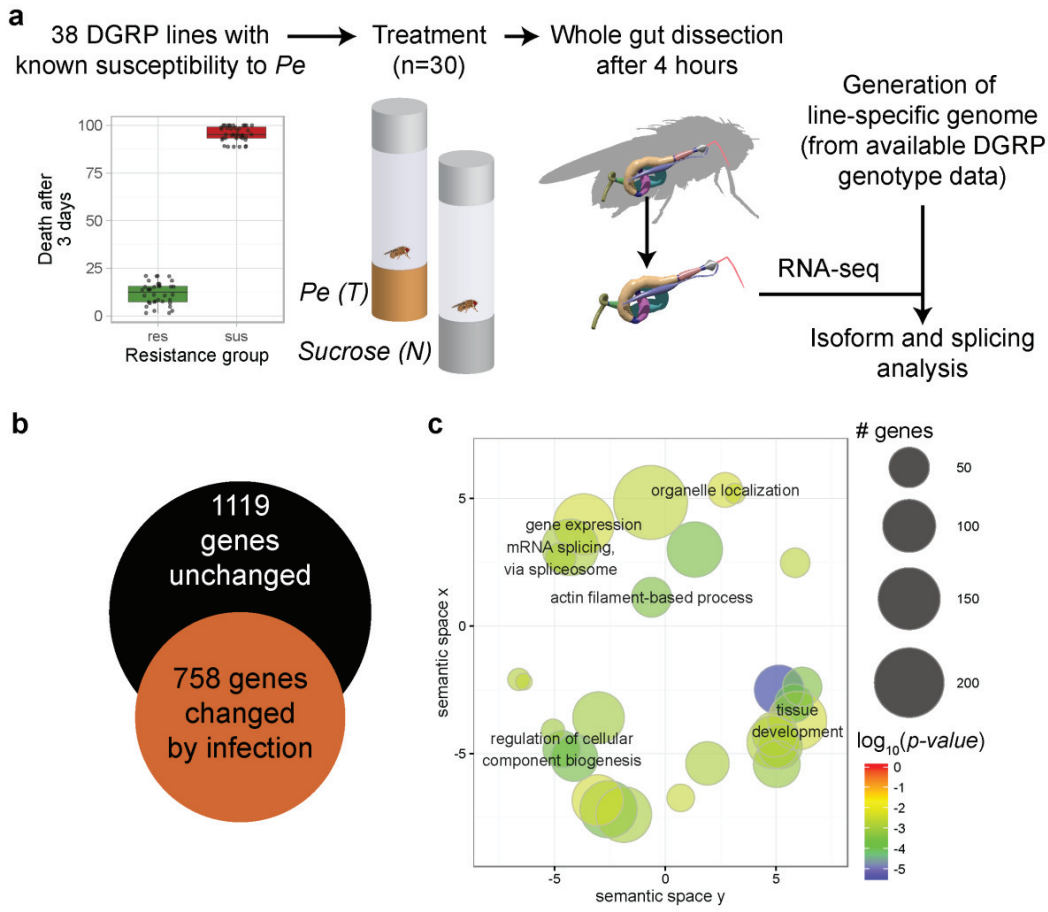


Figure 4:2 **Enteric infection leads to extensive changes in transcript isoform ratios**

(a) The general experimental design of the study. Adult female flies from 38 DGRP lines already known to belong to two susceptibility classes to *P.e.*(Bou Sleiman, Osman et al. 2015) were either fed *P.e.* or sucrose. After 4 hours, total RNA of whole guts was extracted and sequenced. The resulting data was mapped to individualized genomes that have been generated using the known DGRP freeze 2 genotypes(Huang, Massouras et al. 2014). The resulting alignments were used for further analyses. **(b)** Venn diagram of the number of genes whose isoform ratios are significantly altered after infection. MISO(Katz, Wang et al. 2010) was used to calculate the ratios of different annotated isoforms and afterwards, the rasp package (Gonzàlez-Porta, Calvo et al. 2012) was used to determine significance (p-homogeneity > 0.05, BH adjusted p-value < 0.05, effect size > 0.1). **(c)** Graphical representation of enriched biological process gene ontology terms based on the list of genes whose isoform ratios are differentially expressed after infection. The GO analysis was performed using the GOstats (Falcon and Gentleman 2007) R package (Hypergeometric test p-value < 0.005), and REVIGO(Supek, Bo?njak et al. 2011) was used to reduce redundancy in the ontology groups and plot them by semantic similarity (allowed similarity = 0.7). The size of the circle indicates the number of genes belonging to a certain GO category, and the color indicates enrichment significance.

4.2.3 The transcriptional response is characterized by higher isoform diversity

We next examined the effect of infection on the diversity of the transcriptome by calculating the gene-based Shannon entropy for each sample. We found that infection leads to a general increase in diversity in the infected state (**Fig. 4:3a-c, Supplementary Fig. 4:3a-b**). The density plot of Shannon entropies shows that after infection, there is an increase in number of genes with a higher diversity, and consequently fewer genes with low diversity (**Fig. 4:2a**). The average diversity per sample also shows a trend towards higher diversity in the treated state (**Fig. 4:2b**). Interestingly, a breakdown by isoform number reveals that for genes with 2,3, or 4 isoforms, resistant lines exhibit a higher level of diversity than susceptible lines (**Fig. 4:2c, Supplementary Fig. 4:3b**). These observations suggest that upon infection, the transcript output of many genes is less dominated by a single or few isoforms. The functional relevance of this increase in diversity is not clear, and requires further study. It is important to note that transcript diversity does not necessarily lead to protein diversity, since different isoforms could only differ in their UTRs but not necessarily the coding sequence.

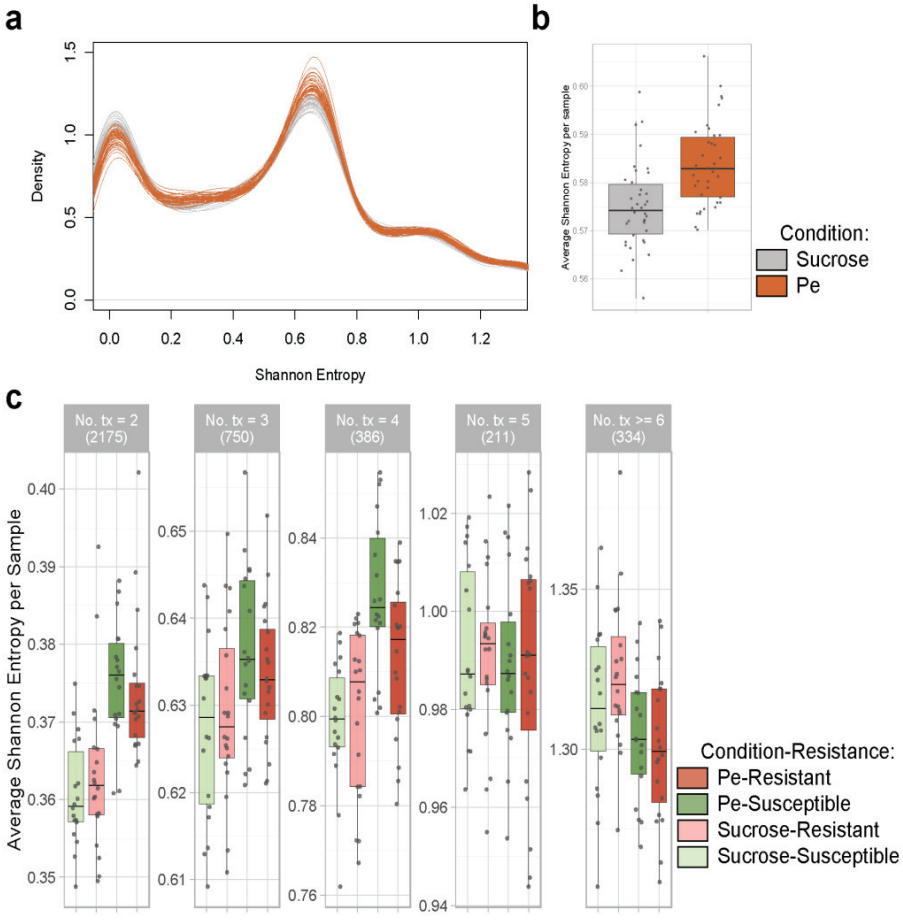


Figure 4:3 **The gut transcriptional response to infection is characterized by higher isoform diversity**

(a) The distribution of Shannon entropies of transcript ratios of each gene per sample. Uninfected and infected samples are in grey and brown, respectively. **(b)** Boxplot of the average Shannon entropy per sample treatment condition. **(c)** Breakdown of average Shannon entropy by isoform number, susceptibility class, and treatment condition.

4.2.4 Post-infection transcripts tend to be longer, mainly due to the production of longer 5' UTR

We next sought to characterize the effect of the splicing differences on the length of the produced transcripts. In order to do that, we estimated an effective length measure for each gene. Briefly, for each gene in each sample, we estimated the effective transcript length weighted mean of its individual transcripts (taking into account the effect of insertions and deletions) by their expression ratios. Similarly, we extended this method to individual features within the transcript, namely the 5'UTR, 3' UTR, and the coding sequence. Then we compared the effective lengths before and after infection to obtain the number

of genes who have an increased, decreased, and an unchanged effective length. Additionally, we generated a null distribution of effective length differences by performing 100 permutations of the data, by randomly assigning infection status to the samples, and compared this to our observed set using G-tests. Interestingly, while the effect of natural variation, namely insertions and deletions, on the coefficient of variation in feature length was most prominent in 3' UTRs, the effect of infection on the effective length of genes was strongest in 5' UTRs **(Supplementary Fig. 4:4a)**. Furthermore, 3' UTR lengths differed the most from the null expectation, but the proportion of those that increase in effective length is close to those that decrease (23.2% vs. 24.1 respectively, **Fig. 4:4a**). On the other hand, we found that there are around 7% more genes that increase in transcript and 5' UTR effective length than those that decrease. Predicted polypeptide length, however, did not show differences from the null distribution nor any skew. The distribution of this shift in effective length is consistent across the DGRP lines, with transcripts and 5'UTRs having an excess of increased effective lengths (**Supplementary Fig. 4:4b-c**). To show which feature contributes to the effective length change the most, we performed a similar analysis, this time calculating the transcript length effective change differences after the removal of a certain feature. Indeed, the removal of 5'UTR length and not the predicted polypeptide or 3' UTR abolished this skew in the proportions (**Fig. 4:4b**). Together, these results suggest that infection-induced differences in transcript ratios affect 5' UTRs the most and favor the production of the isoforms with longer 5'UTRs.

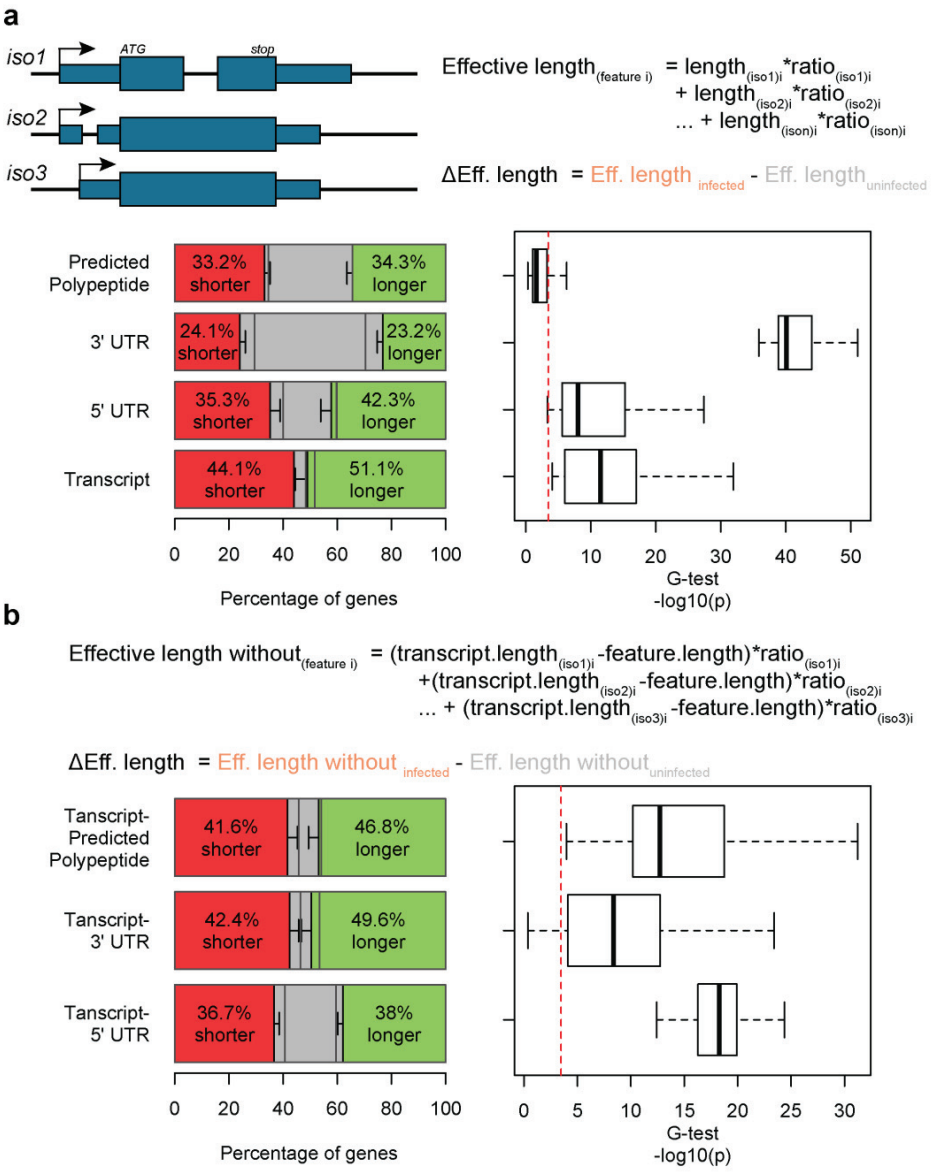


Figure 4:4 **Post-infection transcripts tend to be longer, mainly due to the production of longer 5' UTR.**

(a) The line specific effective length of each gene's transcript, CDS, 5' UTR, and 3' UTR lengths was obtained by calculating the weighted mean of each feature by its isoform ratio. The difference in length between the *P.e.* infected state and the uninfected state was then calculated for each line. The figure shows whether the feature increased, decreased, or didn't change in average length after infection. Error bars are the standard deviation. A null distribution was generated by performing 100 permutations by randomly shuffling the samples. The grey bars indicate the average obtained by permutations. Repeated G-tests were used to compare the feature length change in each line to the null distribution. The boxplots show the $-\log_{10}$(p-values) of the tests, with the dotted red line representing a Bonferroni-

corrected *p-value* threshold **(b)** Similar to previous panel, but this time the effective length of each transcript without either the predicted polypeptide, 3'UTR, or 5'UTR was calculated.

4.2.5 The effect of natural variation on splicing is increased after infection.

We have thus far established that transcript ratios of a large set of genes is significantly affected by infection status, that diversity is increased, and that the effective length of multi-isoform genes is increased. We next sought to establish a link between genetic variation and these transcript levels. To achieve this, we identified splicing quantitative trait loci (sQTLs) in the two infection states. Specifically, for each gene, we looked variations within a 10kb window, that correlate with the shift in its isoform ratios. For that, we used SQTLseekeR (Monlong, Calvo et al. 2014), which employs a similar statistical methodology as the one we used to detect significant differences in transcript ratios (see **Methods**). We identified 499 and 839 naïve- and treated-specific sQTLs, and 395 sQTLs that are common to both conditions (**Fig. 4:5a**). Interestingly, there were around 50% more sQTLs in the treated state. Additionally, the number of genes affected by sQTLs in the treated condition is almost double that of the naïve condition (108 *vs.* 65 genes). However, there is a similar number of genes with significantly different post-infection splicing ratios (as in **Fig. 4:5b**) that are in the naïve (13), treated (16), and shared group (20), indicating that infection response genes are not more likely to be affected by sQTLs upon infection. Together, this suggests that the effect of natural variation on splicing is more pronounced after infection, and that line-specific differences can be more readily detected in the infected state.

To obtain insights into which biological processes are affected by variation in splicing ratios, we performed separate gene ontology enrichment of the three sets of genes. **Figure 4:5b** shows a single graphical representation of the three GO enrichment results. In the naïve state, GO terms related to transcription and splicing as well as development and nitrogen compound metabolic processes are enriched. In the treated state, other categories emerge, namely the detection of stimulus, cell adhesion, and carbohydrate metabolic processes. Both conditions share categories related to cellular homeostasis (specifically ion homeostasis) and energy derivation by oxidation of organic compounds.

Next, we examined the locations of the sQTLs in relation to the gene they are associated with. We used two approaches to obtain metaplots: a gene-centric and intron-centric approach. Since natural variation density

along genes is not uniform, and tends to be higher towards the 5' ends, we generated sets of randomly selected variants with a matching allele frequency spectrum to the sQTLs 10kb around genes. Indeed, both the random samples and the observed sQTL distribution show a peak around the TSS of genes (**Fig. 4:5c**). However, while the random sample distribution shows a single symmetrical peak with wide tails, the sQTL density shows a higher density at the main 5' peak, as well as an elevated plateau along the metagene body. These results suggest that it is more likely to find sQTLs at 5' ends of genes, as well as within the gene bodies. The density distribution could be intuitively explained as being a mixture of two sQTL classes. The first class could be mediating differences in alternative TSS selection as well as splicing, hence the 5' peak. The second class could be acting through co- or post-transcriptional splicing choices in the nascent transcript, where variations within transcript sequence are likely to affect splicing.

To gain insights into how a causal sQTL could be mediating differences in splicing, we calculated the density distribution around the closest intron to each sQTL as well as a suitable null-distribution. Interestingly, we observe a pattern that is very distinct from the random sample. While the random random sample shows a wide peak that is centered around the 5' end of the intron, the sQTLs exhibit a sharp peak at the 5' end, with the highest peak immediately upstream of the intron (**Fig. 4:5d**). There are more sQTLs upstream than downstream, and the number of sQTLs drops sharply right after the intron. This data suggests that natural variation affecting splicing could be doing so by causing differences in the signals required for splicing, predominantly around the 5' splice site. One such example of sQTL is in the gene *fbl6*, which has multiple sQTLs, one of which is exactly at the 5' splice site (**Fig 4:5e-g**). The lines with different alleles at that locus show markedly different splicing patterns, with a clear shift in the major isoform produced in both conditions. However, not all sQTLs could be assigned such a direct mechanism of action as this example, and some might have subtler effects by affecting exonic and intronic splicing enhancers (ESEs and ISEs). To assess this possibility, we asked whether it is more likely that an sQTL overlaps with an ESE or ISE than other random variations. Since these splicing enhancer sequences are short hexamers and numerous in *Drosophila* (Brooks, Aspden et al. 2011), predicting them along the genome leads to many false positives. Nevertheless, we took a set of 330 published enhancers (Brooks, Aspden et al. 2011), and looked for matches along all the gene bodies. Then we counted the overlaps between the sQTLs and a 100 random sets of variants with a matching allele frequency spectrum. Interestingly, 70% of the sQTLs overlapped a predicted enhancer, which is 10% higher than the maximum predicted through permutations (**Supplementary Fig. 5**). This enrichment suggests that it is possible that some sQTLs that lie within ESEs and ISEs could be affecting their enhancer functions. Taken together, our sQTL data shows that we can detect effects of natural variation on

splicing, even more in the infected state, and that these effects could be due to direct changes in splice sites, as well as other mechanisms predominantly at or around the splice donor site.

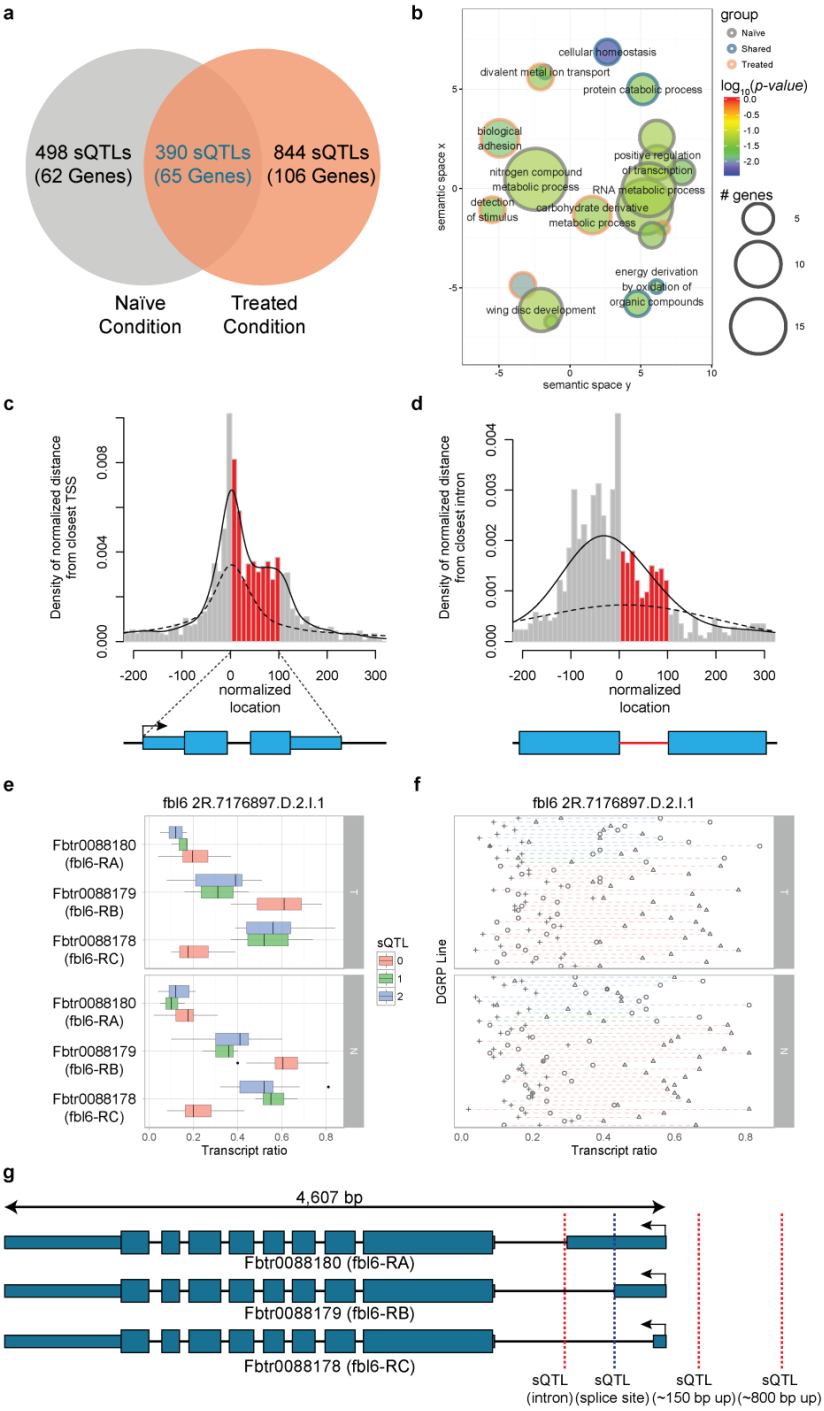


Figure 4:5 **The effect of natural variation on splicing is increased after infection.**

(**a**) Venn diagram showing the result of the *cis*-sQTL analysis (and number of associated genes) using sQTLseekeR(Monlong, Calvo et al. 2014) (BH adjusted p-value < 0.01, maximum difference in ratio > 0.1). **(b)** GO enrichment of the genes in the *cis*-sQTL results. The analysis is similar to that in figure 1, but the three groups in **(a)** were tested separately, then the GO categories were pooled in REVIGO. The shape of the point indicates the gene subset that is enriched with a specific term. **(c)** Metaplot of the pooled *cis*-sQTL results with respect to normalized gene length, and **(d)** intron length. Solid lines represent the density of *cis*-sQTLs, while dashed lines represent a random sample of 500,000 variants that are within 10kb of a gene. **(e)** The isoform ratios of a gene (*fbl6*) that has a *cis*-sQTL on one of its splice sites. The expression levels are grouped by allele of the sQTL, with 0,1,2 being reference, heterozygous, and alternate alleles, respectively. N and T are naïve and treated conditions respectively. **(f)** The isoform ratios by DGRP line in the two conditions. The shape of the point indicates the isoform and the colour of the dashed line indicates genotype. **(g)** Gene diagram of *fbl6* showing its multiple linked *cis*-sQTLs (blue dashed line corresponds to the *cis*-sQTL plotted in the previous panels.).

4.2.6 Intron retention is increased following infection across a natural population

To show that intron retention differences are not unique to the w^{1118} strain, we performed the same analyses for intron retention as in **Fig. 4:1**, this time on each DGRP line. Interestingy, the same pattern emerges across all lines, with more intron retention occurring after infection. **Fig. 4:6a** shows intron retention events that are significant in more than 4 lines (also see **Supplementary Fig. 4:6a**). There is a high degree of overlap among the DGRP lines, as well as between the DGPR and the w^{1118} data (**Supplementary Fig. 4:6b**), suggesting that this phenomenon is not random across the genome, but affects a specific set of introns. Interstingly, a metaplot of the location of retained and spliced introns shows that the density of retained introns is very high at the 5' end of transcripts, which could at least partly explain why longer UTRs are being produced after infection (**Fig. 4:6c**).

4.2.7 Retained introns have exon-like characteristics

We next opted to characterize the retained and spliced introns. Specifically, we sought to compare their length and GC content, both of which are known signals determining exon and intron specification (Amit, Donyo et al. 2012, De Conti, Baralle et al. 2013). In terms of length, retained introns tended to be shorter than their spliced counterparts (**Fig. 4:6d, Supplementary Fig. 4:6b**). In addition to that, their GC

contents were higher than those of the spliced introns, and consequently the difference in GC content between the intons and their flanking exons is lower (**Fig. 4:6e**). Interestingly, the retained introns also had a higher RNA polymerase II occupancy before and after infection (**Fig. 4:6f**, **Supplementary Fig. 4:6c**, see **Methods**). These observations suggest that the retained introns have exon-like characteristics which might explain why they are more prone to be retained.

Exactly why infection leads to more retention of those introns is still not evident. One possibility is that RNA-binding proteins could be differentially reacting to infection, thus leading to the observed differences. RNA-binding proteins contribute to splicing by binding specific targets in nascent transcripts in a context dependent manner (Glisovic, Bachorik et al. 2008, Fu and Ares Jr 2014). This is why we next sought to assess enrichment of RNA-binding motif (RBM) sites in the retained and spliced introns, compared to all the introns that do not change significantly. We used AME (McLeay and Bailey 2010), from the MEME suite (Bailey, Johnson et al. 2015), to look for enrichment of experimentally-derived RBMs (Ray, Kazan et al. 2013). Interestingly, we found enrichment of many RBMs in the spliced introns, but very few RBMs in the retained ones (**Fig. 4:6g**). This is in line with our previous findings that retained introns generally have weaker splicing signals and thus their splicing could be compromised by the radical infection-induced differences in transcription. Interestingly, Braunschweig and colleagues have shown that there is widerspread intron retention in humans and mouse samples (under steady-state conditions) that is coupled to RNA Pol II pausing (Braunschweig, Barbosa-Morais et al. 2014). In addition, they show that reduced intron length and higher GC content are predictors of intron retention. These parallels suggest that intron-retention is a conserved mechanism that has functional implication in normal and diseased physiology.

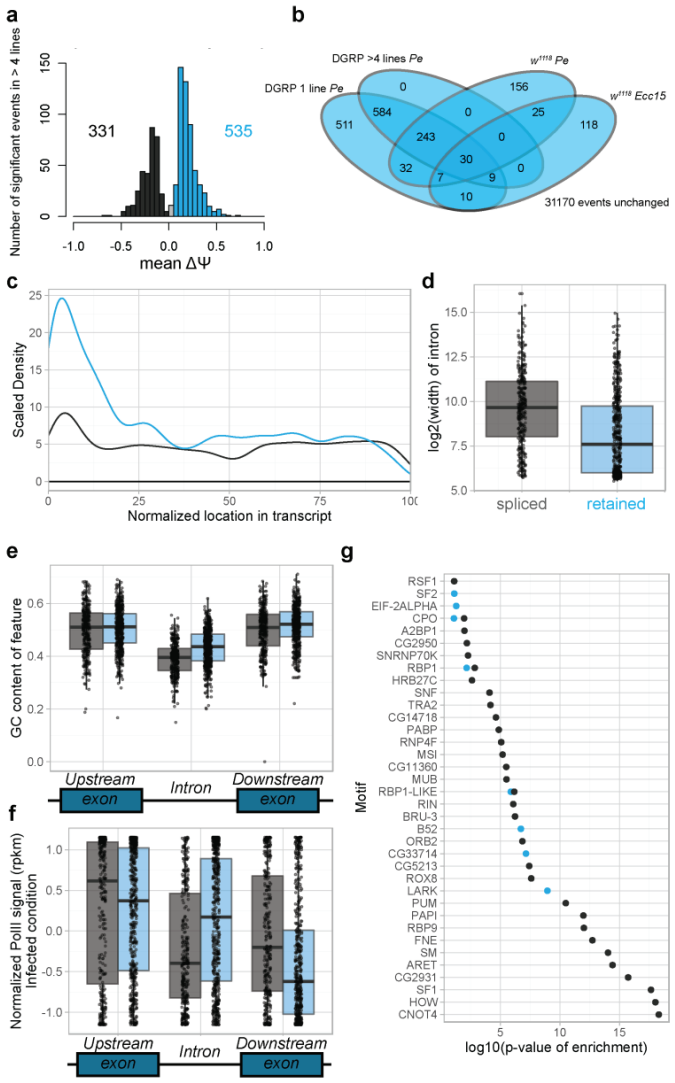


Figure 4:6 **Retained introns have exon-like characteristics**

Throughout the figure, blue and grey represent retained and spliced out introns, respectively. For each sample DGRP line, delta PSI values for different splicing events (McManus, Coolon et al. 2014) were calculated by subtracting the PSI value of the uninfected sample from that of the infected one. **(a)** Histogram of delta PSI values of intron retention events whose PSI values are significantly different after infection in at least 4 DGRP lines. **(b)** Venn diagram of the overlap between the sets of events that are significant in 1 DGRP line, at least 4 DGRP lines, w^{1118} strain infected with *Pe*, and w^{1118} strain infected with *Ecc15.* **(c)** The density of the intron retention events along the normalized length of the gene. **(d)** Length of introns (in log2) in significant intron retention events. **(e)** GC content of those introns and their flanking exons. **(f)** Normalized PolII ChIP-seq signal of these introns and their flanking exons in the *P.e.* infected state. **(g)** The enrichment of *Drosophila melanogaster* RNA binding motifs (Ray, Kazan et al. 2013) calculated using AME (McLeay and Bailey 2010), in the MEME suite (Bailey, Johnson et al. 2015). Blue and grey points indicate enrichment among the sequences of retained introns and spliced introns, respectively.

4.2.8 The RNA-binding protein lark/RBM4 is involved in the defense response

Since the Lark RBM was the most enriched in the sequences of the retained introns, we sought to investigate its possible involvement in the gut response. Lark is the ortholog of human RBM4, a protein that is implicated in splicing, translation, and the stress response. In humans, it has been shown to be activated through phosphorylation by the p38 MAPK pathway in response to stress, where it shuttles out of the nucleus and affects translation of different proteins. In Drosophila, the MAPK pathway, specifically through *p38c* has been shown to mediate the gut immune response to enteric infection through its effect on Atf-2 transcription factor (Chakrabarti, Poidevin et al. 2014). In the DGRP lines, *lark* seems to be mainly induced following infection, with a subset of the susceptible lines having higher induction (**Fig. 4:7a**).

We pursued two strategies to investigate Lark's involvement. The first is by looking at the effect of p-element insertions within or upstream of the lark locus on infection susceptibility and the second by overexpressing it and knocking it down specifically in the adult gut enterocytes. We observed that a reduction in *lark* levels due to p-element insertions in either its 5'UTR or around 300 bases upstream lead to enhanced survival to infection (**Fig. 4:7b**). We used RT-qPCR to show that lark induction is reduced in these lines compared to a wild-type control (**Fig. 4:7b**). However, we were surprised to see that both knockdown and overexpression of *lark* in adult enterocytes resulted in strikingly enhanced survival (**Fig. 4:7c**). We validated lark knockdown and overexpression by performing RT-qPCR on dissected guts and found that indeed, there was up to 80% knockdown and 80-100 times overexpression in comparison to control levels. Collectively, the observations point to a significant contribution of lark in the gut response to infection, however the mechanism of action of this gene remains to be elucidated.

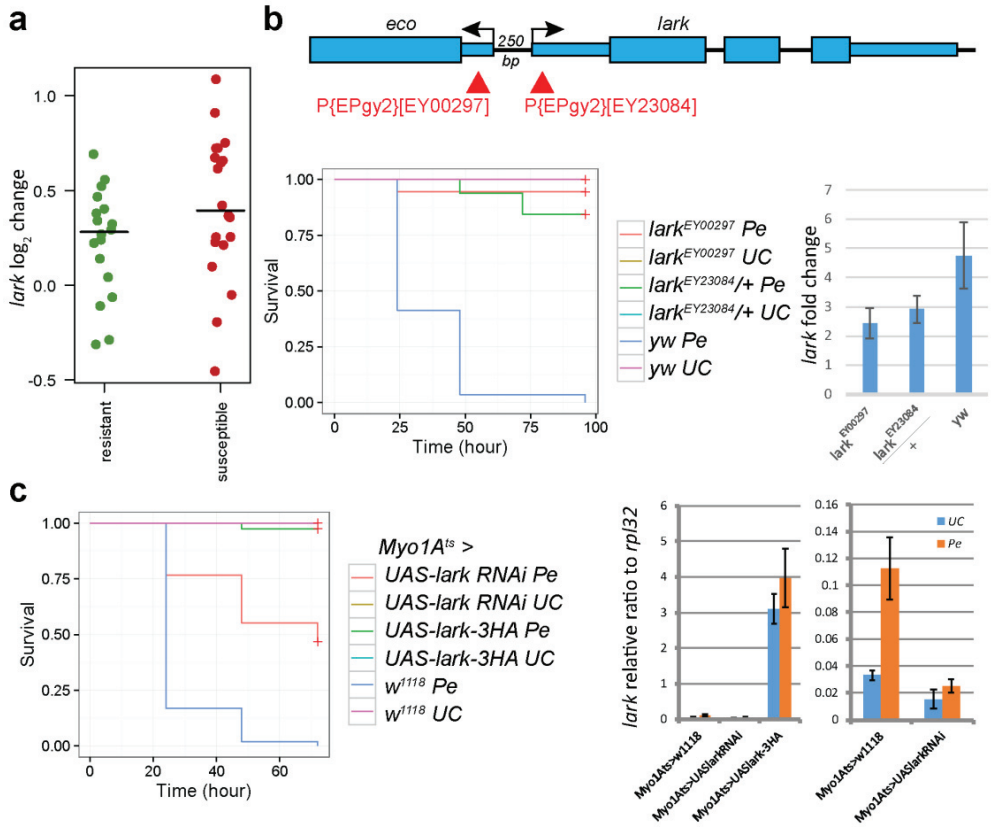


Figure 4:7 **The RNA-binding protein lark/RBM4 is involved in the defence response**

(a) Change in $\log_2$(cpm) of lark levels upon infection in the 38 DGRP lines separated by resistance class. **(b) Top panel:** diagram showing the locations of the p-element insertions in or around the lark locus. **Lower left:** Survival curves of the lark p-element lines compared to a *yw* wildtype. **Lower right:** RT-qPCR-based fold change of *lark* levels in dissected guts of those flies four hours post infection with *Pe*. **(c)** Survival of lark overexpression and knockdown flies driven by the *Myo1A*^{ts} Gal4 driver. Myo1Ats virgins were crossed to either *UAS-lark RNAi*, *UAS-lark-3HA*, or *w*^{1118} males and their F1 progeny were maintained at 18°C. After eclosion, adults were kept at 29°C for 7 days, then infected with *Pe*. Left panel: Survival curves the F1 flies after infection with Pe. **Right panel:** relative ratio of *lark* in dissected guts of those flies 4 hours after infection with Pe. The left and right graphs are based on the same data, but the right one excludes the overexpression construct for visual clarity. (All experiments were performed with three biological replicates and n>30 flies or guts)

4.3 Discussion

The gut response to infection and stress is a collection of concerted mechanisms that optimally lead to the clearance of the pathogen and the restoration of homeostasis. An organism must quickly and reversibly adapt to the challenge to ensure survival. Transcription factors that act in response to stimuli, such as Relish (the IMD pathway), Atf-2 (MAPK pathway), STAT92E (JAK/STAT pathway) have all been studied in the context of gut infection, damage, and regeneration (Lemaitre and Hoffmann 2007, Buchon, Broderick et al. 2009, Buchon, Broderick et al. 2013, Kuraishi, Hori et al. 2013, Lemaitre and Miguel-Aliaga 2013). We and others have catalogued possible targets of those transcription factors by using high-throughput techniques as well as classical ones. However, a major aspect of gene regulation, splicing, has largely been ignored. The fact that only 25% of the genes in our data that have significant splicing differences are also differentially expressed is strong motivation to comprehensively explore the role of splicing.

While there are several examples of interactions between splicing and cell stress (Biamonti and Caceres , Lin, Hsu et al. 2007, Dutertre, Sanchez et al. 2011, Ip, Schmidt et al. 2011, Shalgi, Hurt et al. 2014), there have been very few genome-wide studies addressing the issue (Shalgi, Hurt et al. 2014). In this study, we show that infection leads to a widespread and consistent splicing changes in 39 *Drosophila* strains. Many of the major differences we observe are at the level of 5' UTRs, which means that infection-induced splicing changes could have consequences on regulation, rather than strictly generating protein diversity. In times of stress, the gut might be producing transcripts coding for the same protein species, albeit with different spatial and temporal dynamics.

One important aspect of the gut response to pathogenic bacteria is the general inhibition of translation, which has been previously shown to be dependent on the activation of GCN2 kinase. Activated GCN2 kinase phosphorylates the alpha subunit of the eukaryotic initiation factor (eIF2α), which leads to inhibition of translation initiation. Paradoxically, and specifically after cellular stress, some proteins like ATF4 and ATF5 rely on upstream open reading frames (uORFs) to circumvent translational inhibition (Vattem and Wek 2004, Watatani, Ichikawa et al. 2008, Hatano, Umemura et al. 2013). The presence of uORFs generally inhibits the main ORF, unless they are found in specific configurations and in certain cellular conditions, like in the cases of ATF4 and ATF5 after stress-induced eiF2α phosphorylation. It is possible that the production of longer 5' UTRs, through intron retention or alternative TSS choice, could introduce upstream open reading frames (uORFs), as well as other elements, further contributing to this inhibition of translation (Calvo,

Pagliarini et al. 2009, Waern and Snyder 2013, Wethmar 2014, Johnstone, Bazzini et al. 2016). This also opens the possibility for the production of isoforms that are resistant to inhibition of translation or even isoforms whose translation efficiency is enhanced in stress conditions. For instance, it has been shown that the presence of uORFs in 5'UTRs could affect the recruitment of an isoform to polyribosomes, thus contributing the the translation efficiency (Sterne-Weiler, Martinez-Nunez et al. 2013). Therefore, the poor correlations observed between transcript levels and protein abundances in other systems, could be due to the fact that splicing has been consistently ignored. Therefore, proteomics and ribosomal profiling studies on the fly gut in the infected and non-infected state, paired to the extensive transcriptomic knowledge we have amassed, would be very helpful in bridging the gap between the transcriptomic and proteomic aspects of the gut response.

The observation that retained introns were enriched for the lark motif lead us to investigate the involvement of lark/RBM4 in the gut defense response. In the fly, this gene has mostly been studied in the context of the circadian clock and eye development (Newby and Jackson 1996, Huang, McNeil et al. 2014). In mammals, however, many reports have been published implicating it in regulation of splicing, transcript stability, and translation control. Importantly, it has been shown to be phosphorylated by the p38 MAPK in response to stress, where it translocates out of the nucleus and inhibits Cap-dependent translation while enhancing IRES-dependent translation (Lin, Hsu et al. 2007). Our intuition that higher lark levels would lead to infection susceptibility were proven wrong, as we saw that both lower levels and higher levels of lark, compared to the wild-type, significantly enhanced resistance. Lark/RBM4 seems to be intimately involved in the gut response, yet its exact mechanism of action is still unclear, and merits further investigation. Specifically, the genome-wide effect of lark overexpression and knockdown on intron retention would be a good starting point, followed by cross-linking immunoprecipitation and sequencing (CLIP-seq) to identify lark targets, especially at the intron level in nascent transcripts.

4.4 Materials and Methods

4.4.1 Fly Stocks and infection experiments

For the RNA-seq on the DGRP lines, the same samples of **Chapter 3** were used. We used w1118 and yw flies as wildtype. The UAS-lark RNAi line was obtained from the Transgenic RNAi Project (TRiP.JF02783) and the UAS-lark-3HA line was obtained from Bloomington stock center (stock # 7125). For specific knockdown or overexpression of *lark* in the adult gut enterocyte, F1 lines carrying a copy of the *MyoIA-Gal4* and *tub-Gal80^{ts}* transgenes (Jiang, Patel et al. 2009), as well as one copy of either the *UAS-IR* or the UAS-ORF were kept at 18°C for three days post-eclosion, and then moved to 29°C for 8 days to activate the *UAS* transgenes. Flies were subsequently infected with *P.e.* using the standard oral infection protocol. The P-element insertion lines in lark were obtained from Bloomington stock center (stock #15287 and #22604). Survival was counted every 24 hours as previously described.

4.4.2 RNA extraction

For the all samples, guts from 30 adult female flies were freshly dissected in PBS after four hours of treatment. RNA extraction was performed using Trizol Reagent (Invitrogen) using the standard protocol.

4.4.3 RT-qPCR

cDNA was then synthesized from 1 ug total RNA using *SuperScript II* enzyme (Invitrogen). qPCR experiments were performed on a StepOnePlus Real-Time PCR system (Applied Biosystems) using Power SYBR® Green PCR Master Mix (Applied Biosystems). Relative gene expression was calculated after normalization to the control *RpL32* mRNA.

4.4.4 RNA-seq

Library preparation and sequencing: For the *w^{1118}* samples, paired-end Illumina Truseq libraries were generated and sequenced on an Illumina HiSeq 2500 for 75 cycles in the Gene Expression Core Facility at EPFL. As for the 80 DGRP samples, single-end Illumina Truseq libraries were sequenced for 100 cycles on an Illumina HiSeq 2000 at the Genomics Technology Platform of the University of Lausanne.

Quality control: For the 76 single end DGRP samples, the same quality control measures were applied as in **Chapter 2**. As for the paired-end w1118 samples, we used cutadapt version 1.8 to remove adapter sequences as well as bases with a quality score inferior to 20. FastQC version 0.11.2 was used to assess the result of the trimming.

Mapping to individualized transcriptomes: Refer to **Chapter 3** Methods

Transcript ratio estimation and comparisons: We used MISO version 0.5.3 to obtain transcript ratios (PSI values) from each of the individualized genomes and annotations. We used the Ensembl BDGP 5.25 as annotation. We also extracted the assigned counts for each transcript from the MISO outputs. For the detection of genes with significantly altered isoform ratios after infection, we used the rasp package (https://www.isglobal.org/en/web/guest/statistical-software), a distance-based non-parametric multivariate approach as described in (Gonzàlez-Porta, Calvo et al. 2012). We slightly modified the package script in order to obtain the effect sizes of infection on the isoform ratios of each gene, which are normally calculated but not reported.

Intron retention analyses: We used available annotations for intron retention analysis from the Graveley lab (McManus, Coolon et al. 2014) to estimate the PSI value of each event in MISO. Then we used the miso-compare function on each sample pair (treated and naïve) to detect statistically significant differences due to infection. Events with a Bayes factor greater than 10 and a PSI difference greater than 0.2 were considered significant.

4.4.5 ChIP-seq

RNA Polymerase II ChIP-seq: For each condition, 100 w^{1118} adult female flies were killed by submerging them in liquid nitrogen. Guts were dissected on ice and stored at -80°C. On the day of the experiments, guts were homogenized in NE Buffer (15mM HEPES, 10mM KCl, 0.1mM EDTA, 0.5 mM EGTA, 350mM Sucrose, 0.1% Tween-20, 5mM MgCl2, 1mM DTT, 1mM PMSF, protease inhibitor tablet) supplemented with 1% formaldehyde using a douncer and pestle. After 10 minutes, crosslinking was quenched by the addition of Glycine for a final concentration of 0.125M. Samples were cleared by centrifugating for 4 min at 4000 rpm

and 4°C. Samples were washed twice with ice-cold NE buffer and twice with ice-cold RIPA buffer (25mM Tris-HCl pH7.6, 150mM NaCl, 0.5% Na-deoxycholate, 0.5mM DTT, 0.1% SDS, 1% NP-40, protease inhibitor tablet). Finally, samples were resuspended in 130 µl RIPA buffer and sonicated in Covaris E-220 (30 seconds, Intensity: 175, Cycles per burst 200, Duty 20%, Water level: 10). Samples were then cleared by centrifugation for 10 min, 4°C max speed. At this point, 1% of the total volume was separated as input and stored at 4°C, then, the remaining amount was diluted 1:5 in IP Dilution buffer (2.8 ml H2O, 3 µl 10% SDS, 7.2 µl 0.5M EDTA, 33 µl Triton X-100, 50.1 µl Tris-HCl pH 8.1, 100.2 µl 5M NaCl). We then added 1 µg of antibody (Abcam ab5408) and incubated the sample overnight at 4°C on a rotating platform. The next day, the sample was transferred to a tube containing 50 µl of magnetic beads (M-280 Sheep Anti-Mouse IgG) blocked overnight in Beads Blocking Buffer (8.77ml PBS 1x, 1 ml BSA 1%, 10 µl Triton X-100, 220 µl 45% Fish Gelatin) and the mixture was incubated for 2 hours at 4°C on a magnetic platform. Using a magnetic racks, beads were washed once with Low Salt Buffer (20mM Tris-HCl pH 8.1, 150 mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100), twice with High Salt Buffer (20mM Tris-HCl pH 8.1, 500 mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100), LiCl Buffer (10 mM Tris-HCl pH 8.1, 250 mM LiCl, 1mM EDTA, 1% NP-40, 1% NA-deoxycholate) and TE-NaCl buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 50 mM NaCl). In between each wash, beads were incubated 10 min at 4°C on a rotating platform. After the last wash, beads are resuspended in 500 µl of Elution Buffer (3.24 mL H2O, 50 µl Tris-HCl pH 7.5 1M, 10 µl EDTA 0.5M, 1 mL NaHCO3 0.5M, 500 µl 10% SDS, 200 µl NaCl 5M) and the input sample was supplemented with the same amount. From then on, both the input and the IP were treated similarly. We first incubated them at 37°C for 30 min with 900 rpm shaking in the presence of 7.5 µl RNAse A 20 mg/ml. We then added 10 µl of Proteinase K and incubated the sample at 55°C overnight. The next day, we added 10 µl of Proteinase K and incubated for 1h at 45°C. Samples were then spin down for 5 min at room temperature and 2000 rpm, finally, we used 500 µl of samples as starting material for Qiagen PCR purification kit, following the manufacturer instructions. We eluted the the IP and the inpiut in 30 µl. We used the Qubit dsDNA HS kit to measure the DNA load.

Library preparation: 10 ng of DNA were transferred to a low binding tube and completed to 55.5 µl with H2O. We added 3 µl of NEBNext Ultra End Repair / dA-Tailing Module Enzyme mix and 6.5 µl of Reaction buffer and incubated each tube at 20°C for 30 min, then 65°C for 30 min. The product of the reaction was purified using the Qiagen MinElute PCR Purification Kit, elution was made in 12.5 µl of Elution Buffer. For each tube, an adapter with a different barcode was selected. We used the DNA Quick ligase kit, using 15 µl of 2x buffer, 1.5 µl of DNA quick ligase and 1 µl of adapter hybrid primer. Mixture was incubated at 22°C for 30 min. The reaction was purified using the Qiagen MinElute PCR Purification Kit, elution was made in 50 µl of Elution Buffer. Samples were purified using AMPure beads in a 1:1 ratio, washed twice with 80% EtOH

and resuspended in 20 µl of Elution Buffer. Using 1 µl, we perfrom a qPCR using the KAPA SYBR green kit 50 µl total volume to determine the number of cycle for each samples. We then amplify each sample by PCR using the KAPA master mix. We then perform a size selection using AMPure beads, first using a 0.6.1 ratio and excluding the bound fraction followed by a 1:1 ratio selection, washing twice with 80% EtOH and re-suspending in 20 µl Elution Buffer. We used in 1 µl to measure the DNA load with Qubit dsDNA HS assay and 1 µl to assess the fragment profile using the Agilent Bio-analyzer DNA 12000 kit.

Sequencing: All 6 samples were sequenced on Illumina HiSeq 2500.

Mapping and analysis: The sequencing reads were mapped to the reference genome using STAR, then the counts for every intron retention event (the flanking exons as well as the intron) was counted using the regionCounts function in the R csaw package. The count data was converted to RPKM and quantile normalized prior to the analyses. Since the RNA pol II coverage decays from the 5' to the 3' end of a gene, we converted the RPKM values to the standard normal distribution for each intron retention event (the flanking exons and intron) when we were comparing the retained and the spliced events.

4.4.6 Statistical and Computational analyses

Shannon diversity: For each gene, the Shannon diversity was calculated based on the transcript ratios of its annotated isoforms in R. This was done for each RNA-seq sample. The Delta Shannon for each DGRP line was calculated by subtracting the Naïve Shannon diversity from the treated one.

Effective length calculations: We first generated tables of transcript, 5' UTR, 3' UTR, and CDS lengths for each line, taking into account the insertions and deletions in those lines. Then, for each line and condition, we calculated the effective length of a gene as the sum of the products of the length and the corresponding isoform ratio (**Fig. 4:3**).

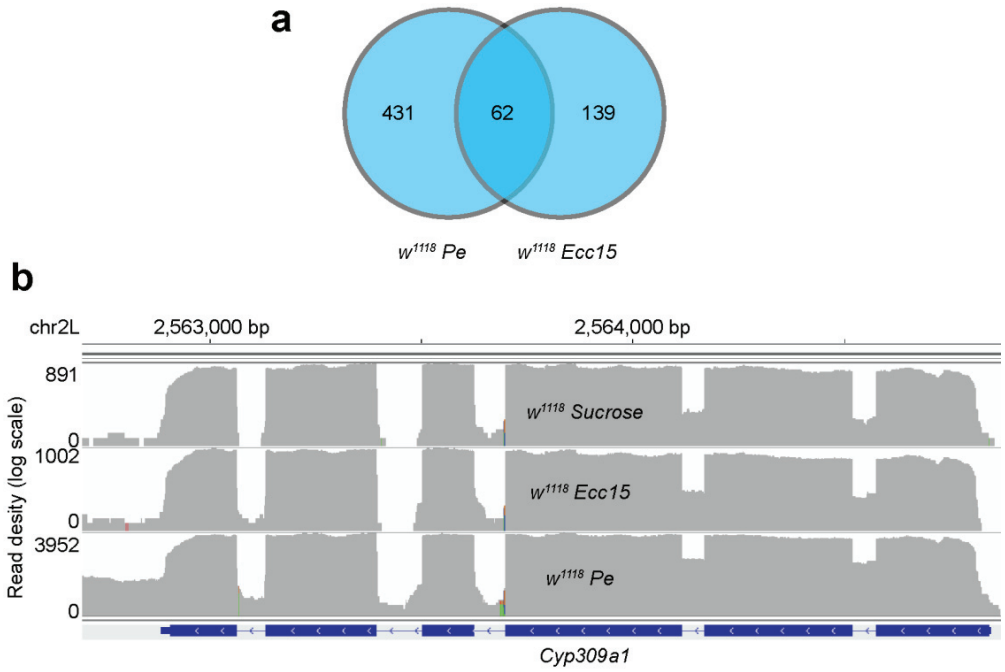
sQTL analysis: sQTL analysis was performed using sQTLSeekR (Monlong, Calvo et al. 2014) using the transcript ratios and genetic variants 10 kb around each expressed gene with multiple isoforms. We performed

slight modifications on the package script in order to extract information about the effect size of sQTLs which was normally calculated but not reported.

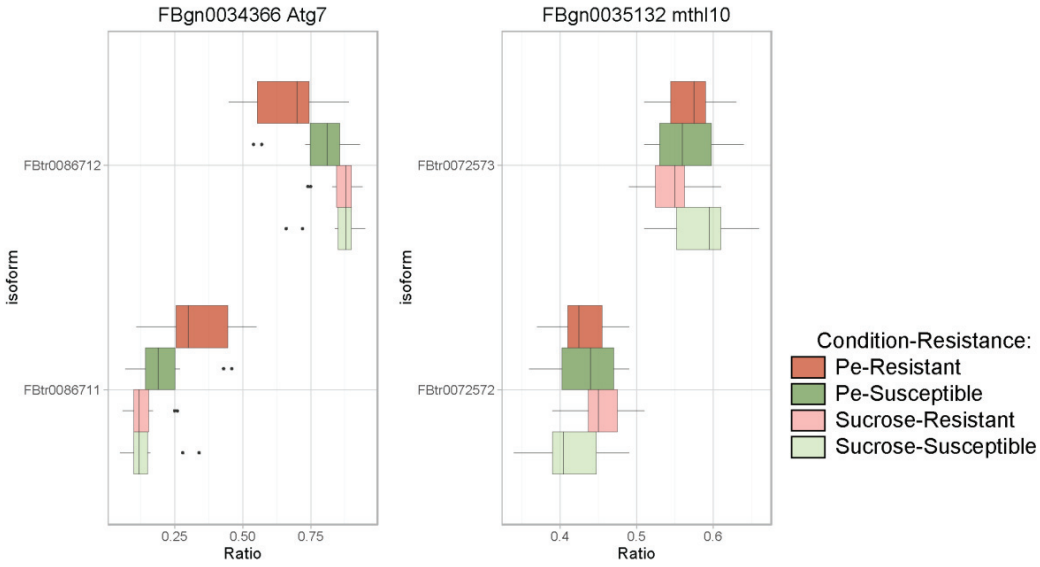
ESE and ISE analyses: We used a published set of 330 intronic and exonic splicing enhancers and pattern matching through the BSgenome and Biostrings R packages to catalogue all the possible locations of those elements within the gene bodies of the reference genome. We then calculated the percentage of sQTLs that overlap with a predicted element. To assess the overlap expected by chance, we randomly sampled, 100 times, sets of variants that are within 10kb of expressed genes that have a similar allele frequency spectrum as the sQTLs.

RNA-binding motif analyses: We used AME, from the MEME suite, to look for all binding motifs of RNA binding proteins using *Drosophila*-specific PWM scores from (Ray, Kazan et al. 2013) in retained or spliced introns. For both searches, we used the of introns that do not change significantly after infection as background.

4.5 Supplementary Materials

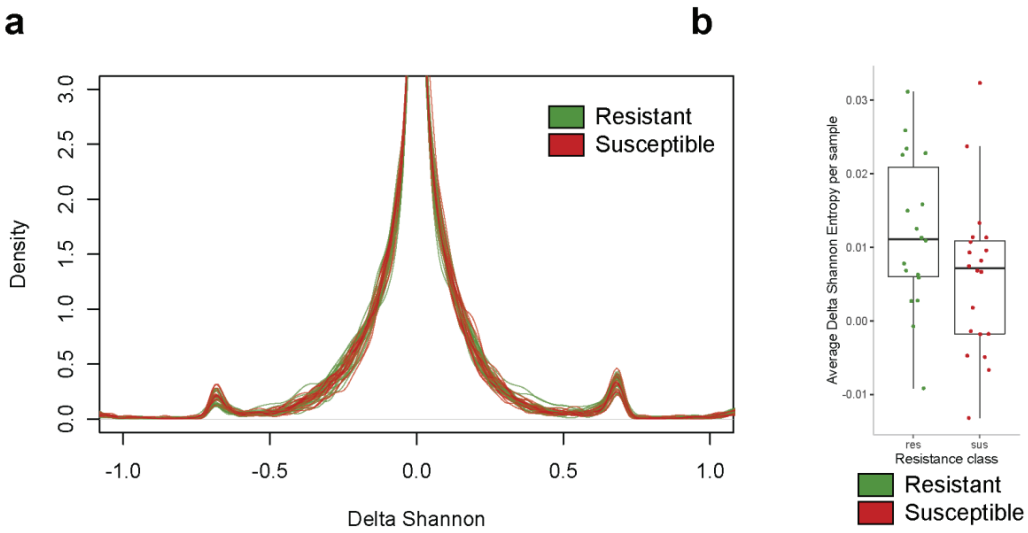


Supplementary Figure 4:1 **Enteric infection with different pathogens leads to widespread changes in intron retention** **(a)** Venn diagram showing the intersection of the significant intron retention events under the two conditions (*Pe* and *Ecc15*). **(b)** Illustration of multiple intron retention events within a single gene *Cyp309a1*. Snapshot was obtained using IGV.



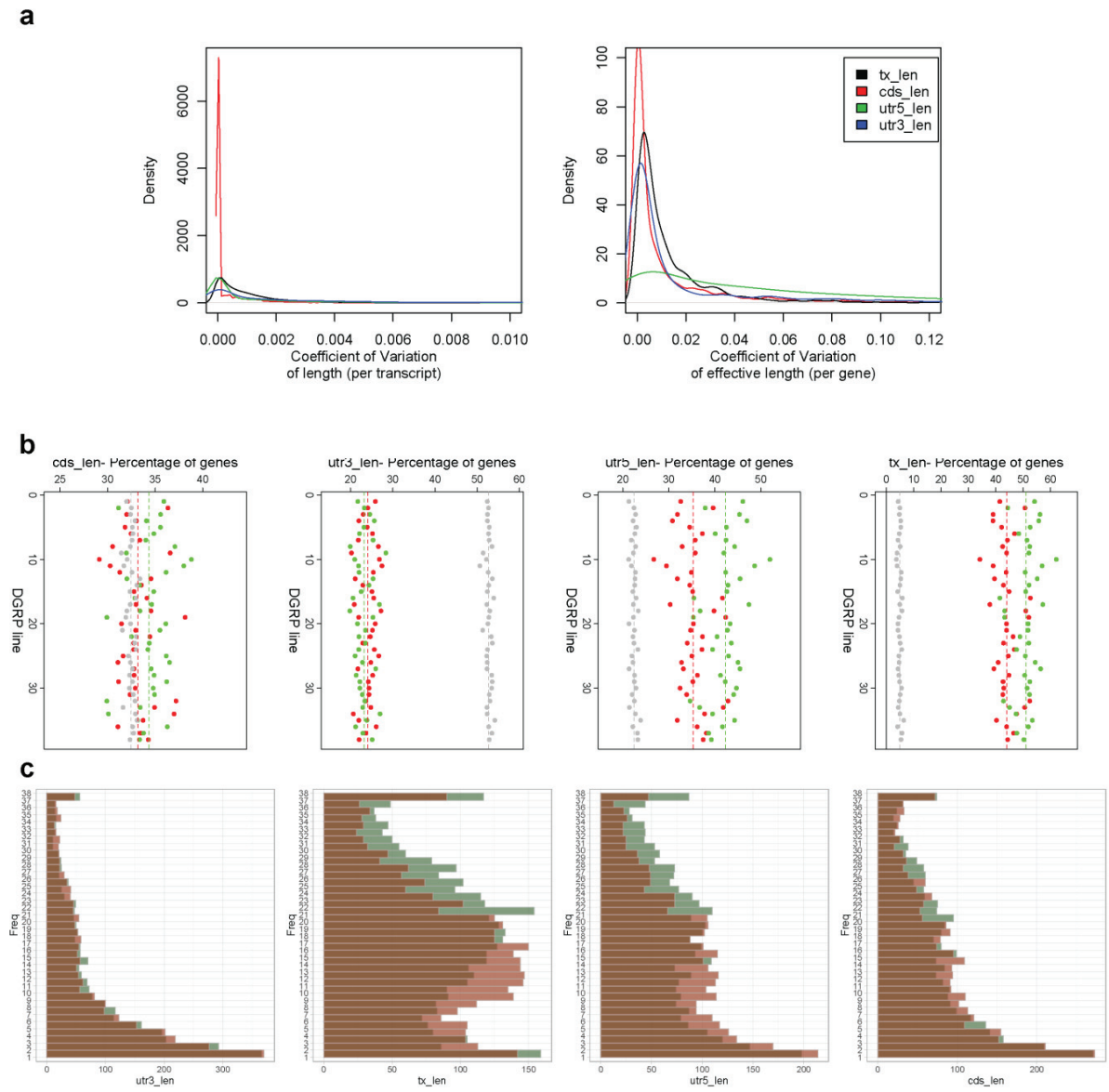
Supplementary Figure 4:2 **Enteric infection leads to extensive changes in transcript isoform ratios**

Examples of gene isoform ratios in the different conditions and susceptibility groups. Atg7 shows a shift in isoform usage upon infection, whereas mthl10 exhibits a slight difference in isoform usage between resistant and susceptible lines in the uninfected (sucrose) state.



Supplementary Figure 4:3 **The transcriptional response is characterized by higher isoform diversity**

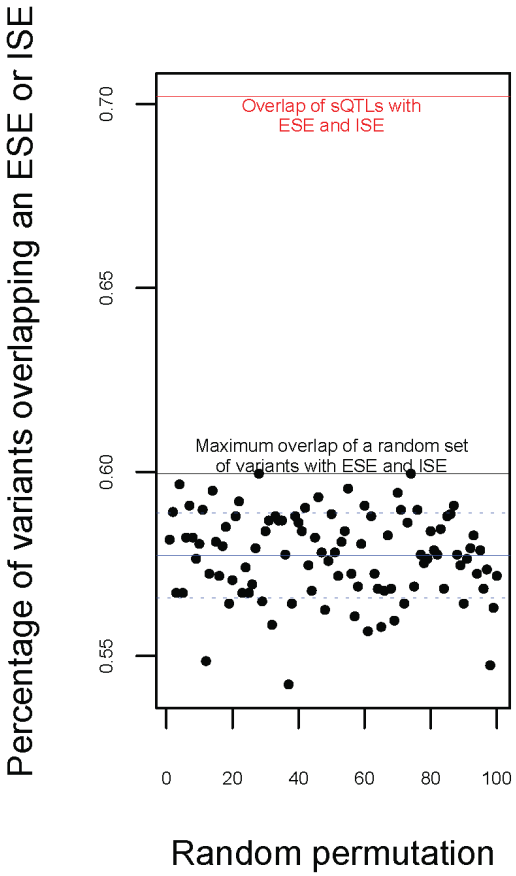
(a) Distribution of delta Shannon entropy values (Shannon entropy in infected minus uninfected state) per gene per DGRP line. **(b)** Boxplot of average delta Shannon entropy per DGRP line, separated by resistance class (one-tailed t-test p-value < 0.05).



Supplementary Figure 4:4 **Post-infection transcripts tend to be longer, mainly due to the production of longer 5' UTRs**

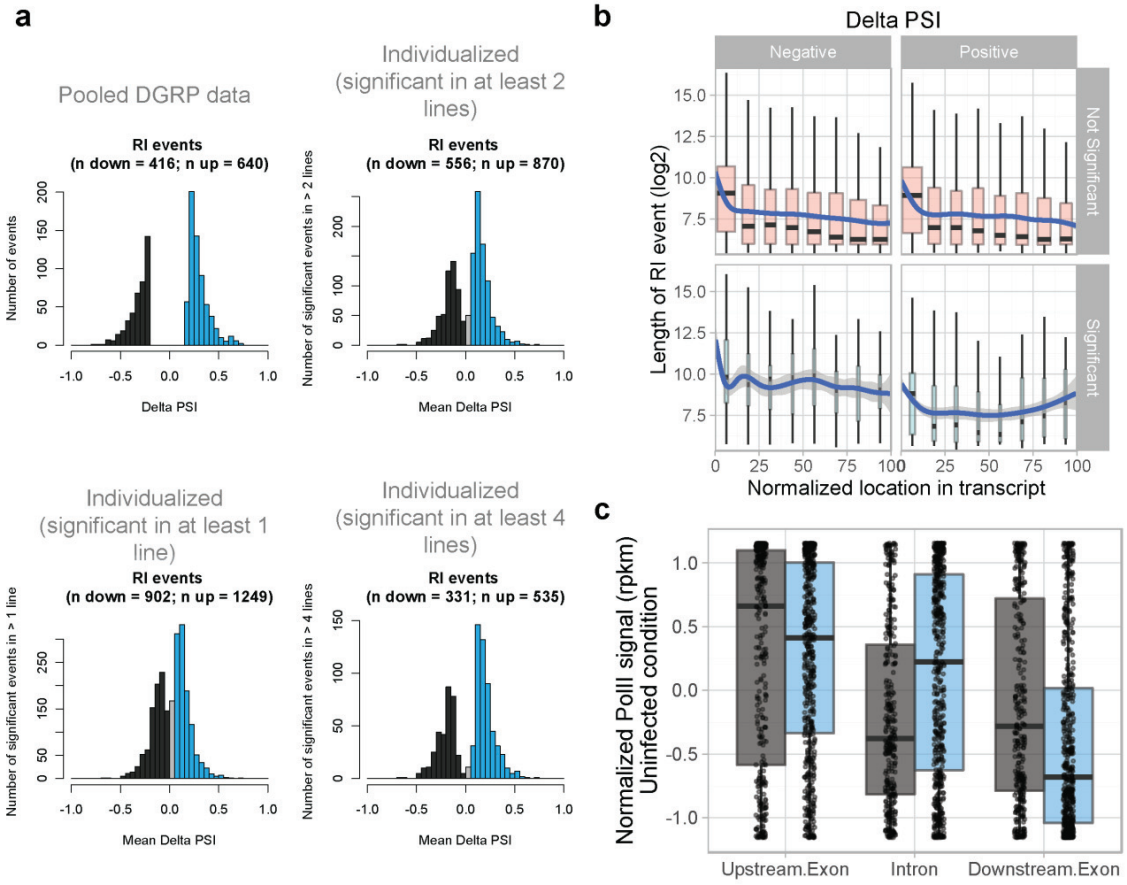
(a) Left panel: The distribution of coefficients of variation in length of each transcript or feature due to natural variation within the DGRP lines. **Right panel:** The distribution of coefficients of variation in effective length of each gene due to natural variation within the DGRP lines, isoform transcript ratios, and infection. **(b)** Breakdown of number of genes whose effective length increases (in green), decreases (in red) or stays constant (in grey) based on a certain feature (From left to right: the predicted polypeptide- (cds), 3' UTR-, 5' UTR-, and the transcript-based effective length change). **(c)** The per-gene frequency distribution among the DGRP lines of the number of genes that increase (green)

or decrease (red) in effective length based on the features (from left to right: 3' UTR, transcript, 5' UTR, and predicted polypeptide)



Supplementary Figure 4:5 **Predicted Exonic and Intronic Splicing Enhancers (ESE and ISE) are enriched for sQTLs.**

ESE and ISE locations were predicted along all gene bodies using pattern matching to the reference genome. Then the percentage of the sQTLs that overlap a predicted element was computed and plotted in red. A null distribution of the percentage overlap was produced by randomly picking variants within gene bodies with a similar allele frequency distribution as the sQTLs. This was repeated 100 times and the percentage, as well as the mean (solid line) and standard deviations (dashed lines) were computed. A solid line shows the maximum overlap obtained through random permutations.



Supplementary Figure 4:6 **Retained introns have exon-like characteristics**

(a) Histograms of delta PSI values of significantly different intron retention events after infection in the pooled alignments of all DGRP sequencing data aligned to the reference sequence (top left), significantly different events in alignments to individualized genomes in at least 1 DGRP line (bottom left), 2 DGRP lines (top right), and 4 DGRP lines (bottom right). **(b)** Distribution of intron lengths (in $\log_2$ scale) as a function of location within the transcript for non-significant (upped panel, red) and significant in at least 4 DGRP lines (lowe panel, blue) as well as retained (positive delta PSI, right) and spliced (negative delta PSI, left) introns. The blue lines are loess smoothing curves with 95% confidence intervals. **(c)** Standardized RNA polymerase II signal density in $\log_2$(rpkm) in introns and flanking exons for all significant intron retention events in the uninfected state.

Chapter 5 Conclusion

Understanding how information in the genome gives rise to complex phenotypes is a major question in modern biology. Specifically, a good understanding of how genetic variation mediates quantitative differences in traits is of great interest. Insights from such studies are highly relevant in human complex disease risk research, animal and plant breeding, and pest control.

A systems genetics study of the variability in the *Drosophila* gut response to enteric infection was presented in this thesis. The genetic architecture of this response was unraveled and novel mediators of resistance to enteric infection are presented. Using GWAS, many loci with modest effect on the phenotype were identified, some of which we validated experimentally. Using transcriptomic module analysis, as well as experimental validation, ROS metabolism was identified as an important determinant of resistance class. Large-scale transcriptomics on phenotypic extremes were analysed and used to predict the survival phenotype of *Drosophila* lines. The effect of natural variation on RNA expression and splicing was assessed. A novel player in the gut defense response, *nutcracker,* was identified. Finally, an in-depth analysis of splicing differences after infection was performed and a light was cast on lark/RBM4 as a potentially important mediator in response to enteric infection.

The presented studies are mostly hypothesis-generating, and are not designed to understand the mechanism of action of a certain gene or pathway. This is why more focused studies should be designed in order to gain a deeper understanding of the biology involved. In addition, the results of the analyses allow the greater scientific community to adopt and validate some of the hypotheses and candidates.

5.1 The Reductionist Approach *versus* Systems Genetics

One of the most valuable lessons learned from the phenotypic diversity in *Drosophila* lines is that biologists should adopt a holistic as well as the standard reductionist approach. Conclusions based on one genetic background might not be generalizable since they completely disregard natural genetic variation (Hartman, Garvik et al. 2001). For example, we show that many DGRP lines are highly resistant to *Pseudomonas entomophila* infection, with infection-induced inhibition of translation occuring in susceptible but not resistant lines. Thinking retrospectively, the *Pseudomonas entomophila* L48 strain, which was part of a panel of isolates from fruit flies or decaying fruits from the Island of Guadeloupe (Vodovar, Vinals et al. 2005), might have not been identified as a lethal bacterium if a strain other than Oregon^{R} – for instance a resistant DGRP line - were to have been used in the screen. For the same reason, it is equally possible that some other pathogenic bacteria were missed in the screen. This reasoning could be applied on almost every genetic study or screen performed in the last century.

One has to acknowledge that testing on several strains increases cost and complexity of experiments, especially large scale screening, and that it is not clear whether the advantage would outweigh the added cost. It is also impossible to estimate the knowledge that was *not* acquired after a century of reductionist approach. Including multiple genetic backgrounds in every study might prove to be counterproductive, leading to slower research progress on all fronts. What is certain is that scientists should be open to results that challenge their understanding of some basic phenomena. One pertinent example in human genetics is that some humans carrying known mutations leading to severe Mendelian childhood diseases do not have clinical manifestations (Chen, Shi et al. 2016). The simplest explanation for this resilience is the existence of variants with large effect that buffer the deleterious effect of the loss of function mutation. This is not a new biological concept. In fact, classical modifier screens in *Drosophila* and other systems exploit the interaction between alleles to identify novel members of a genetic pathway (St Johnston 2002). While these screens rely on artificial induction of mutations, this phenomenon might be pervasive in natural populations and systems genetics could be help understanding it.

For all those reasons, twenty-first century biological research should exploit as many high-throughput techniques as possible to reach generalizable conclusions.

5.2 Lessons from the Genetic Architecture of Resistance to Enteric Infection

Results in this thesis show that pathways expected or known to be involved in a certain process do not necessarily have to play a role in phenotypic differences at the population level. This is one of the most striking take-home messages in this thesis. In contrast to studies in which different fly lines (from a panel of chromosome 2 substitution lines) were inoculated with a Gram-negative bacterium, where resistance to the bacterium was shown to be mediated by variations in signal transduction and pathogen recognition genes (Lazzaro, Sceurman et al. 2004, Lazzaro, Sackton et al. 2006), we were consistently failing to detect genetic or transcriptomic differences in canonical immune response pathways of resistant and susceptible flies. Except for one tested susceptible line with a loss-of-function mutation in the *Dredd* gene, all lines responded very similarly to infection at the transcriptomic level, even though they had very different survival rates. It is therefore likely that the major players in resistance processes, such as the response to Gram-negative bacteria, behave similarly across the population due to evolutionary and functional restraints, so that many secondary or tolerance processes collectively contribute to the inter-individual differences. Taken individually, these secondary processes or factors might contribute very little to the phenotype and would therefore be very hard to detect with a reductionist approach.

It is not clear to what extent this genetic architecture could be generalized to other enteric infection models or other host-pathogen interactions, such as septic injury or natural fungal infections. The lack of canonical immune pathways in the genetic association or transcriptomic results could mean that susceptibility to enteric infection with *P.e.* is a proxy to susceptibility to general stress, specifically ROS-induced stress. It therefore suggests that tolerance, and not resistance, mechanisms are the dominant players determining resistance. Indeed, when we measured resistance to paraquat in eight lines from the phenotypic extremes, the lines susceptible to *P.e.* where also susceptible to paraquat treatment. Our survival data on the full DGRP panel, however, does not correlate with resistance to paraquat-induced oxidative stress (Weber, Khan et al. 2012). This is not surprising, since in the same study, the genetic correlation between resistance to two oxidative stress-inducing agents, namely paraquat and menadione sodium bisulfite were not high. Additionally, the GWAS results of the two treatments had very little overlap, indicating that different genomic loci contribute to the resistance to the two substances. In the light of this, the lack of correlation with *P.e.* resistance is not surprising, knowing that enteric infection not only induces a burst in ROS but also activates the immune response. What is common, however, is the complex genetic architecture of both treatments, and indeed that of survival to *P.e.* infection.

This complex genetic architecture is in stark contrast to that of resistance to viral infection. Studies in the DGRP as well as another unrelated panel of flies showed that the genetic architecture is much simpler (Magwire, Fabian et al. 2012, Cogni, Cao et al. 2016). This is consistent with theoretical models where selection pressures exerted by pathogens lead the increase in frequency of major-effect resistance alleles, and consequently to a simple resistance genetic architecture (Hill 2012). These resistance alleles are expected to be specific to a certain pathogen and therefore this model only applies to co-evolved pathogens, like the Drosophila C virus and the Sigma virus. Host-pathogen co-evolution and its effect on the genetic architecture of resistance is not specific to viruses and flies. For example, a study on cholera susceptibility in a human population from the Granges River Delta, the historic epicenter of cholera, showed that many of the genes that are positively-selected are also associated with cholera susceptibility (Karlsson, Harris et al. 2013). On the other hand, *P.e.* is not a *Drosophila melanogaster*-specific pathogen and like other members of the *Pseudomonas* genus, is an opportunistic pathogen (Boucias and Pendland 1998). Consequently, it is conceivable that no specific large effect resistance alleles have been selected for in the fly. We believe that small-effect variants predominantly affecting tolerance mechanisms, including ROS metabolism, mediate variation in resistance.

Another argument as to why variability in resistance to enteric infection is not mediated by immune processes is that since the fly has no adaptive immune system (Lemaitre and Hoffmann 2007), and consequently limited specificity in the response, it has to maintain the balance between the response to infectious bacteria and tolerance to gut microbiota (Ryu, Kim et al. 2008, Paredes, Welchman et al. 2011, Bosco-Drayon, Poidevin et al. 2012, Lee, Kim et al. 2013). Any genetic variation causing high variability in the population could therefore interfere with this finely-tuned system, leading to undesirable consequences. For these reasons, variation in immune capability, like Imd pathway activation, is likely to be highly restrained, only leaving room for variation in other aspects such as stress tolerance mechanisms like ROS metabolism.

In conclusion, the origin, history, and specificity of the host-pathogen interaction defines the genetic architecture of host resistance. Understanding those aspects using systems genetics could lead to a more complete and unbiased understanding of resistance to infectious disease.

5.3 Lessons from Gene Expression Profiling and Prospects

We were initially interested in finding genes that could explain the differences between susceptible and resistant lines. To our dismay, we found that molecular differences even between phenotypic extremes do not have to be consistent. Despite the clear phenotypic differences, differential gene expression surprisingly failed to detect genes that are consistently different between susceptible and resistant lines. We had to resort to module-level analyses and to machine-learning approaches to identify gene signatures of resistance class. We believe that different lines have various combinations of risk factors that collectively lead to a certain survival phenotype. It is the system as a whole, and not one or few genes that define resistance to infection. A departure from a one-gene-at-a-time approach such as module analysis should therefore be standard procedure for studying gene expression.

We identified many associations between gene expression levels and genomic variants in the eQTL analyses, many of which were shared between the infected and non-infected state. To what extent does genetic variation lead to variation in resistance through its effect on gene expression has yet to be systematically assessed. Nevertheless, we focus on *ntc*, as it is the most differentially expressed gene with respect to resistance class. Interestingly, mutants in *ntc* are more susceptible to P.e. infection. Also, the fact that it has *cis*-eQTLs around it suggest that we might be close to indentifying a causal variant in an enhancer element affecting resistance to infection. The specific cell type(s) where *ntc* exerts its effects are still unknown. For this reason, cell-specific knockdown and overexpression of *ntc* is underway. Furthermore, the *ntc* eQTLs will be validated using reporter assays, transcription factor-DNA interaction assays, and allele specific RT-qPCR. Allele-specific GAL4 reporters of the regions around the *cis*-eQTLs would be useful to implicate variations in the putative *ntc cis*-regulatory elements in the determination of *ntc* levels. Furthermore, the interaction between the polymorphic DNA sequences around the eQTLs and the respective predicted transcription factors using MITOMI microfluidic technology will be measured (Rockel, Geertz et al. 2012). This will allow the assessment of the effect of the eQTL variants on binding affinity. Allele-specific RT-qPCR on DGRP F1 hybrids harboring the two alleles of each of the eQTLs will help confirm the *cis*-effect of the regions around the eQTLs. ChIP sequencing of the *daughterless*, *broad-complex*, and *relish* transcription factors in the gut before and after infection would also show whether these transcription factors bind to the regions overlapping the eQTLs. The characterization of the function and regulation of *ntc* will provide a unique example of regulatory variation affecting an ecologically-relevant complex trait like enteric infection susceptibility.

Another aspect of gene expression that should not be ignored is RNA splicing (Levanon and Sorek 2003). Not only it is plays an important part in general gene expression, but also in gene regulation during stress responses (Biamonti and Caceres , Yost and Lindquist 1986, Ali and Reddy 2008, Dutertre, Sanchez et al. 2011). In this thesis, the potential importance of splicing is brought forward due to many interesting observations. First enteric infection leads to widespread differences in isoforms of genes, disproportionately affecting splice sites at the 5' end of transcripts, and leading to generally longer 5' UTRs. The functional relevance of these observations, however, remains to be assessed in future work, especially since splicing is not isolated from transcription and nuclear RNA export (Reed 2003, Kornblihtt, De La Mata et al. 2004). Furthermore, the impact of splicing changes on the gut proteome is still not measured. We (Michael Frochaux and I) have recently started generating large-scale gut proteomics data (up to 4500 proteins) for the DGRP lines as well as for reference strains in the normal and infected condition. Since the projects are still in their infancies, I chose not to present the preliminary data in this thesis. The results of this project will add a missing link between the genotype and phenotype. Studies in other systems, as well as our preliminary studies, show little correlation between transcript levels and protein levels. Integrating the RNA-sequencing data with the proteomics data will therefore be of great interest to the scientific community working on the gut. Furthermore, it will open the door for validating many hypotheses relating to the alternate 5' UTR-mediated effect on translation efficiency. In addition to that, my colleague Michael Frochaux is working on optimizing ribosomal profiling experiments in order to directly assess ribosomal occupancy of different transcripts. This will open many avenues in understanding the role that splicing has in enteric infection, specifically exploring whether different isoforms are preferentially recruited to ribosomes after infection.

We also focus on *lark/RBM4* as a potentially relevant factor in post-infection splicing regulation. First, we find that sequences with retained introns after infection are enriched for its binding motif and depleted of other, more common motifs. Second, *lark* overexpression and knockdown lead to increased resistance to *P.e*. infection. Last but not least, the mammalian ortholog of lark, RBM4, has already been implicated in RNA splicing and regulation of translation in normal and stressed conditions. In order to understand the effect of lark on splicing, RNA-sequencing *lark* knockdown and overexpression will be especially important. CLIP-sequencing could identify and validate its binding partners, specifically the introns it binds before and after infection.

Bibliography

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y.-H. C. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, Miklos, J. F. Abril, A. Agbayani, H.-J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. d. Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M.-H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. C. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Sidén-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z.-Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, T. Woodage, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R.-F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin and J. C. Venter (2000). "The Genome Sequence of Drosophila melanogaster." Science **287**(5461): 2185-2195.

Akira, S., S. Uematsu and O. Takeuchi (2006). "Pathogen Recognition and Innate Immunity." Cell **124**(4): 783-801.

Albert, F. W. and L. Kruglyak (2015). "The role of regulatory variation in complex traits and disease." Nat Rev Genet **16**(4): 197-212.

Ali, G. S. and A. S. N. Reddy (2008). Regulation of Alternative Splicing of Pre-mRNAs by Stresses. Nuclear pre-mRNA Processing in Plants. A. S. N. Reddy and M. Golovkin. Berlin, Heidelberg, Springer Berlin Heidelberg**:** 257-275.

Altenburg, E. and H. J. Muller (1920). "The Genetic Basis of Truncate Wing,—an Inconstant and Modifiable Character in Drosophila." Genetics **5**(1): 1-59.

Amancio, R. T., A. M. Japiassu, R. N. Gomes, E. C. Mesquita, E. F. Assis, D. M. Medeiros, B. Grinsztejn, P. T. Bozza, H. C. Castro-Faria Neto and F. A. Bozza (2013). "The innate immune response in HIV/AIDS septic shock patients: a comparative study." PLoS One **8**(7): e68730.

Amcheslavsky, A., J. Jiang and Y. T. Ip (2009). "Tissue Damage-Induced Intestinal Stem Cell Division in Drosophila." Cell Stem Cell **4**(1): 49-61.

Amit, M., M. Donyo, D. Hollander, A. Goren, E. Kim, S. Gelfman, G. Lev-Maor, D. Burstein, S. Schwartz, B. Postolsky, T. Pupko and G. Ast (2012). "Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition." Cell Reports **1**(5): 543-556.

Anderson, M. J. (2001). "A new method for non-parametric multivariate analysis of variance." Austral Ecology **26**(1): 32-46.

Apidianakis, Y. and L. G. Rahme (2011). "Drosophila melanogaster as a model for human intestinal infection and pathology." Dis Model Mech **4**(1): 21-30.

Ayroles, J. F., M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman, M. M. Magwire, S. M. Rollmann, L. H. Duncan, F. Lawrence, R. R. H. Anholt and T. F. C. Mackay (2009). "Systems genetics of complex traits in Drosophila melanogaster." Nat Genet **41**(3): 299-307.

Bader, M., E. Arama and H. Steller (2010). "A novel F-box protein is required for caspase activation during cellular remodeling in Drosophila." Development (Cambridge, England) **137**(10): 1679-1688.

Bader, M., S. Benjamin, O. L. Wapinski, D. M. Smith, A. L. Goldberg and H. Steller (2011). "A conserved F-box–regulatory complex controls proteasome activity in Drosophila." Cell **145**(3): 371-382.

Bailey, T. L., J. Johnson, C. E. Grant and W. S. Noble (2015). "The MEME Suite." Nucleic Acids Research.

Barreiro, L. B. and L. Quintana-Murci (2010). "From evolutionary genetics to human immunology: how selection shapes host defence genes." Nat Rev Genet **11**(1): 17-30.

Bateson, W., G. Mendel and W. M. Wheeler (1902). Mendel's principles of heredity; a defence by W. Bateson ... With a translation of Mendel's original papers on hybridisation. Cambridge, University press.

Baye, T. M., T. Abebe and R. A. Wilke (2011). "Genotype–environment interactions and their translational implications." Personalized medicine **8**(1): 59-70.

Bayne, C. J. (2003). "Origins and Evolutionary Relationships Between the Innate and Adaptive Arms of Immune Systems." Integrative and Comparative Biology **43**(2): 293-299.

Beavis, W. D. (1998). "QTL analyses: power, precision, and accuracy." Molecular dissection of complex traits **1998**: 145-162.

Biamonti, G. and J. F. Caceres "Cellular stress and RNA splicing." Trends in Biochemical Sciences **34**(3): 146-153.

Bogdan, C., M. Röllinghoff and A. Diefenbach (2000). "Reactive oxygen and reactive nitrogen intermediates in innate and specific immunity." Current Opinion in Immunology **12**(1): 64-76.

Bonnay, F., E. Cohen-Berros, M. Hoffmann, S. Y. Kim, G. L. Boulianne, J. A. Hoffmann, N. Matt and J.-M. Reichhart (2013). "big bang gene modulates gut immune tolerance in Drosophila." Proceedings of the National Academy of Sciences.

Boomsma, D., A. Busjahn and L. Peltonen (2002). "Classical twin studies and beyond." Nat Rev Genet **3**(11): 872-882.

Bosco-Drayon, V., M. Poidevin, Ivo G. Boneca, K. Narbonne-Reveau, J. Royet and B. Charroux (2012). "Peptidoglycan Sensing by the Receptor PGRP-LE in the Drosophila Gut Induces Immune Responses to Infectious Bacteria and Tolerance to Microbiota." Cell Host & Microbe **12**(2): 153-165.

Bou Sleiman, M. S., D. Osman, A. Massouras, A. A. Hoffmann, B. Lemaitre and B. Deplancke (2015). "Genetic, molecular and physiological basis of variation in Drosophila gut immunocompetence." Nat Commun **6**.

Boucias, D. G. and J. C. Pendland (1998). Insect Pathogenic Bacteria. Principles of Insect Pathology. Boston, MA, Springer US**:** 177-216.

Brand, A. H. and N. Perrimon (1993). "Targeted gene expression as a means of altering cell fates and generating dominant phenotypes." Development **118**(2): 401-415.

Braunschweig, U., N. L. Barbosa-Morais, Q. Pan, E. N. Nachman, B. Alipanahi, T. Gonatopoulos-Pournatzis, B. Frey, M. Irimia and B. J. Blencowe (2014). "Widespread intron retention in mammals functionally tunes transcriptomes." Genome Research **24**(11): 1774-1786.

Brem, R. B., G. Yvert, R. Clinton and L. Kruglyak (2002). "Genetic Dissection of Transcriptional Regulation in Budding Yeast." Science **296**(5568): 752-755.

Brooks, A. N., J. L. Aspden, A. I. Podgornaia, D. C. Rio and S. E. Brenner (2011). "Identification and experimental validation of splicing regulatory elements in Drosophila melanogaster reveals functionally conserved splicing enhancers in metazoans." RNA **17**(10): 1884-1894.

Buchon, N., N. A. Broderick and B. Lemaitre (2013). "Gut homeostasis in a microbial world: insights from Drosophila melanogaster." Nat Rev Micro **11**(9): 615-626.

Buchon, N., N. A. Broderick, M. Poidevin, S. Pradervand and B. Lemaitre (2009). "Drosophila Intestinal Response to Bacterial Infection: Activation of Host Defense and Stem Cell Proliferation." Cell Host & Microbe **5**(2): 200-211.

Buchon, N., D. Osman, Fabrice P. A. David, H. Yu Fang, J.-P. Boquete, B. Deplancke and B. Lemaitre (2013). "Morphological and Molecular Characterization of Adult Midgut Compartmentalization in Drosophila." Cell Reports **3**(5): 1725-1738.

Bush, W. S. and J. H. Moore (2012). "Chapter 11: Genome-Wide Association Studies." PLoS Comput Biol **8**(12): e1002822.

Calvo, S. E., D. J. Pagliarini and V. K. Mootha (2009). "Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans." Proceedings of the National Academy of Sciences of the United States of America **106**(18): 7507-7512.

Castle, W. E. and C. C. Little (1910). "On a modified Mendelian ratio among yellow mice." Science **32**(833): 868-870.

Chakrabarti, S., P. Liehl, N. Buchon and B. Lemaitre (2012). "Infection-Induced Host Translational Blockage Inhibits Immune Responses and Epithelial Renewal in the Drosophila Gut." Cell Host Microbe **12**(1): 60-70.

Chakrabarti, S., M. Poidevin and B. Lemaitre (2014). "The <italic>Drosophila</italic> MAPK p38c Regulates Oxidative Stress and Lipid Homeostasis in the Intestine." PLoS Genet **10**(9): e1004659.

Chapman, R. F., S. J. Simpson and A. E. Douglas (2013). The insects : structure and function.

Charroux, B. and J. Royet (2012). "Gut-microbiota interactions in non-mammals: What can we learn from Drosophila?" Seminars in Immunology **24**(1): 17-24.

Chen, R., L. Shi, J. Hakenberg, B. Naughton, P. Sklar, J. Zhang, H. Zhou, L. Tian, O. Prakash, M. Lemire, P. Sleiman, W.-y. Cheng, W. Chen, H. Shah, Y. Shen, M. Fromer, L. Omberg, M. A. Deardorff, E. Zackai, J. R. Bobe, E. Levin, T. J. Hudson, L. Groop, J. Wang, H. Hakonarson, A. Wojcicki, G. A. Diaz, L. Edelmann, E. E. Schadt and S. H. Friend (2016). "Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases." Nat Biotech **34**(5): 531-538.

Civelek, M. and A. J. Lusis (2014). "Systems genetics approaches to understand complex traits." Nat Rev Genet **15**(1): 34-48.

Cockerham, C. C. and B. S. Weir (1977). "Quadratic Analyses of Reciprocal Crosses." Biometrics **33**(1): 187-203.

Cogni, R., C. Cao, J. P. Day, C. Bridson and F. M. Jiggins (2016). "The genetic architecture of resistance to virus infection in Drosophila." Molecular Ecology: n/a-n/a.

Coyne, J. A. and H. A. Orr (1989). "Patterns of speciation in Drosophila." Evolution: 362-381.

Crick, F. H. (1958). On protein synthesis. Symp Soc Exp Biol.

Cubillos, F. A., V. Coustham and O. Loudet (2012). "Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants." Current Opinion in Plant Biology **15**(2): 192-198.

Darwin, C. (1871). On the origin of species. New York :, D. Appleton and Co.

David, C. J., A. R. Boyne, S. R. Millhouse and J. L. Manley (2011). "The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65–Prp19 complex." Genes & Development **25**(9): 972-983.

De Conti, L., M. Baralle and E. Buratti (2013). "Exon and intron definition in pre-mRNA splicing." Wiley Interdisciplinary Reviews: RNA **4**(1): 49-60.

Di Fruscio, M., S. Styhler, E. Wikholm, M.-C. Boulanger, P. Lasko and S. Richard (2003). "kep1 interacts genetically with dredd/Caspase-8, and kep1 mutants alter the balance of dredd isoforms." Proceedings of the National Academy of Sciences **100**(4): 1814-1819.

Doss, S., E. E. Schadt, T. A. Drake and A. J. Lusis (2005). "Cis-acting expression quantitative trait loci in mice." Genome Research **15**(5): 681-691.

Driscoll, J., S. Brody and M. Kollef (2007). "The Epidemiology, Pathogenesis and Treatment of Pseudomonas aeruginosa Infections." Drugs **67**(3): 351-368.

Drosophila 12 Genomes Consortium (2007). "Evolution of genes and genomes on the Drosophila phylogeny." Nature **450**(7167): 203-218.

Dutertre, M., G. Sanchez, J. Barbier, L. Corcos and D. Auboeuf (2011). "The emerging role of pre-messenger RNA splicing in stress responses: Sending alternative messages and silent messengers." RNA Biology **8**(5): 740-747.

Dutta, D., Adam J. Dobson, Philip L. Houtz, C. Gläßer, J. Revah, J. Korzelius, P. H. Patel, Bruce A. Edgar and N. Buchon "Regional Cell-Specific Transcriptome Mapping Reveals Regulatory Complexity in the Adult Drosophila Midgut." Cell Reports **12**(2): 346-358.

Enuameh, M. S., Y. Asriyan, A. Richards, R. G. Christensen, V. L. Hall, M. Kazemian, C. Zhu, H. Pham, Q. Cheng, C. Blatti, J. A. Brasefield, M. D. Basciotta, J. Ou, J. C. McNulty, L. J. Zhu, S. E. Celniker, S. Sinha, G. D. Stormo, M. H. Brodsky and S. A. Wolfe (2013). "Global analysis of Drosophila Cys2-His2 zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants." Genome Research **23**(6): 928-940.

Fairfax, B. P., P. Humburg, S. Makino, V. Naranbhai, D. Wong, E. Lau, L. Jostins, K. Plant, R. Andrews, C. McGee and J. C. Knight (2014). "Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression." Science **343**(6175): 1246949.

Fairfax, B. P., S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg and J. C. Knight (2012). "GENETICS OF GENE EXPRESSION IN PRIMARY IMMUNE CELLS IDENTIFIES CELL-SPECIFIC MASTER REGULATORS AND ROLES OF HLA ALLELES." Nature genetics **44**(5): 502-510.

Falcon, S. and R. Gentleman (2007). "Using GOstats to test gene lists for GO term association." Bioinformatics **23**(2): 257-258.

Falconer, D. S. (1952). "The Problem of Environment and Selection." The American Naturalist **86**(830): 293-298.

Falconer, D. S. and T. F. C. Mackay (1996). Introduction to Quantitative Genetics.

Fisher, R. A. (1918). "The Correlation Between Relatives on the Supposition of Mendelian Inheritance." Trans. R. Soc. Edinburgh **52**.

Flint, J. and T. F. C. Mackay (2009). "Genetic architecture of quantitative traits in mice, flies, and humans." Genome Research **19**(5): 723-733.

Flores, J. and P. C. Okhuysen (2009). "Genetics of susceptibility to infection with enteric pathogens." Curr Opin Infect Dis **22**(5): 471-476.

Flori, D. A. and T. A. Mousseau (1987). "Quantitative genetics and fitness: lessons from Drosophila." Heredity **58**: 103-118.

Francesconi, M. and B. Lehner (2014). "The effects of genetic variation on gene expression dynamics during development." Nature **505**(7482): 208-211.

Franklin, A., A. W. F. Edwards, D. J. Fairbanks, D. L. Hartl and T. Seidenfeld (2008). Ending the Mendel-Fisher controversy, University of Pittsburgh Press.

Fu, X.-D. and M. Ares Jr (2014). "Context-dependent control of alternative splicing by RNA-binding proteins." Nat Rev Genet **15**(10): 689-701.

Gaffney, D. J., J.-B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad and J. K. Pritchard (2012). "Dissecting the regulatory architecture of gene expression QTLs." Genome Biology **13**(1): R7-R7.

Gerrits, A., Y. Li, B. M. Tesson, L. V. Bystrykh, E. Weersing, A. Ausema, B. Dontje, X. Wang, R. Breitling, R. C. Jansen and G. de Haan (2009). "Expression Quantitative Trait Loci Are Highly Sensitive to Cellular Differentiation State." PLoS Genetics **5**(10): e1000692.

Getz, G. S. (2005). "Bridging the innate and adaptive immune systems." Journal of Lipid Research **46**(4): 619-622.

Gibson, G. and I. Dworkin (2004). "Uncovering cryptic genetic variation." Nat Rev Genet **5**(9): 681-690.

Gibson, G., J. E. Powell and U. M. Marigorta (2015). "Expression quantitative trait locus analysis for translational medicine." Genome Medicine **7**(1): 60.

Glisovic, T., J. L. Bachorik, J. Yong and G. Dreyfuss (2008). "RNA-binding proteins and post-transcriptional gene regulation." FEBS Letters **582**(14): 1977-1986.

Goddard, M. E. and B. J. Hayes (2007). "Genomic selection." J Anim Breed Genet **124**(6): 323-330.

Gonzàlez-Porta, M., M. Calvo, M. Sammeth and R. Guigó (2012). "Estimation of alternative splicing variability in human populations." Genome Research **22**(3): 528-538.

Greenspan, R. J. (2004). Fly pushing : the theory and practice of drosophila genetics. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.

Griffing, B. (1956). "Concept of General and Specific Combining Ability in Relation to Diallel Crossing Systems." Australian Journal of Biological Sciences **9**(4): 463-493.

Ha, E.-M., K.-A. Lee, S. H. Park, S.-H. Kim, H.-J. Nam, H.-Y. Lee, D. Kang and W.-J. Lee (2009). "Regulation of DUOX by the Gαq-Phospholipase Cβ-Ca2+ Pathway in Drosophila Gut Immunity." Developmental Cell **16**(3): 386-397.

Ha, E.-M., C.-T. Oh, J.-H. Ryu, Y.-S. Bae, S.-W. Kang, I.-h. Jang, P. T. Brey and W.-J. Lee (2005). "An Antioxidant System Required for Host Protection against Gut Infection in Drosophila." Developmental Cell **8**(1): 125-132.

Ha, E. M., C. T. Oh, Y. S. Bae and W. J. Lee (2005). "A direct role for dual oxidase in Drosophila gut immunity." Science **310**(5749): 847-850.

Hales, K. G., C. A. Korey, A. M. Larracuente and D. M. Roberts (2015). " Genetics on the Fly: A Primer on the Drosophila Model System ." Genetics **201**(3): 815-842.

Hansen, T. F. (2006). "The Evolution of Genetic Architecture." Annual Review of Ecology, Evolution, and Systematics **37**(1): 123-157.

Harbison, S., L. McCoy and T. Mackay (2013). "Genome-wide association study of sleep in Drosophila melanogaster." BMC Genomics **14**(1): 281.

Hartman, J. L., B. Garvik and L. Hartwell (2001). "Principles for the buffering of genetic variation." Science **291**(5506): 1001-1004.

Hatano, M., M. Umemura, N. Kimura, T. Yamazaki, H. Takeda, H. Nakano, S. Takahashi and Y. Takahashi (2013). "The 5'-untranslated region regulates ATF5 mRNA stability via nonsense-mediated mRNA decay in response to environmental stress." FEBS Journal **280**(18): 4693-4707.

Hedengren, M., BengtÅsling, M. S. Dushay, I. Ando, S. Ekengren, M. Wihlborg and D. Hultmark (1999). "Relish, a Central Factor in the Control of Humoral but Not Cellular Immunity in Drosophila." Molecular Cell **4**(5): 827-837.

Hill, A. V. S. (2012). "Evolution, revolution and heresy in the genetics of infectious disease susceptibility." Philosophical Transactions of the Royal Society B: Biological Sciences **367**(1590): 840-849.

Houle, D., B. Morikawa and M. Lynch (1996). "Comparing mutational variabilities." Genetics **143**(3): 1467-1483.

Huang da, W., B. T. Sherman and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.

Huang, W., M. A. Carbone, M. M. Magwire, J. A. Peiffer, R. F. Lyman, E. A. Stone, R. R. H. Anholt and T. F. C. Mackay (2015). "Genetic basis of transcriptome diversity in Drosophila melanogaster." Proceedings of the National Academy of Sciences **112**(44): E6010-E6019.

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Rámia, A. Tarone, L. Turlapati, T. Zichner, D. Zhu, R. Lyman, M. Magwire, K. Blankenburg, M. A. Carbone, K. Chang, L. Ellis, S. Fernandez, Y. Han, G. Highnam, C. Hjelmen, J. Jack, M. Javaid, J. Jayaseelan, D. Kalra, S. Lee, L. Lewis, M. Munidasa, F. Ongeri, S. Patel, L. Perales, A. Perez, L. Pu, S. Rollmann, R. Ruth, N. Saada, C. Warner, A. Williams, Y.-Q. Wu, A. Yamamoto,

Y. Zhang, Y. Zhu, R. Anholt, J. Korbel, D. Mittelman, D. Muzny, R. Gibbs, A. Barbadilla, S. Johnston, E. Stone, S. Richards, B. Deplancke and T. Mackay (2014). "Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines." Genome Research.

Huang, Y., G. P. McNeil and F. R. Jackson (2014). "Translational Regulation of the DOUBLETIME/CKI?/? Kinase by LARK Contributes to Circadian Period Modulation." PLoS Genet **10**(9): e1004536.

Hughes, T. A. (2006). "Regulation of gene expression by alternative untranslated regions." Trends in Genetics **22**(3): 119-122.

Huguet, C., P. Crepieux and V. Laudet (1997). "Rel/NF-kB transcription factors and IkB inhibitors: evolution from a unique common ancestor." Oncogene **15**(24): 2965-2974.

Ip, J. Y., D. Schmidt, Q. Pan, A. K. Ramani, A. G. Fraser, D. T. Odom and B. J. Blencowe (2011). "Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation." Genome Research **21**(3): 390-401.

Janeway, C. A., P. Travers, M. Walport and M. J. Shlomchik (1997). Immunobiology: the immune system in health and disease, Current Biology.

Jennings, B. H. (2011). "Drosophila – a versatile model in biology & medicine." Materials Today **14**(5): 190-195.

Jiang, H., P. H. Patel, A. Kohlmaier, M. O. Grenley, D. G. McEwen and B. A. Edgar (2009). "Cytokine/Jak/Stat Signaling Mediates Regeneration and Homeostasis in the Drosophila Midgut." Cell **137**(7): 1343-1355.

Johnstone, T. G., A. A. Bazzini and A. J. Giraldez (2016). "Upstream ORFs are prevalent translational repressors in vertebrates." The EMBO Journal **35**(7): 706-723.

Karasov, W. H., C. Martínez del Rio and E. Caviedes-Vidal (2011). "Ecological physiology of diet and digestive systems." Annual review of physiology **73**: 69-93.

Karlsson, E. K., J. B. Harris, S. Tabrizi, A. Rahman, I. Shlyakhter, N. Patterson, C. O'Dushlaine, S. F. Schaffner, S. Gupta, F. Chowdhury, A. Sheikh, O. S. Shin, C. Ellis, C. E. Becker, L. M. Stuart, S. B. Calderwood, E. T. Ryan, F. Qadri, P. C. Sabeti and R. C. LaRocque (2013). "Natural Selection in a Bangladeshi Population from the Cholera-Endemic Ganges River Delta." Science Translational Medicine **5**(192): 192ra186-192ra186.

Katz, Y., E. T. Wang, E. M. Airoldi and C. B. Burge (2010). "Analysis and design of RNA sequencing experiments for identifying isoform regulation." Nat Meth **7**(12): 1009-1015.

Kaushik, L. S. and P. D. Puri (1984). "A Study of Maternal and Maternal Interaction Effects in Diallel Crosses." Biometrical Journal **26**(7): 771-777.

Khush, R. S., W. D. Cornwell, J. N. Uram and B. Lemaitre (2002). "A ubiquitin-proteasome pathway represses the Drosophila immune deficiency signaling cascade." Curr Biol **12**(20): 1728-1737.

Kilpinen, H., S. M. Waszak, A. R. Gschwind, S. K. Raghav, R. M. Witwicki, A. Orioli, E. Migliavacca, M. Wiederkehr, M. Gutierrez-Arcelus, N. I. Panousis, A. Yurovsky, T. Lappalainen, L. Romano-Palumbo, A.

Planchon, D. Bielser, J. Bryois, I. Padioleau, G. Udin, S. Thurnheer, D. Hacker, L. J. Core, J. T. Lis, N. Hernandez, A. Reymond, B. Deplancke and E. T. Dermitzakis (2013). "Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription." Science **342**(6159): 744-747.

Kimbrell, D. A. and B. Beutler (2001). "The evolution and genetics of innate immunity." Nature Reviews Genetics **2**(4): 256-267.

King, E. G., C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen, K. W. Broman, A. D. Long and S. J. Macdonald (2012). "Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource." Genome research **22**(8): 1558-1566.

Kirienko, N. V. and D. S. Fay (2010). "SLR-2 and JMJC-1 regulate an evolutionarily conserved stress-response network." EMBO J **29**(4): 727-739.

Klepsatel, P., M. Gáliková, N. De Maio, C. D. Huber, C. Schlötterer and T. Flatt (2013). "VARIATION IN THERMAL PERFORMANCE AND REACTION NORMS AMONG POPULATIONS OF DROSOPHILA MELANOGASTER." Evolution **67**(12): 3573-3587.

Kopp, E. B. and S. Ghosh (1995). "NF-kB and Rel proteins in innate immunity." Advances in immunology **58**(1).

Kornblihtt, A. R., M. De La Mata, J. P. Fededa, M. J. MuÑOz and G. NoguÉS (2004). "Multiple links between transcription and splicing." RNA **10**(10): 1489-1498.

Kumar, A., H. Wu, L. S. Collier-Hyams, J. M. Hansen, T. Li, K. Yamoah, Z.-Q. Pan, D. P. Jones and A. S. Neish (2007). "Commensal bacteria modulate cullin-dependent signaling via generation of reactive oxygen species." The EMBO Journal **26**(21): 4457-4466.

Kuraishi, T., A. Hori and S. Kurata (2013). "Host-microbe interactions in the gut of Drosophila melanogaster." Frontiers in Physiology **4**.

Kwan, T., D. Benovoy, C. Dias, S. Gurd, D. Serre, H. Zuzan, T. A. Clark, A. Schweitzer, M. K. Staples, H. Wang, J. E. Blume, T. J. Hudson, R. Sladek and J. Majewski (2007). "Heritability of alternative splicing in the human genome." Genome Research **17**(8): 1210-1218.

Lazzaro, B., B. Sceurman and A. Clark (2004). "Genetic basis of natural variation in D. melanogaster antibacterial immunity." Science **303**(5665): 1873 - 1876.

Lazzaro, B. P. and T. J. Little (2009). Immunity in a variable world.

Lazzaro, B. P., T. B. Sackton and A. G. Clark (2006). "Genetic Variation in Drosophila melanogaster Resistance to Infection: A Comparison Across Bacteria." Genetics **174**(3): 1539-1554.

Lee, K.-A., S.-H. Kim, E.-K. Kim, E.-M. Ha, H. You, B. Kim, M.-J. Kim, Y. Kwon, J.-H. Ryu and W.-J. Lee (2013). "Bacterial-Derived Uracil as a Modulator of Mucosal Immunity and Gut-Microbe Homeostasis in Drosophila." Cell **153**(4): 797-811.

Lee, M. N., C. Ye, A.-C. Villani, T. Raj, W. Li, T. M. Eisenhaure, S. H. Imboywa, P. I. Chipendo, F. A. Ran, K. Slowikowski, L. D. Ward, K. Raddassi, C. McCabe, M. H. Lee, I. Y. Frohlich, D. A. Hafler, M. Kellis, S.

Raychaudhuri, F. Zhang, B. E. Stranger, C. O. Benoist, P. L. De Jager, A. Regev and N. Hacohen (2014). "Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells." Science **343**(6175).

Lee, S., Gonçalo R. Abecasis, M. Boehnke and X. Lin (2014). "Rare-Variant Association Analysis: Study Designs and Statistical Tests." American Journal of Human Genetics **95**(1): 5-23.

Lehner, B. (2013). "Genotype to phenotype: lessons from model organisms for human genetics." Nat Rev Genet **14**(3): 168-178.

Lemaitre, B. and J. Hoffmann (2007). "The Host Defense of Drosophila melanogaster." Annual Review of Immunology **25**(1): 697-743.

Lemaitre, B. and I. Miguel-Aliaga (2013). "The Digestive Tract of Drosophila melanogaster." Annual Review of Genetics **47**(1): 377-404.

Leoni, G., L. Le Pera, F. Ferrè, D. Raimondo and A. Tramontano (2011). "Coding potential of the products of alternative splicing in human." Genome Biology **12**(1): R9-R9.

Leulier, F., A. Rodriguez, R. S. Khush, J. M. Abrams and B. Lemaitre (2000). "The Drosophila caspase Dredd is required to resist Gram-negative bacterial infection." EMBO reports **1**(4): 353-358.

Levanon, E. Y. and R. Sorek (2003). "The importance of alternative splicing in the drug discovery process." TARGETS **2**(3): 109-114.

Lhocine, N., P. S. Ribeiro, N. Buchon, A. Wepf, R. Wilson, T. Tenev, B. Lemaitre, M. Gstaiger, P. Meier and F. Leulier (2008). "PIMS modulates immune tolerance by negatively regulating Drosophila innate immune signaling." Cell host & microbe **4**(2): 147-158.

Lin, J.-C., M. Hsu and W.-Y. Tarn (2007). "Cell stress modulates the function of splicing regulatory protein RBM4 in translation control." Proceedings of the National Academy of Sciences of the United States of America **104**(7): 2235-2240.

Linder, C. C. (2001). "The influence of genetic background on spontaneous and genetically engineered mouse models of complex diseases." Lab animal **30**(5): 34-39.

Litman, G. W., J. P. Rast and S. D. Fugmann (2010). "The origins of vertebrate adaptive immunity." Nat Rev Immunol **10**(8): 543-553.

Long, A. D., S. J. Macdonald and E. G. King (2014). "Dissecting Complex Traits Using the Drosophila Synthetic Population Resource." Trends in genetics : TIG **30**(11): 488-495.

Lynch, M. and B. Walsh (1998). Genetics and analysis of quantitative traits.

Ma, J. and M. Ptashne (1987). "The carboxy-terminal 30 amino acids of GAL4 are recognized by GAL80." Cell **50**(1): 137-142.

Mackay, T. F. C. and R. F. Lyman (2005). "Drosophila bristles and the nature of quantitative genetic variation." Philosophical Transactions of the Royal Society B: Biological Sciences **360**(1459): 1513-1527.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barron, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ramia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman and R. A. Gibbs (2012). "The Drosophila melanogaster Genetic Reference Panel." Nature **482**(7384): 173-178.

Mackay, T. F. C., E. A. Stone and J. F. Ayroles (2009). "The genetics of quantitative traits: challenges and prospects." Nat Rev Genet **10**(8): 565-577.

Magwire, M. M., D. K. Fabian, H. Schweyen, C. Cao, B. Longdon, F. Bayer and F. M. Jiggins (2012). "Genome-Wide Association Studies Reveal a Simple Genetic Basis of Resistance to Naturally Coevolving Viruses in <italic>Drosophila melanogaster</italic>." PLoS Genet **8**(11): e1003057.

Magwire, M. M., D. K. Fabian, H. Schweyen, C. Cao, B. Longdon, F. Bayer and F. M. Jiggins (2012). "Genome-Wide Association Studies Reveal a Simple Genetic Basis of Resistance to Naturally Coevolving Viruses in Drosophila melanogaster." PLoS Genet **8**(11): e1003057.

Mann, E. A. and S. A. Saeed (2012). "Gastrointestinal infection as a trigger for inflammatory bowel disease." Curr Opin Gastroenterol **28**(1): 24-29.

Manolio , T. A. (2010). "Genomewide Association Studies and Assessment of the Risk of Disease." New England Journal of Medicine **363**(2): 166-176.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll and P. M. Visscher (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

Marianes, A. and A. C. Spradling (2013). "Physiological and stem cell compartmentalization within the Drosophila midgut." eLife **2**: e00886.

Markou, P. and Y. Apidianakis (2013). "Pathogenesis of intestinal Pseudomonas aeruginosa infection in patients with cancer." Frontiers in Cellular and Infection Microbiology **3**: 115.

Martins, N. E., V. G. Faria, L. Teixeira, S. Magalhães and É. Sucena (2013). "Host Adaptation Is Contingent upon the Infection Route Taken by Pathogens." PLoS Pathog **9**(9): e1003601.

Massouras, A., S. M. Waszak, M. Albarca-Aguilera, K. Hens, W. Holcombe, J. F. Ayroles, E. T. Dermitzakis, E. A. Stone, J. D. Jensen, T. F. Mackay and B. Deplancke (2012). "Genomic variation and its impact on gene expression in Drosophila melanogaster." PLoS Genet **8**(11): e1003055.

Mayr, E. (1982). The growth of biological thought : diversity, evolution, and inheritance / Ernst Mayr. Cambridge, Mass, Belknap Press.

McLeay, R. C. and T. L. Bailey (2010). "Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data." BMC Bioinformatics **11**(1): 1-11.

McManus, C. J., J. D. Coolon, J. Eipper-Mains, P. J. Wittkopp and B. R. Graveley (2014). "Evolution of splicing regulatory networks in Drosophila." Genome Research **24**(5): 786-796.

McQuilton, P., S. E. St Pierre and J. Thurmond (2012). "FlyBase 101--the basics of navigating FlyBase." Nucleic Acids Res **40**.

Mendel, G. (1965). Experiments in plant hybridisation : Mendel's original paper in English translation, with commentary and assessment by the late Sir Ronald A. Fisher, together with a reprint of W. Bateson's biographical notice of Mendel / Gregor Mendel ; edited by J.H. Bennett. Edinburgh, Oliver & Boyd.

Merila, J. and B. C. Sheldon (1999). "Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas." Heredity **83**(2): 103-109.

Micchelli, C. A. and N. Perrimon (2006). "Evidence that stem cells reside in the adult Drosophila midgut epithelium." Nature **439**(7075): 475-479.

Monlong, J., M. Calvo, P. G. Ferreira and R. Guigó (2014). "Identification of genetic variants associated with alternative splicing using sQTLseekeR." Nat Commun **5**.

Morgan, T. H. (1911). "THE ORIGIN OF NINE WING MUTATIONS IN DROSOPHILA." Science **33**(848): 496-499.

Morgan, T. H. (1911). "RANDOM SEGREGATION VERSUS COUPLING IN MENDELIAN INHERITANCE." Science **34**(873): 384.

Morgan, T. H. (1915). The mechanism of Mendelian heredity. New York, Holt.

Muszynski, J. A., R. Nofziger, K. Greathouse, J. Nateri, L. Hanson-Huber, L. Steele, K. Nicol, J. I. Groner, G. E. Besner, C. Raffel, S. Geyer, O. El-Assal and M. W. Hall (2014). "Innate immune function predicts the development of nosocomial infection in critically injured children." Shock **42**(4): 313-321.

Nebert, D. W., G. Zhang and E. S. Vesell (2008). "From Human Genetics and Genomics to Pharmacogenetics and Pharmacogenomics: Past Lessons, Future Directions." Drug metabolism reviews **40**(2): 187-224.

Newby, L. M. and F. R. Jackson (1996). "Regulation of a specific circadian clock output pathway by lark, a putative RNA-binding protein with repressor activity." Journal of Neurobiology **31**(1): 117-128.

Nica, A. C. and E. T. Dermitzakis (2013). "Expression quantitative trait loci: present and future." Philosophical Transactions of the Royal Society B: Biological Sciences **368**(1620): 20120362.

Nojima, T., T. Gomes, Ana Rita F. Grosso, H. Kimura, Michael J. Dye, S. Dhir, M. Carmo-Fonseca and Nicholas J. Proudfoot "Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing." Cell **161**(3): 526-540.

O'Brien, L. E., S. S. Soliman, X. Li and D. Bilder (2011). "Altered modes of stem cell division drive adaptive intestinal growth." Cell **147**(3): 603-614.

Obbard, D. J., J. J. Welch, K.-W. Kim and F. M. Jiggins (2009). "Quantifying Adaptive Evolution in the Drosophila Immune System." PLoS Genet **5**(10): e1000698.

Ohlstein, B. and A. Spradling (2006). "The adult Drosophila posterior midgut is maintained by pluripotent stem cells." Nature **439**(7075): 470-474.

Ohlstein, B. and A. Spradling (2007). "Multipotent Drosophila intestinal stem cells specify daughter cell fates by differential notch signaling." Science **315**(5814): 988-992.

Opota, O., I. Vallet-Gely, R. Vincentelli, C. Kellenberger, I. Iacovache, M. R. Gonzalez, A. Roussel, F. G. van der Goot and B. Lemaitre (2011). "Monalysin, a novel ss-pore-forming toxin from the Drosophila pathogen Pseudomonas entomophila, contributes to host intestinal damage and lethality." PLoS Pathog **7**(9): e1002259.

Orozco, Luz D., Brian J. Bennett, Charles R. Farber, A. Ghazalpour, C. Pan, N. Che, P. Wen, Hong X. Qi, A. Mutukulu, N. Siemers, I. Neuhaus, R. Yordanova, P. Gargalovic, M. Pellegrini, T. Kirchgessner and Aldons J. Lusis "Unraveling Inflammatory Responses using Systems Genetics and Gene-Environment Interactions in Macrophages." Cell **151**(3): 658-670.

Overend, G., P. Cabrero, A. X. Guo, S. Sebastian, M. Cundall, H. Armstrong, I. Mertens, L. Schoofs, J. A. Dow and S. A. Davies (2012). "The receptor guanylate cyclase Gyc76C and a peptide ligand, NPLP1-VQQ, modulate the innate immune IMD pathway in response to salt stress." Peptides **34**(1): 209-218.

Ozsolak, F. and P. M. Milos (2011). "RNA sequencing: advances, challenges and opportunities." Nat Rev Genet **12**(2): 87-98.

Paredes, J. C., D. P. Welchman, M. Poidevin and B. Lemaitre (2011). "Negative regulation by amidase PGRPs shapes the Drosophila antibacterial response and protects the fly from innocuous infection." Immunity **35**(5): 770-779.

Passador-Gurgel, G., W.-P. Hsieh, P. Hunt, N. Deighton and G. Gibson (2007). "Quantitative trait transcripts for nicotine resistance in Drosophila melanogaster." Nat Genet **39**(2): 264-268.

Punnett, R. C. (1911). Mendelism. New York, The Macmillan Company.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, Manuel A R. Ferreira, D. Bender, J. Maller, P. Sklar, Paul I W. de Bakker, Mark J. Daly and Pak C. Sham (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." American Journal of Human Genetics **81**(3): 559-575.

Ray, D., H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris and T. R. Hughes (2013). "A compendium of RNA-binding motifs for decoding gene regulation." Nature **499**(7457): 172-177.

Reed, R. (2003). "Coupling transcription, splicing and mRNA export." Current Opinion in Cell Biology **15**(3): 326-331.

Reiter, L. T., L. Potocki, S. Chien, M. Gribskov and E. Bier (2001). "A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster." Genome Res **11**(6): 1114-1125.

Rockel, S., M. Geertz and S. J. Maerkl (2012). "MITOMI: A Microfluidic Platform for In Vitro Characterization of Transcription Factor–DNA Interaction." Gene Regulatory Networks: Methods and Protocols: 97-114.

Rodríguez, A. d. V., D. Didiano and C. Desplan (2011). "Power tools for gene expression and clonal analysis in Drosophila." Nature methods **9**(1): 47-55.

Rushton, A. R. (2000). "Nettleship, Pearson and Bateson: The Biometric-Mendelian Debate in a Medical Context." Journal of the History of Medicine and Allied Sciences **55**(2): 134-157.

Ryu, J.-H., S.-H. Kim, H.-Y. Lee, J. Y. Bai, Y.-D. Nam, J.-W. Bae, D. G. Lee, S. C. Shin, E.-M. Ha and W.-J. Lee (2008). "Innate Immune Homeostasis by the Homeobox Gene Caudal and Commensal-Gut Mutualism in Drosophila." Science **319**(5864): 777-782.

Sax, K. (1923). "The Association of Size Differences with Seed-Coat Pattern and Pigmentation in PHASEOLUS VULGARIS." Genetics **8**(6): 552-560.

Schaeffer, L. R. (2006). "Strategy for applying genome-wide selection in dairy cattle." Journal of Animal Breeding and Genetics **123**(4): 218-223.

Schatz, D. G. (2004). "V (d) j recombination." Immunological reviews **200**(1): 5-11.

Schieber, M. and Navdeep S. Chandel "ROS Function in Redox Signaling and Oxidative Stress." Current Biology **24**(10): R453-R462.

Schneider, D. S. and J. S. Ayres (2008). "Two ways to survive infection: what resistance and tolerance can teach us about treating infectious diseases." Nat Rev Immunol **8**(11): 889-895.

Seidel, G. E. (2009). "Brief introduction to whole-genome selection in cattle using single nucleotide polymorphisms." Reproduction, Fertility and Development **22**(1): 138-144.

Shabalin, A. A. (2012). "Matrix eQTL: ultra fast eQTL analysis via large matrix operations." Bioinformatics **28**(10): 1353-1358.

Shalgi, R., Jessica A. Hurt, S. Lindquist and Christopher B. Burge (2014). "Widespread Inhibition of Posttranscriptional Splicing Shapes the Cellular Transcriptome following Heat Shock." Cell Reports **7**(5): 1362-1370.

Shazman, S., H. Lee, Y. Socol, R. S. Mann and B. Honig (2013). "OnTheFly: a database of Drosophila melanogaster transcription factors and their binding sites." Nucleic Acids Research.

Silventoinen, K., P. K. E. Magnusson, P. Tynelius, J. Kaprio and F. Rasmussen (2008). "Heritability of body size and muscle strength in young adulthood: a study of one million Swedish men." Genetic Epidemiology **32**(4): 341-349.

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. R. Gentleman, V. Carey, W. Huber, R. Irizarry and S. Dudoit, Springer New York**:** 397-420.

St Johnston, D. (2002). "The art and design of genetic screens: Drosophila melanogaster." Nat Rev Genet **3**(3): 176-188.

St Pierre, S. E., L. Ponting, R. Stefancsik and P. McQuilton (2014). "FlyBase 102--advanced approaches to interrogating FlyBase." Nucleic Acids Res **42**.

Stainier, D. Y. R. (2005). "No Organ Left Behind: Tales of Gut Development and Evolution." Science **307**(5717): 1902-1904.

Sterne-Weiler, T., R. T. Martinez-Nunez, J. M. Howard, I. Cvitovik, S. Katzman, M. A. Tariq, N. Pourmand and J. R. Sanford (2013). "Frac-seq reveals isoform-specific recruitment to polyribosomes." Genome Research **23**(10): 1615-1623.

Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles and E. T. Dermitzakis (2007). "Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes." Science **315**(5813): 848-853.

Sturtevant, A. H. (1913). "The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association." Journal of Experimental Zoology **14**(1): 43-59.

Supek, F., M. Bo?njak, N. kunca and T. muc (2011). "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms." PLoS ONE **6**(7): e21800.

Swami, M. (2010). "Complex traits: Using genetic architecture to improve predictions." Nat Rev Genet **11**(11): 748-748.

Takeuchi, O. and S. Akira (2010). "Pattern recognition receptors and inflammation." Cell **140**.

The GTEx Consortium (2013). "The Genotype-Tissue Expression (GTEx) project." Nature genetics **45**(6): 580-585.

Thevenon, D., E. Engel, A. Avet-Rochex, M. Gottar, E. Bergeret, H. Tricoire, C. Benaud, J. Baudier, E. Taillebourg and M. O. Fauvarque (2009). "The Drosophila ubiquitin-specific protease dUSP36/Scny targets IMD to prevent constitutive immune signaling." Cell Host Microbe **6**(4): 309-320.

Thoday, J. M. (1961). "Location of Polygenes." Nature **191**(4786): 368-370.

Tinsley, M. C., S. Blanford and F. M. Jiggins (2006). "Genetic variation in Drosophila melanogaster pathogen susceptibility." Parasitology **132**(06): 767-773.

Tsuda, M., R. Ootaka, C. Ohkura, Y. Kishita, K. H. Seong, T. Matsuo and T. Aigaki (2010). "Loss of Trx-2 enhances oxidative stress-dependent phenotypes in Drosophila." FEBS Lett **584**(15): 3398-3401.

Turro, E., S.-Y. Su, Â. Gonçalves, L. J. Coin, S. Richardson and A. Lewin (2011). "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." Genome Biology **12**(2): 1-15.

Tzou, P., S. Ohresser, D. Ferrandon, M. Capovilla, J.-M. Reichhart, B. Lemaitre, J. A. Hoffmann and J.-L. Imler (2000). "Tissue-Specific Inducible Expression of Antimicrobial Peptide Genes in Drosophila Surface Epithelia." Immunity **13**(5): 737-748.

Vattem, K. M. and R. C. Wek (2004). "Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells." Proceedings of the National Academy of Sciences of the United States of America **101**(31): 11269-11274.

Veenstra, J. A., H.-J. Agricola and A. Sellami (2008). "Regulatory peptides in fruit fly midgut." Cell and Tissue Research **334**(3): 499-516.

Visscher, P. M. (2008). "Sizing up human height variation." Nat Genet **40**(5): 489-490.

Visscher, Peter M., S. Macgregor, B. Benyamin, G. Zhu, S. Gordon, S. Medland, William G. Hill, J.-J. Hottenga, G. Willemsen, Dorret I. Boomsma, Y.-Z. Liu, H.-W. Deng, Grant W. Montgomery and Nicholas G. Martin (2007). "Genome Partitioning of Genetic Variation for Height from 11,214 Sibling Pairs." American Journal of Human Genetics **81**(5): 1104-1110.

Vodovar, N., M. Vinals, P. Liehl, A. Basset, J. Degrouard, P. Spellman, F. Boccard and B. Lemaitre (2005). "Drosophila host defense after oral infection by an entomopathogenic Pseudomonas species." Proceedings of the National Academy of Sciences of the United States of America **102**(32): 11414-11419.

Waern, K. and M. Snyder (2013). "Extensive Transcript Diversity and Novel Upstream Open Reading Frame Regulation in Yeast." G3: Genes|Genomes|Genetics **3**(2): 343-352.

Watatani, Y., K. Ichikawa, N. Nakanishi, M. Fujimoto, H. Takeda, N. Kimura, H. Hirose, S. Takahashi and Y. Takahashi (2008). "Stress-induced Translation of ATF5 mRNA Is Regulated by the 5'-Untranslated Region." Journal of Biological Chemistry **283**(5): 2543-2553.

Weber, A. L., G. F. Khan, M. M. Magwire, C. L. Tabor, T. F. C. Mackay and R. R. H. Anholt (2012). "Genome-Wide Association Analysis of Oxidative Stress Resistance in <italic>Drosophila melanogaster</italic>." PLoS ONE **7**(4): e34745.

Wethmar, K. (2014). "The regulatory potential of upstream open reading frames in eukaryotic gene expression." Wiley Interdisciplinary Reviews: RNA **5**(6): 765-768.

Wolfer, D. P., W. E. Crusio and H.-P. Lipp (2002). "Knockout mice: simple solutions to the problems of genetic background and flanking genes." Trends in Neurosciences **25**(7): 336-340.

Woolhouse, M. E. J., J. P. Webster, E. Domingo, B. Charlesworth and B. R. Levin (2002). "Biological and biomedical implications of the co-evolution of pathogens and their hosts." Nat Genet **32**(4): 569-577.

Wright, S. (1921). "Systems of mating. I. The biometric relations between parent and offspring." Genetics **6**(2): 111.

Yagi, Y., Y. M. Lim, L. Tsuda and Y. Nishida (2013). "fat facets induces polyubiquitination of Imd and inhibits the innate immune response in Drosophila." Genes Cells **18**(11): 934-945.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard and P. M. Visscher (2010). "Common SNPs explain a large proportion of heritability for human height." Nature genetics **42**(7): 565-569.

Yost, H. J. and S. Lindquist (1986). "RNA splicing is interrupted by heat shock and is rescued by heat shock protein synthesis." Cell **45**(2): 185-193.

Zaitlen, N., B. Pasaniuc, S. Sankararaman, G. Bhatia, J. Zhang, A. Gusev, T. Young, A. Tandon, S. Pollack, B. J. Vilhjálmsson, T. L. Assimes, S. I. Berndt, W. J. Blot, S. Chanock, N. Franceschini, P. G. Goodman, J. He, A. J. M. Hennis, A. Hsing, S. A. Ingles, W. Isaacs, R. A. Kittles, E. A. Klein, L. A. Lange, B. Nemesure, N. Patterson, D. Reich, B. A. Rybicki, J. L. Stanford, V. L. Stevens, S. S. Strom, E. A. Whitsel, J. S. Witte, J. Xu, C. Haiman, J. G. Wilson, C. Kooperberg, D. Stram, A. P. Reiner, H. Tang and A. L. Price (2014). "Leveraging population admixture to explain missing heritability of complex traits." Nature genetics **46**(12): 1356-1362.

Zasloff, M. (2002). "Antimicrobial peptides of multicellular organisms." nature **415**(6870): 389-395.

Zhang, G. (2015). "Genetic Architecture of Complex Human Traits: What Have We Learned from Genome-Wide Association Studies?" Current Genetic Medicine Reports **3**(4): 143-150.

Zhang, X., R. Joehanes, B. H. Chen, T. Huan, S. Ying, P. J. Munson, A. D. Johnson, D. Levy and C. J. O'Donnell (2015). "Identification of common genetic variants controlling transcript isoform variation in human whole blood." Nat Genet **47**(4): 345-352.

Zhou, R., N. Silverman, M. Hong, D. S. Liao, Y. Chung, Z. J. Chen and T. Maniatis (2005). "The Role of Ubiquitination in Drosophila Innate Immunity." Journal of Biological Chemistry **280**(40): 34048-34055.

Zichner, T., D. A. Garfield, T. Rausch, A. M. Stütz, E. Cannavó, M. Braun, E. E. M. Furlong and J. O. Korbel (2013). "Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing." Genome Research **23**(3): 568-579.

Zuk, O., E. Hechter, S. R. Sunyaev and E. S. Lander (2012). "The mystery of missing heritability: Genetic interactions create phantom heritability." Proceedings of the National Academy of Sciences **109**(4): 1193-1198.

Curriculum Vitae – Maroun Bou Sleiman

CONTACT INFORMATION

Address : Chemin de Renens 41, 1004 Lausanne, Switzerland
Email : marounbs@gmail.com
Tel : 0041762033215

PERSONAL INFORMATION

Nationality: Lebanese
Date and place of birth: 03 February 1988, Lebanon

EDUCATION

- **Saint Joseph School, Cornet Chehwan, Lebanon** (1991-2005)
  - Baccaleaureate in Life Sciences (2005)
- **American University of Beirut, Lebanon** (2006-2012)
  - Bachelor of Science in Biology (2006-2009)
  - Masters of Science in Biology (2009-2012)
    - Masters thesis project performed in the Laboratory of Prof. Bruno Lemaitre at EPFL.
    - Thesis advisor: Prof. Colin Andrew Smith
    - Thesis title: Identifying polymorphisms underlying the ability to survive infection: a quantitative genetics and evolutionary approach in drosophila melanogaster.
- **Ecole Polytechnique Fédérale de Lausanne** (2012-present)
  - PhD student in Molecular Life Sciences.
  - Co-advised by Prof. Bruno Lemaitre and Prof. Bart Deplancke.
  - Thesis title: The Drosophila gut response to infection: a systems approach

PUBLICATIONS

Possik EJ, Bou Sleiman MS, Ghattas IR, Smith CA. Randomized codon mutagenesis reveals that the HIV Rev arginine-rich motif is robust to substitutions and that double substitution of two critical residues alters specificity. *Journal of Molecular Recognition* **26**, 286-296 (2013).

Chng W-bin A, Bou Sleiman MS, Schüpfer F, Lemaitre B. Transforming Growth Factor β/Activin Signaling Functions as a Sugar-Sensing Feedback Loop to Regulate Digestive Enzyme Expression. *Cell Reports* **9**, 336-348 (2014).

Bou Sleiman MS, Osman D, Massouras A, Hoffmann AA, Lemaitre B, Deplancke B. Genetic, molecular and physiological basis of variation in Drosophila gut immunocompetence. *Nat Commun* **6**, (2015).

Neyen C, Binggeli O, Roversi P, Bertin L, Bou Sleiman MS, Lemaitre B. The Black cells phenotype is caused by a point mutation in the Drosophila pro-phenoloxidase 1 gene that triggers melanization and hematopoietic defects1. *Developmental & Comparative Immunology* **50**, 166-174 (2015).

SELECTED CONFERENCE TALKS

Bou Sleiman *et al.* (2013): Understanding the Basis of Natural Variation in Response to Intestinal Infection in a *Drosophila* Population. In: The European Drosophila Research Conference, Barcelona, Spain.

Bou Sleiman *et al.*, (2016): The Genomics of Drosophila Gut Immunocompetence Variation. In: The European Drosophila Research Conference, Heidelberg, Germany.

Bou Sleiman et al., (2016): The Genomics of Drosophila Gut Immunocompetence Variation. In: The Lausanne Fly Meeting, Lausanne, Switzerland.