ORTHOGONAL APPROACHES FOR SURVEYING GENETIC VARIATION AND ITS CONSEQUENCES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Maria Florencia Schlamp

August 2019

ORTHOGONAL APPROACHES FOR SURVEYING GENETIC VARIATION AND ITS CONSEQUENCES

Maria Florencia Schlamp, Ph.D.

Cornell University 2019

Our morphological traits, responses to stimuli, and the composition of our microbiomes are all phenotypic adaptations influenced by the genetic variation that defines us. Understanding this multimodal network of relationships requires the analysis of a multitude of orthogonal biological systems. Tailoring our approach to the individual biological outputs and systems allows us to reach a deeper understanding of the evolution, regulation, and interactions among biological processes.

When available, we can use genomic data from large populations to establish links between genetic variation and phenotypic adaptation. For instance, positive selection can be inferred from variation computationally and statistically via evidence of selective sweeps. In **Chapter 2**, I evaluate eight selection scans to detect selective sweeps in domestic dogs, a population with well-documented selection pressures imposed by human preferences for specific morphologies and other traits.

Pathogen-driven selective pressures modulate adaptation in the immune response, because hosts must keep up in the host-pathogen arms race. The high energetic cost of mounting an immune response reduces resource availability to other

physiological processes. To explore these trade-offs, in **Chapter 3** I profile the transcription dynamics of the *Drosophila melanogaster* innate immune response in a dense time course and I apply a broad range of statistical methods, including temporal clustering, gene set expression analysis, and Granger causality to construct putative gene interactions networks.

The interaction of hosts with mutualistic symbionts can drive genetic adaptation in hosts through mutually-beneficial processes. In humans, the gut microbiome provides a wealth of symbiotic interactions. To address whether this mutualistic relationship drives host adaptation, in **Chapter 4** I study the influence of host genetics on microbiome composition by performing high-resolution QTL mapping to identify genetic variation in Diversity Outbred mice significantly associated with specific bacterial abundances.

This thesis presents three orthogonal approaches for surveying genetic variation and its consequences, using a combination of data collected through three sequencing methods: population genomic data using genotyping, global transcriptome dynamics using RNA-sequencing, and microbiome composition using 16S rRNA gene sequencing.

BIOGRAPHICAL SKETCH

Florencia Schlamp was born in a small town in Argentina in 1990. At the early age of 8, Florencia would already tell her siblings that her favorite subject was Natural Sciences. Back then, little did she know, that 20 years later, she would be completing her Doctoral studies in Genomics and Computational Biology at Cornell University.

Her interest in Biology strengthened when she moved to Brazil as a teenager and was able to develop her love for SCUBA diving. Her love for science shaped while taking courses in High School, grew to the extent that she decided to pursue her college studies in Biology. In 2010 she joined the inaugural class of New York University Abu Dhabi, which would change her life both as a person and as a future scientist. Florencia continued to develop her research skills under the supervision of Dr. John Burt and Dr. Michael Purugganan.

While in NYUAD, she met amazing faculty, staff, and students. She met lifelong friends and her life partner, Adam Dolan. Together they decided to pursue Graduate studies at Cornell University. For the following years, under the supervision of Dr. Andrew Clark, Florencia was fortunate to pursue multidisciplinary and diverse research projects, while also developing teaching and mentoring skills.

Looking to the future, Florencia is passionate about research, and excited to try opportunities in academia, where she can make use of her passion for sharing knowledge and teaching science.

To my family and friends, and everyone who believed in me…

ACKNOWLEDGMENTS

First and foremost, I want to thank my parents and siblings, whose support and love have built me up to who I am. My mom and dad have been an amazing influence on my life. Their examples are always with me and will continue to shape who I hope to be in the future. This achievement belongs to them, too. To my siblings, Sofia and Guillermo, thank you for your encouragement and love. I am so very proud of you both and I cannot wait to see what we accomplish in the future.

I would like to thank Andy for being my PI and providing me with the most amazing group of people he managed to get together in his lab. I appreciate the patience and encouragement he has shown me and I will always remember the sailing trips we took. To Philipp I extend my gratitude for providing guidance, particularly during the early stages of my exploration into population genetics and computational biology. Mariana has been a source of great advice and I am thankful to have her on my committee. The collaboration I have had with Sumanta has been extraordinarily productive and I know that my time-course project would not have turned out as well without his involvement. I want to particularly thank him for his help, teachings, and advice.

I want to thank the beautiful humans and scientists in the Clark lab for making it a joy to work there every day. I have found lasting friendships, sharp minds, and kind hearts there. I cannot mention everyone with descriptions of how much they mean to me, so an incomplete list of the great people from the Clark Lab will have to suffice: Manisha, Sri, Elissa, Andrea, Emily, Yasir, Arvid, Ian, Tram, Andrew, Amanda, Lori.

Thank you to Cornell University and in particular the GGD Field, for providing me with an amazing community whose support and friendship have been vital these five years. Again, I am faced with the task of making an incomplete list, but I must mention a few: Roman, Dan, Zach, Mike, Ian, Afrah – Thank you for everything.

Thank you to all my students, mentees, and undergraduate interns I have supervised, but in particular to David, whose hard work and continuous desire to learn served as a constant reminder of why I love science, research, and teaching so much.

Thank you to my dearest lifelong friends I made at NYUAD, especially those who kept me company with DnD over the internet. I am so grateful to have you all as a continuing part of my life. Special thanks to Juan Felipe for all his help and support. I am so thankful for the privilege of having you around both at NYUAD and Cornell.

And thank you Veronica for your unconditional friendship and support across time and distance.

Above all, thank you to my amazing life partner, Adam. I am so proud of what we have accomplished, and I am excited to see what comes next. And thank you to my amazing cat, Numi, for all the cuddles.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

BIOLOGICAL MOTIVATIONS

A common pursuit in evolutionary biology is to understand the causal link between genetic variation and phenotypic adaptation. The development of increasingly swift and cost-effective sequencing technologies has spurred explosive growth in data, analytical tools, and population genetics theory that can be used to characterize these genotype-phenotype relationships. Despite this growth, identifying the genetic loci responsible for a given phenotype is far from a solved problem. The detection of selection within a population can be performed with a variety of tools whose performance depends on the time window when selection occurred, its strength, the population demographics, and the polygenicity of the selected trait (VITTI *et al.* 2013; WEIGAND AND LEESE 2018). Additionally, "phenotypes" include a staggering variety of data types including -but not limited to- trait categories, the magnitude of response to stimuli over time, and the composition of symbiotic species within an organism. Tailoring our approach to these different biological outputs allows us to reach a deeper understanding of the evolution, regulation, and interactions among biological processes.

When genomic data from a large population are available, we can establish links between genetic variation and phenotypic adaptation by measuring patterns of variation among individuals. For instance, positive selection can be inferred using computational and statistical methods (VITTI *et al.* 2013; WEIGAND AND LEESE 2018) by looking for evidence of selective sweeps: regions of the genome with reduced variation in the nucleotides neighboring a common mutation. These signatures suggest the presence of a beneficial mutation, which has been driven to a higher frequency by natural selection. In **Chapter 2** of this thesis, I use dogs as a model system to characterize the performance of eight different methods for detecting selection scans. Dogs are especially useful for this due to the presence of known selective pressures on morphology and behavior imposed by human preference.

Selection, however, does not always lead to a reduction in genetic variation, but can sometimes increase it instead. This is the case in host-pathogen interactions, a very interesting instance of phenotypic adaptation that benefits from persistent genetic variation. In this system, balancing selection leads genomic regions that drive immune response processes to be maintained at high genetic variation in a population, instead of allowing one particular allele to reach fixation (CROZE *et al.* 2016). The over-homogenization of a species can cripple its ability to respond to change, as was the case with the clonal banana cultivar "Gros Michel" which experienced population collapse due to shared susceptibility to a pathogen infection the 1950's (PLOETZ 1994). Thus, pathogen-driven selective pressures modulate the process of adaptation

in the immune response, where hosts constantly need to keep up in the host-pathogen arms race.

The link between genetic variation and fitness in host-pathogen interactions is further complicated by the presence of trade-offs within the host. Due to the high energetic cost of mounting an immune response (LAZZARO AND GALAC 2006), allocating resources to the immune system reduces resource availability to other life processes (ZEROFSKY *et al.* 2005; DIANGELO *et al.* 2009). Therefore, organisms must tune their immune responses to be effective, while also balancing resource trade-offs with other biological processes. This tuning is likely to be mediated through a series of regulatory and feedback circuits in the immune system, which are yet to be fully understood. To elucidate this, in **Chapter 3** of this thesis I profile the transcription dynamics of the *Drosophila melanogaster* innate immune response using a dense gene expression time course. With this work, I unveil distinct temporal patterns of transient and sustained responses to infection that occur over different time scales, I provide several novel functional annotations for previously uncharacterized genes, and suggest new interactions governing temporal gene regulation of the immune response and trade-offs with metabolism and repair.

The host-pathogen arms race is not the only type of host-microbe coevolution we can observe in nature. The interaction of hosts with mutualistic symbionts can drive genetic adaptation in hosts through a mutually beneficial process instead of one driven by conflict (SHAPIRA 2016). Some well-known examples of how mutualistic

symbionts affect host evolution include squids developing a light organ to host bioluminescent bacteria, and aphids - sap-feeding insects - developing a vertical transmission of bacteria that break down and provide nutrients otherwise inaccessible to aphids (SHAPIRA 2016). In humans, the gut microbiome provides a wealth of symbiotic interactions, and how or whether this relationship is driving host genomic adaptation is currently under study. To address this, in **Chapter 4** of this thesis I study the influence of host genetics on gut microbiome composition using the Diversity Outbred mouse panel, a population designed to be the most genetically diverse mouse resource currently available (CHURCHILL *et al.* 2012). In this work, I performed a high-resolution QTL mapping that identifies genetic variation in Diversity Outbred mice significantly associated with specific bacterial taxon abundances.

In this introduction I discuss three orthogonal approaches for surveying genetic variation and its consequences. I use a combination of data collected through three sequencing methods: population genomic data using genotyping, global transcriptome dynamics using RNA-sequencing, and microbiome composition using 16S rRNA sequencing.

INFERRING REGIONS OF POSITIVE SELECTION IN POPULATION GENOMIC DATA

The appearance of new mutations, the selective forces of natural selection, and the stochastic effects of genetic drift, all contribute to drive allele frequency changes in populations across generations. Knowing how these allele frequencies are changing throughout time can allow us to infer how evolution works by identifying the parts of the genome under selection. These changes could be directly measured if we were to continually sequence the genomes of a population over generational timescales and observe the resulting changes in allele frequency brought on by selective pressures (MALASPINAS *et al.* 2012; BANK *et al.* 2014; FOLL *et al.* 2015). And while some work has already been done in a similar fashion in organisms with short generation times (yeast (KRYAZHIMSKIY *et al.* 2014), bacteria (GOOD *et al.* 2017), *Daphnia* (ZBINDEN *et al.* 2008), *Drosophila* (HOULE *et al.* 2017)), it is not yet feasible to easily achieve on organisms with longer generation times. Fortunately, genomes accumulate signatures of their evolutionary history, which allow us to infer what happened in the past by reverse engineering the process of natural selection from these signatures found in genome sequence variation in extant population samples.

Among the most drastic and recognizable signatures of positive selection are hard selective sweeps. Here, a single acquired beneficial mutation with strong selective advantage on a population will quickly increase in frequency over time (HERMISSON AND PENNINGS 2005) possibly reaching fixation, which is when a specific variant is

shared by the entire population (**Figure 1.1**). As this happens, the regions linked to this beneficial mutation will 'hitchhike', also increasing in frequency alongside the new beneficial mutation. Strong selective sweeps can occur too fast for recombination to break down the statistical association (linkage disequilibrium) across neighboring alleles, thus leaving characteristic signatures of haplotype homozygosity around the selected region. However, positive selection will not always produce the clean signature of a hard selective sweep. Instead, a second beneficial mutation could be established before the first one reaches fixation (**Figure 1.1**) (WILSON *et al.* 2014). Alternatively, adaptation could occur from multiple alleles already present in the population as standing genetic variation. In these cases, beneficial alleles will be found in different haplotype backgrounds which will all increase in frequency, leaving a signature known as soft selective sweeps.

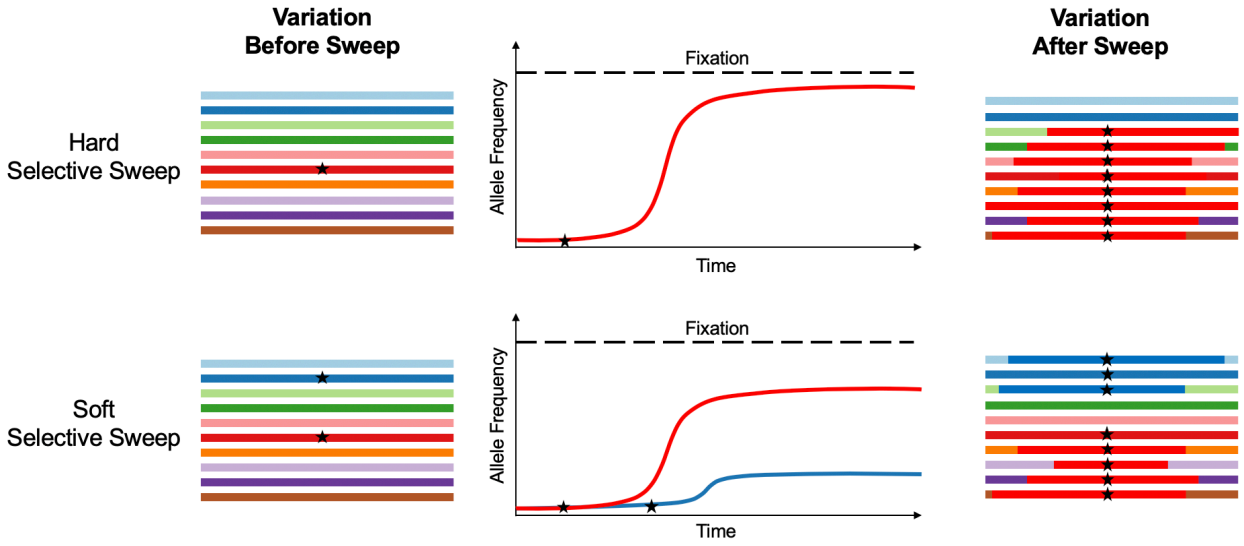


Figure 1.1. Hard versus soft selective sweeps. Variation signatures seen among population haplotypes (on the right) can help us distinguish between hard and soft selective sweeps.

These selective sweep signatures can be detected by directly comparing the genomic regions of individuals within a population and between populations. This kind of population genomic data can be obtained by genotyping or sequencing whole genomes of multiple individuals in a population. Whole genome sequencing provides the highest resolution of genomic data, but it can still be costly when sampling large population numbers. A more cost effective alternative is to only measure the genetic variations of single nucleotide polymorphisms (SNPs) with a method called SNP genotyping, where hundreds of thousands of probes within a chip array hybridize multiple SNPs at the same time. This is a particularly effective method of surveying genomes that are already well characterized. SNP chip genotyping facilitates the selection of the most informative SNPs within a population, phasing the resulting data, and calculating the frequency of alleles in the surveyed population.

Once population genomic data is collected, there are multiple data analysis methods that aim to detect regions under selection based on patterns observed in the genetic variation. These methods are based on the distribution or spectrum of allele frequencies, the lengths and frequencies of shared haplotypes, or runs of identity-by-descent. Allele frequency based methods (such as π, Tajima's D (TAJIMA 1989), and CLR (NIELSEN *et al.* 2005; PAVLIDIS *et al.* 2013)) analyze the distortion in site frequency spectra, where certain SNP frequencies are lower or higher than expected under a neutral model. On the other hand, haplotype based statistics (such as iHS (VOIGHT *et al.* 2006), nSL (FERRER-ADMETLLA *et al.* 2014), H12 (GARUD *et al.* 2015),

H (SCHLAMP *et al.* 2016), EHH, and IBD (CAI *et al.* 2011; HAN AND ABNEY 2013)) search for levels of haplotype homozygosity that are much more elevated than expected under neutrality. Furthermore, cross-population methods (such as F_{ST} (HOLSINGER AND WEIR 2009), XP-EHH (SABETI *et al.* 2007), and HapFLK (FARIELLO *et al.* 2013)) extend their analysis by using differences in allele frequency between populations, which allows the detection of other types of selection, such as negative selection (VITTI *et al.* 2013). Finally, there are composite methods (such as CLR (NIELSEN *et al.* 2005; PAVLIDIS *et al.* 2013), XP-CLR (CHEN *et al.* 2010), and CMS (GROSSMAN *et al.* 2010; GROSSMAN *et al.* 2013)) which combine multiple test scores to improve resolution and power, and reduce false discovery rate. All these methods have different approaches to finding selective sweeps, and as such they have varying degrees of success, depending on the data type, parameters, and assumptions. Therefore, it is of high interest to properly evaluate and benchmark these methods using positive controls from real population data.

Evaluating the performance of selection scans to detect selective sweeps in domestic dogs

The domestic dog is a great system to study genomic signatures of positive selection, because they provide a large number of distinct and well defined breeds that have been artificially selected throughout centuries for very different physical traits, such as body size, coat color, and skull shape; as well as behavioral traits, such as

obedience, herding, and hunting (FREEDMAN *et al.* 2016). This provides well characterized positive control loci where we already know the causal mutations which are under specific selective pressures through known breeding strategies. For example, 46-52.5% of body size variance in dog breeds can be explained by variants at just six genes (RIMBAULT *et al.* 2013), and most coat phenotypes in purebred dogs in the United States are governed by variants in three genes (CADIEU *et al.* 2009). Multiple additional variants have been identified to affect specific morphological variations in dogs, such as lips and ears (BOYKO *et al.* 2010) and skull shape (SCHOENEBECK AND OSTRANDER 2013).

In **Chapter 2**, I present the evaluation of multiple selection scans' performance to detect selective sweeps in domestic dog populations. I developed a custom pipeline that integrates eight statistics (HapFLK, iHS, nSL, H, H12, CLR, Tajima's D, and π) in order to detect signatures of soft and hard sweeps, and used it to confirm positive selection in 12 positive control loci already known to have experienced positive selection in specific dog breeds due to their association to desirable morphological phenotypes. This work successfully detected signature patterns of haplotype and nucleotide polymorphism left by artificial selection during dog domestication, and demonstrated the power and limitations of different selection scans and choice of analysis parameters used. Since then, I have adapted this pipeline to detect regions of positive selection in other genomic population datasets such as Persian Arabian and Iranian horse breeds (SADEGHI *et al.* 2018).

PROFILING TRANSCRIPTION DYNAMICS USING RNA SEQUENCING TIME SERIES

The genome encodes all the information needed to build and maintain an organism throughout its life. This information, however, is not all needed at the same time, so while the genome is a rather static source of genetic information on the individual level, not all of it is being transcribed and used at any given time. Certain genes and regulatory elements play vital and transient roles during different developmental stages or in response to environmental stimuli such as infections, mating, and temperature changes. Likewise, certain genes and regulatory elements play roles in only certain cells or tissues. While the genome is the same in all cells, many genes are only transcriptionally active in certain cell types, contributing to cell differentiation and specialized functions. Many genes need to be transcribed from the genome only at a particular moment or at a particular place. Thus, proper positive and negative regulation of transcription is vital in many processes such as during development, where if genes continue to be expressed after they are not needed anymore it can lead to developmental malfunctions.

Technologies such as microarrays and RNA sequencing (RNA-seq) allow us to get a snapshot of what the transcriptional landscape looks like at any given time and in a specific tissue in an individual organism. These technologies measure the type and abundance of mRNAs present in the organism at the moment of sample collection. Thanks to this, we can get a much clearer picture of how transcripts correlated with

certain processes, such as cell differentiation and reproduction; or reacting to external stimuli, such as a drug treatment or pathogen infection. Through the study of transcriptomics, we have been able to annotate the function of genes and transcription factors, as well as elucidate biological processes, and determine pathways for many diseases.

Nonetheless, most biological processes are dynamic and a single snapshot of the process only shows a small part of the picture. In order to properly study and analyze dynamic processes over time, it is key to sample transcription in a time-series manner. Getting multiple snapshots throughout the process allows us to characterize the dynamics of gene expression, whether responses are transient or sustained, and determine the slope and kinetics of expression change. As sequencing technologies have increased accuracy and decreased costs, collecting time series expression data has become considerably more feasible, and multiple studies have already started characterizing the dynamics of development (GEIJER *et al.* 2012; BATUT AND GINGERAS 2017; WHITE *et al.* 2017) and disease (CHO *et al.* 2015; CHEN *et al.* 2016a; BENDJILALI *et al.* 2017).

Although data collection of gene expression is increasingly cost effective, cost is still one of the main limiting factors when designing an RNA-seq time series experiment. For this reason, thoughtful experimental design in deciding the duration and sampling rate are extremely important (BAR-JOSEPH *et al.* 2012). Cyclic, developmental, and response processes will require very different strategies, both in

sampling rate, start and end points, and choice of replicates and controls (BAR-JOSEPH *et al.* 2012). For example, an experiment studying a developmental process will not be able to have time and age matched controls in the same way a drug treatment experiment could. Furthermore, with cost as a main limiting factor, the need for enough replicates will often mean having fewer time points, although experimental data and theoretical analysis has shown that under reasonable assumptions, sampling time points at higher resolution is preferred over having more replicates (SEFER *et al.* 2016).

While the collection of gene expression time series data is increasingly viable, the main challenge remains in the data analysis stage. Many studies default to using standard existing methods of analyzing static RNA-seq data, and although these methods are well established and are regularly integrated with other omics data (SPIES *et al.* 2017), there is now more awareness that these methods are not ideal for dealing with time course data (BAR-JOSEPH *et al.* 2012; SPIES *et al.* 2017). Time course data has very different underlying statistical assumptions that are not taken into account when using methods for analyzing static data, such as the temporal dependencies of the data and the correlation of genes between previous and subsequent time points (SPIES AND CIAUDO 2015). Many statistical and computational methods to analyze the exponential increase in data dimensionality and complexity are currently being developed to analyze gene expression data, but they all face non trivial challenges, and no benchmark has been achieved.

Statistical methods developed for microarray data analysis in the early 2000's can be used as a starting point to study *univariate time series* (one gene at a time) of gene expression profiles over time, and identify genes and pathways that are differentially expressed in at least one or more time points over the entire time-course of experiments (SUBRAMANIAN *et al.* 2005; EFRON AND TIBSHIRANI 2007; LAW *et al.* 2014). Recent advances in machine learning methods can also be used to cluster genes based on their temporal expression profiles (MONTERO AND VILAR 2014). However, a main challenge lies in *multivariate time series* analyses to find co-movement and lead-lag patterns between two gene expression trajectories, possibly accounting for spurious association caused by other genes. Some fields outside of genetics have a head start in developing and applying methods for analyzing similar multivariate dense time course data. In neuroimaging, brain scans such as MRIs and EEGs collect data throughout time to analyze connectivity between different brain regions (SETH *et al.* 2015), and diagnose conditions such as seizures, epilepsy, and brain tumors. In the financial sector, time series of the stock market can be used to build connectivity networks between firms and identify risk propagation (BILLIO *et al.* 2012; BASU *et al.* 2017). These advancements in time series analysis can and should still be used to inform methods and potential challenges of analyzing transcriptome dynamics.

In biological systems, gene interactions are also dynamic, implying that temporal gene expression profiles should be able to unveil causal dependencies among genes and pathways. Given this domain knowledge of transcriptomics and the

types of results that would be most salient to researchers in the field, we chose to model the underlying behavior from first principles using Granger Causality (GC) analysis (GRANGER 1969). This type of analysis was developed from the field of econometrics, a branch of economics where mathematical and statistical models are used to describe economic systems. The causality concept proposed by GC is based on predictability, where if the prediction of future values for a time series is improved by having knowledge of the past of a second time series, then this second time series is said to be Granger causal for the first one (GRANGER 1969). This method facilitates the discovery of time series systems that are correlated with a lag in time, and infers causality from that lagged correlation. The defined causality between two systems can then be treated as an edge in a network, progressively characterizing the relationship of multiple correlated systems.

Dense time course gene expression profiling of the *Drosophila melanogaster* innate immune response

In most organisms, the first line of defense against pathogens is the innate immune response. In vertebrates, the innate immune response plays the vital role of activating the adapted immune response, which is based on antigen-specific selection of antibodies and receptors. Insects, on the other hand, do not have an adaptive immune response and must rely on the innate immune response to recognize common microbial structures - such as peptidoglycans - to mount a generic and

systemic response (ALBERTS *et al.* 2002). This response is characterized by the mass production of antimicrobial peptides (AMPs) by the humoral response, and the internalization and degradation of pathogens by the cellular response (LEMAITRE AND HOFFMANN 2007). Launching an effective immune response is vital to an organism, because if a pathogen is not neutralized and cleared the resulting infection could kill the host. On the other hand, immune responses are energetically costly, giving rise to resource trade-offs between the immune response and other vital processes such as reproduction (SCHWENKE *et al.* 2016). As a consequence of these trade-offs, the activation and repression of the immune response is tightly regulated (AGGARWAL AND SILVERMAN 2008), and this tuning is likely to be mediated through a series of regulatory and feedback properties of the immune system.

In **Chapter 3**, I present the transcriptome dynamics profiling of the *Drosophila melanogaster* innate immune response. I performed a dense time-course RNA-seq experiment, and analyzed it by applying a broad range of statistical methods, including temporal clustering and gene set expression analysis, and used novel applications of Granger causality to construct putative gene interaction networks. The fruit fly *Drosophila melanogaster* is an ideal model organism to study transcriptome dynamics because it is a highly tractable laboratory system, allowing a diverse range of genetic, genomic, and molecular tools to aid scientific research on the system. *Drosophila* also possess many immune genes and pathways that are homologous to those of other organisms such as mammals. My experiment revealed distinct temporal patterns of

transient and sustained responses to infection that occur over different time scales, I provide several novel functional annotations for previously uncharacterized genes, and suggest new interactions governing temporal gene regulation of the immune response and trade-offs with metabolism and repair.

STUDYING THE INFLUENCE OF HOST GENETICS ON GUT MICROBIOME COMPOSITION

Organisms harbor a complex array of microbes, which affect how organisms interact with the world. These microbes interact with biological processes of the host, affecting fitness and disease. Some symbiotic relationships with these microbes are so important the hosts co-evolve with them: squids and their light organs evolved just to house bioluminescent bacteria that help squids camouflage from predators, aphids have bacteria that give them nutrients otherwise inaccessible to them and that transmit vertically to next generations (SHAPIRA 2016). In humans, the gut microbiome is estimated to house 10^{14} bacteria, providing an immense potential diversity of host-symbiont interactions (SHAPIRA 2016). The dynamics and mechanism by which these host-microbe interaction drives co-adaptation is currently under study. In mammals, the gut microbiome gets colonized during birth, and its composition and abundance is affected by multiple factors, such as diet, disease, and antibiotics. This gut microbiome modulates immune and metabolic phenotypes

(ROUND AND MAZMANIAN 2009; TURNBAUGH *et al.* 2009; GARRETT *et al.* 2010; VEIGA *et al.* 2010; RIDAURA *et al.* 2013) as well as and disease incidence. The association between gut microbiome populations and disease has been observed in obesity (LEY *et al.* 2005), heart disease (FAVA *et al.* 2006), diabetes (WEN *et al.* 2008), and liver cancer (YOSHIMOTO *et al.* 2013; SANDUZZI ZAMPARELLI *et al.* 2017), among others.

To accurately characterize the relationship between the gut microbiome and different phenotypic outcomes, it becomes vital to understand the ways in which the microbiome itself can be modulated by both environment and genetic factors. While it has long been clear that the gut microbiome composition is strongly impacted by environmental factors (ROTHSCHILD *et al.* 2018), studies have identified many genetic variants significantly associated with specific bacterial taxa abundances (DAVENPORT *et al.* 2015; BONDER *et al.* 2016; TURPIN *et al.* 2016; WANG *et al.* 2016; GOODRICH *et al.* 2017; IGARTUA *et al.* 2017; ROTHSCHILD *et al.* 2018).

To even begin to be able to study how host and microbiome could be interacting, we need to accurately survey and characterize the microbiome. We can do this by performing 16S rRNA sequencing or metagenomic sequencing (GOODRICH *et al.* 2014a). 16S rRNA sequencing is currently the most common method for surveying bacterial taxonomy. The targeted 16S rRNA gene is present in all bacteria, and codes for the 16S ribosomal RNA which is one of the components of a subunit in a bacterial ribosome. This gene is variable enough across different bacteria that it acts as

a signature sequence ideal for bacterial identification (YANG *et al.* 2016). More specifically, it has nine hypervariable regions (V1-V9) with different degrees of sequence conservation across taxa. The most variable regions allow us to distinguish between different species, while more conserved regions only allow broader taxonomic levels to be distinguished. Sub-regions V4-V6 are specifically recommended to optimize the phylogenetic resolution of bacterial identification (YANG *et al.* 2016). 16S rRNA sequencing, however, only provides data on microbiome membership and abundance. Metagenomics sequencing, on the other hand, surveys the entire genomes of all organisms in a sample, including viruses and fungi, and thus it can also provide information about functional pathways (also known as the hologenome). This means, however, that metagenomics sequencing is more expensive as it also requires more sequencing depth. Therefore, the choice in sequencing method will depend on the number of samples, experimental design, and desired data output.

Once microbiome membership and abundance information is collected, these data can be compared between healthy and disease states (such as cirrhosis (BAJAJ *et al.* 2012; BAJAJ *et al.* 2014), Non-Alcoholic Fatty Liver Disease (JIANG *et al.* 2015, and breast cancer (HIEKEN *et al.* 2016; ZHU *et al.* 2018)). In these scenarios, microbe abundance is analyzed with methods similar to those of transcriptomics, and "differential expression" can be calculated using pairwise comparisons. When studying the influence of the host genome on the microbiome, we can treat and

analyze microbiome composition as a quantitative phenotype, allowing the estimation of heritability of each microbiome attribute and the identification of quantitative trait loci are correlated with a certain bacterial taxa abundance (BENSON *et al.* 2010; MCKNITE *et al.* 2012; O'CONNOR *et al.* 2014; TURPIN *et al.* 2016).

Current human studies face a main limitation in not being able to control environmental factors such as diet, meaning that only the strongest genetic effects can be detected. Strategies to circumvent this limitation include twin studies, where monozygotic and dizygotic twins will have different degrees of genotype sharing, but similar early shared environment on their gut microbiota, such as maternal effect and familial dietary preferences (GOODRICH *et al.* 2014b). Another strategy is to study human populations that live in isolated closed communities such as the Hutterites, who all share the same diet by communally preparing and eating the same meals in a colony dining room (DAVENPORT *et al.* 2015). Mouse studies, on the other hand, provide total control of diet and other environmental factors. This, coupled with well-defined genetic differences among inbred lines, provide a good basis to dissect genetic and environmental factors affecting microbiome composition. Typical lab inbred mice strains, however, have limited genetic variation that does not accurately represent genetically diverse human populations. Heterogeneous mice stocks such as the Diversity Outbred mice were developed specifically to address this issue, providing an exciting new resource for research applications (CHURCHILL *et al.* 2012).

High-resolution QTL mapping with Diversity Outbred mouse strains identifies genetic variants that impact gut microbiome composition

The Diversity Outbred (DO) mouse population was designed to be the most genetically diverse mouse resource available. The DO mouse stock has highly heterogeneous genomes as a result of their very particular breeding scheme, where they are derived from the same eight progenitor lines used to establish the Collaborative Cross (CC) (COLLABORATIVE CROSS CONSORTIUM 2012). The eight progenitor lines included five classical inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, and NZO/HlLtJ), and three wild-derived strains representing different *Mus musculus* subspecies (CAST/EiJ, PWK/PhJ, and WSB/EiJ) (COLLABORATIVE CROSS CONSORTIUM 2012). The DO population was first established by randomly outbreeding CC mice, and is now maintained by randomly assigning breeding pairs for the next generation (CHURCHILL *et al.* 2012). This results in a robust population of highly diverse mice, where each individual DO mouse has a unique combination of segregating alleles, representing a unique mosaic of the original eight progenitor lines (**Figure 1.2**). Thanks to this outbreeding strategy, the DO mouse population has much higher levels of heterozygosity, more similar to the human condition than other inbred mouse resources. This characteristic of the DO mouse population allows a substantially increased genetic mapping resolution (CHURCHILL *et al.* 2012), while still being able to fully control their environment. A limitation of the DO mouse outbreeding design is that because each individual DO mouse is a unique mosaic of

the eight original founder strains, the genotype of each DO mouse is irreproducible. Replicating results within the DO population is thus more difficult than with inbred lines, as it is the case in natural populations. However, this limitation could be partially circumvented by using the reproducible genotypes of the CC and the eight founder lines as a form of validation (SVENSON *et al.* 2012). The CC/DO founder progenitor lines have already successfully been used to identify genetic associations with intestinal microbiome composition (O'CONNOR *et al.* 2014), further motivating the profiling of the gut microbiome in the DO mouse population.

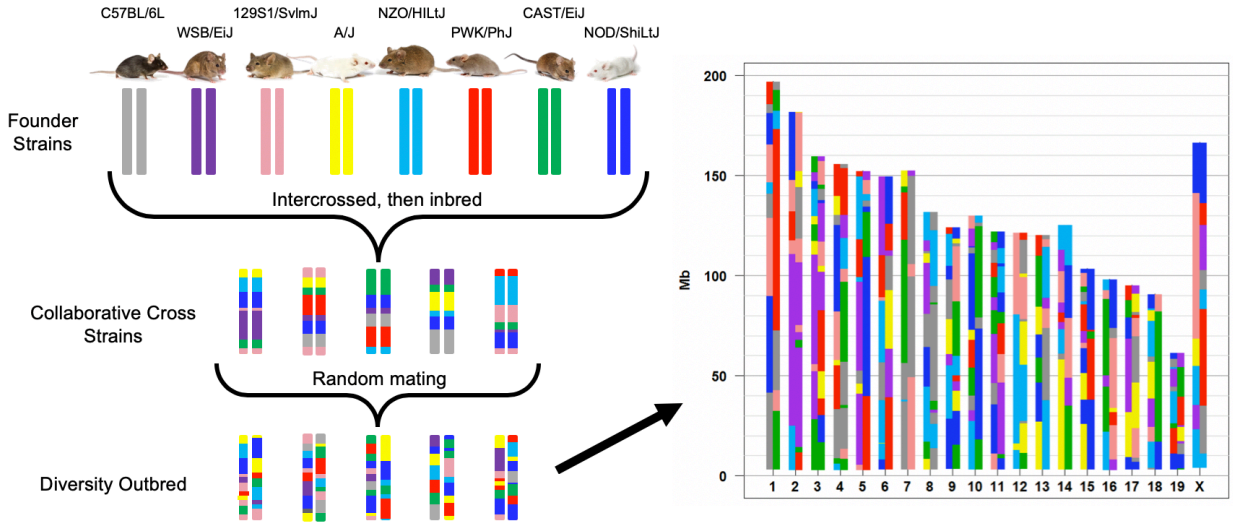


Figure 1.2. Development of Diversity Outbred mouse panel. Simplified crossing scheme from eight founder lines. A representation of the genome of a single DO mouse (on the right) shows the mosaic composition with each color representing each original founder.

In **Chapter 4**, I present the gut microbiota profiling of 247 Diversity Outbred mice using 16S rRNA gene sequencing, and study the influence of host genetics on gut microbiome composition by performing a high-resolution QTL mapping in the Diversity Outbred mouse panel using microbiome abundances as a response variable. This work uncovered strong evidence of host genetic factors associated with specific bacterial taxa abundances and functional molecular pathways, providing insight into the complex dynamics between host genetics and the gut microbiome, and isolating potential associations between microbial taxa and QTLs that may be involved in pathological disease phenotypes.

CHAPTER 2

EVALUATING THE PERFORMANCE OF SELECTION SCANS TO DETECT SELECTIVE SWEEPS IN DOMESTIC DOGS[1]

ABSTRACT

Selective breeding of dogs has resulted in repeated artificial selection on breed-specific morphological phenotypes. A number of quantitative trait loci associated with these phenotypes have been identified in genetic mapping studies. We analyzed the population genomic signatures observed around the causal mutations for 12 of these loci in 25 dog breeds, for which we genotyped 25 individuals in each breed. By measuring the population frequencies of the causal mutations in each breed, we identified those breeds in which specific mutations most likely experienced positive selection. These instances were then used as positive controls for assessing the performance of popular statistics to detect selection from population genomic data. We found that artificial selection during dog domestication has left characteristic signatures in the haplotype and nucleotide polymorphism patterns around selected loci that can be detected in the genotype data from a single population sample.

[1] Published as: Schlamp F, Made J, Stambler R, Chesebrough L, Boyko AR, and Messer PW (2016) Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. Molecular Ecology, 25: 342-356. doi:10.1111/mec.13485

However, the sensitivity and accuracy at which such signatures were detected varied widely between loci, the particular statistic used and the choice of analysis parameters. We observed examples of both hard and soft selective sweeps and detected strong selective events that removed genetic diversity almost entirely over regions >10 Mbp. Our study demonstrates the power and limitations of selection scans in populations with high levels of linkage disequilibrium due to severe founder effects and recent population bottlenecks.

INTRODUCTION

Identifying the molecular targets on which positive selection has acted constitutes one of the key challenges for modern population genetics. Ideally, positive selection is inferred directly from the frequency changes of selected alleles in a population over time (MALASPINAS *et al.* 2012; BANK *et al.* 2014; FOLL *et al.* 2015). However, such approaches require data on historic allele frequencies, otherwise they remain limited to situations of particularly rapid evolution that can be observed in real time.

Positive selection can also be detected from cross-population comparisons, based on the prediction that allele frequencies should differ between subpopulations if positive selection has acted in only one of them (LEWONTIN AND KRAKAUER 1973;

SABETI *et al.* 2007; AKEY *et al.* 2010). While such tests do not require time-course data, they remain limited to scenarios where selection acted only in a subset of individuals.

The most broadly applicable strategy for identifying positive selection is to search for its signatures in a single population sample, taken at a single point in time. Approaches from this category aim to identify the characteristic signatures of selective sweeps (KAPLAN *et al.* 1989; BARTON 2000; MAYNARD AND HAIGH 2007) which include a local trough in genetic diversity around the selected locus (KIM AND STEPHAN 2002), characteristic biases in the frequency distributions of single-nucleotide polymorphisms (SNPs) (BRAVERMAN *et al.* 1995; FAY AND WU 2000) and the presence of a long haplotype that extends much farther than expected under neutrality (SABETI *et al.* 2002). These signatures form the basis for most popular scans for selective sweeps (VITTI *et al.* 2013).

However, positive selection may not always produce selective sweeps. The classic selective sweep model presupposes that adaptation occurs from a single de novo mutation (HERMISSON AND PENNINGS 2005). Yet adaptation could often proceed from alleles that are already present as standing genetic variation (SGV) (ORR AND BETANCOURT 2001; INNAN AND KIM 2004; BARRETT AND SCHLUTER 2008). This should be particularly common in the evolution of polygenic traits, such as body size, where multiple trait-affecting alleles may be segregating in the population at any time (PRITCHARD *et al.* 2010).

Whether adaptation from SGV still produced sweep-like signatures depends on the initial frequency and age of a selected allele at the time when positive selection commences (PRZEWORSKI *et al.* 2005; PENNINGS AND HERMISSON 2006b). If the selected allele has been around long enough to recombine onto different haplotypes prior to the onset of positive selection, several haplotypes may then increase in frequency simultaneously. In this case, diversity is not necessarily reduced in the vicinity of the selected site and SNP frequency spectra can actually become biased towards intermediate frequencies (PRZEWORSKI *et al.* 2005). Very similar patterns are produced when adaptation involves several de novo mutations that independently emerged on distinct haplotypes, which is expected in very large populations or when mutational target sizes are large (PENNINGS AND HERMISSON 2006a; KARASOV *et al.* 2010; MESSER AND PETROV 2013). The patterns generated by adaptation from SGV and recurrent de novo mutation are commonly referred to as soft selective sweeps, in contrast to the classical hard selective sweep, where only a single haplotype rises in frequency (HERMISSON AND PENNINGS 2005).

Most scans for positive selection have been designed and tested exclusively under the assumption of a hard selective sweep model, and we do not know whether they provide a comprehensive picture of the mode and frequency of positive selection, or whether they identify only a subset of instances that is biased towards hard selective sweeps. Simulation studies have shown that selection scans quickly lose power for adaptation from SGV as the initial frequency of the selected allele increases

(PRZEWORSKI *et al.* 2005; TESHIMA *et al.* 2006; GARUD *et al.* 2015). However, it is unclear whether the alleles involved in adaptation from SGV are typically rare or frequent prior to the onset of selection.

Here, we use a set of known quantitative trait loci (QTLs) in the domestic dog (*Canis lupus familiaris*) as positive controls to examine the performance of popular selection scans in a real biological system. Our positive control loci were identified by genomewide association studies, rather than selection scans, and thus are not necessarily biased towards hard selective sweeps from the outset. We focus specifically on a subset of QTLs for which we know the causal mutations and could thus measure their frequencies in individual dog breeds. This information allowed us to assess which mutations have likely experienced positive selection in which breeds.

There are over 400 dog breeds today that have been bred for highly specific and diverse physical traits, including coat color, size, skull shape and behavioral traits such as obedience, herding and hunting. Modern dogs were the first animal to be domesticated, before cattle and horses, and domestication from their wolf ancestors goes back at least 15 000 years. Breeding programs throughout history, however, have resulted in periodic population bottlenecks, inbreeding, high levels of linkage disequilibrium in individual breeds and a prevalence of inherited diseases such as cancer, heart disease and hip dysplasia, among others (LINDBLAD-TOH *et al.* 2005). These features make purebred dogs a particularly challenging system for population genetic analysis.

MATERIALS AND METHODS

Genotyping

Genotyping data are from (SHANNON *et al.* 2015) briefly, blood was collected through cephalic venipuncture under Cornell IACUC # 2005-0151, and genomic DNA was extracted using a standard salt precipitation from EDTA blood samples and stored in the Cornell Veterinary Biobank.

Genotyping was performed using the Illumina 170k CanineHD array, which was developed using the dog reference sequences (generated from a Boxer and a Poodle) and pooled DNA from a series of European and Asian breeds (Irish Wolfhounds, West Highland White Terriers, Belgian Shepherds and Shar-Peis) as well as pooled wolf DNA as described in (VAYSSE *et al.* 2011). We customized this array by adding 12 143 markers ascertained from whole genome sequencing data from mostly Eurasian village dogs (AUTON *et al.* 2013), approximately equally split between East Asian and Western dogs. Markers were preferentially chosen for being in coding regions but poorly tagged by existing array markers. The genotypes were combined with published CanineHD data from (AXELSSON *et al.* 2013). The full SNP panels (3 million SNPs for the CanineHD array design and 14 million SNPs for the custom array content) were pruned for evenness, ability to design probe sequence and efficiency. In general, no effort was made to differentially enrich one source or another in particular regions of the genome, except that a subset of custom SNPs

were specifically included in the IGF1 and MSRB3 regions to facilitate fine mapping of those loci. No such enrichment of markers was made for the other 10 loci.

The unimputed data set contained a call rate over 99.1%, and no locus contained >5% missing data. Imputation was performed because some methods to detect positive selection require no missing data, but the proportion of imputed genotypes is negligible and unlikely to bias the results.

Phasing was performed for all autosomal and X chromosome markers with minor allele frequency (MAF) >0.01 using SHAPEIT (DELANEAU *et al.* 2013). Select regions showing strong evidence of positive selection when comparing allele frequency data across breeds and associated with a known phenotypic effect were chosen for analyzing selection signatures in each population.

Frequency estimates of causal mutations in breeds

Selection signatures were estimated from a randomly selected subset of 25 unrelated individuals per breed. The allele frequency of the causal variant (when known) or the top associated variant was estimated from the entire data set (SHANNON *et al.* 2015) based on a much larger number of individuals genotyped (25–722 dogs per breed).

Selection scans

The hapFLK statistic was calculated using the program HapFLK (version 1.2) (FARIELLO *et al.* 2013), downloaded from: https://forge-dga.jouy.inra.fr/projects/hapflk (August 2015). The population tree was obtained by hapFLK to compute Reynolds distances and the kinship matrix across all 25 breeds genomewide, using Culpeo Fox as the outgroup. The hapFLK scan was run using all 25 breeds genomewide. We used the following parameters: eight clusters (−K 8), 20 EM runs to fit the LD model (−nfit=20), phased data (–phased). Once hapFLK values were generated, we calculated P-values by fitting a standard normal distribution genomewide in R (FARIELLO *et al.* 2013).

iHS scans were performed using the program Selscan (version 1.0.4) (SZPIECH AND HERNANDEZ 2014), downloaded from: http://github.com/szpiech/selscan (April 2015). All scans were run on polarized data with default iHS Selscan parameters: –max-extend 1 000 000 (maximum EHH extension in bp), –max-gap 200 000 (maximum gap allowed between two SNPs in bp), –cut-off 0.05 (EHH decay cut-off). We used the recombination map of Auton et al. (AUTON *et al.* 2013). The output results for each SNP were then frequency-normalized over all chromosomes using the script norm, provided with Selscan. This normalization was also performed using default parameters: –bins 100 (number of frequency bins). The fractions of SNPs with values above 2.0 were calculated over genomic windows of specified sizes (25, 51,

101, 201 neighboring sites on our chip) and the resulting ratio was assigned to the position of the central SNP of the window, as suggested in (VOIGHT *et al.* 2006).

In contrast to iHS, which measures the length of haplotypes in terms of genetic distance and thereby requires specification of a recombination map, the nSL statistic measures haplotype lengths in terms of the number of segregating sites in the sample, making it more robust to recombination rate variations. nSL scans were performed using the original implementation of the statistic (FERRER-ADMETLLA *et al.* 2014), downloaded from: http://cteg.berkeley.edu.proxy.library.cornell.edu/~nielsen/resources/software/ (April 2015). All scans were run using default nSL parameters. The output results were normalized and averaged over windows following the same procedures used for iHS.

The H statistic was estimated using the program h-scan (version 1.3), downloaded from: http://messerlab.org/resources/ (April 2015). The H statistic measures the average length of pairwise haplotype homozygosity tracts around a given genomic position in base pairs. The length of the homozygosity tract $h_{ij}(x)$ for a pair of samples (i, j) at genomic position x is defined as the distance between the first heterozygous site to the left and to the right of x. The value of $H(x)$ at position x is then defined as the average over all pairs in the sample: $H(x) = \frac{2}{n(n-1)} \sum_{i<j} h_{ij}(x)$.

H values were calculated at each SNP position in the data set. All scans were run using default H-scan parameters.

H12, Tajima's D and π values were calculated over windows of a fixed number of SNPs on our genotyping chip (d = 25, 51, 101 and 201). The estimated values of each statistic were then assigned to the position of the central SNP of the window. H12 values were estimated following the definition provided in (GARUD *et al.* 2015). Tajima's D values were variance-normalized according to the formulas given in (TAJIMA 1989). Note that because all scans were run on a fully imputed data set, haplotype clustering for H12 is unambiguous in this study.

CLR is a likelihood-ratio test that compares the SNP frequency spectrum in candidate regions with the genomic background to identify regions with sweep-characteristic deviations. CLR scans were performed using the software SweeD (version 3.1) (PAVLIDIS *et al.* 2013), downloaded from: http://sco.h-its.org/exelixis/web/software/sweed/ (April 2015). For each chromosome CLR was calculated with a resolution of 10 000 bins, assuring that the density of bins is much higher than the density of SNPs in each chromosome. All CLR scans were run on unfolded spectra using the polarized data.

RESULTS

A set of 12 positive controls for studying the signatures of positive selection in dogs

The molecular basis of morphological phenotypes selected during domestication of dog breeds has been extensively studied, and dozens of QTL for breed-specific phenotypes have been identified, which often explain surprisingly high fractions of phenotypic variance (RIMBAULT *et al.* 2013). We compiled a set of 12 known QTLs distributed across nine chromosomes of the dog genome for which we know the specific mutations that are likely causal for breed-specific traits (**Table 3.1**). Our set includes mutations affecting body size [IGF1R, STC2, GHR and IGF1 (RIMBAULT *et al.* 2013)], fur type [MC5R and KRT71 (CADIEU *et al.* 2009)], coat color [MC1R and TYRP1 (SCHMUTZ AND MELEKHOVETS 2012)], hair length [FGF5 (CADIEU *et al.* 2009)], lip morphology (CHRNB1), ear morphology [MSRB3 (BOYKO *et al.* 2010)] and snout length [BMP3 (SCHOENEBECK AND OSTRANDER 2013)]. These loci are representative of loci that show evidence of strong selection based on elevated levels of divergence between breeds (AKEY *et al.* 2010; BOYKO *et al.* 2010; VAYSSE *et al.* 2011). Some QTLs known to be associated with breed-specific morphological traits were intentionally excluded from our analysis, because the causal mutations were either not well-tagged by markers in our data set [e.g. the insertion in the 3'UTR of RSPO2 associated with a furnishings phenotype (CADIEU *et al.* 2009)], or the locus

was very close to another locus [e.g. the size-related locus HMGA2 (RIMBAULT *et al.* 2013) that is only 300 kbp away from MSRB3].

Table 2.1. Set of known QTLs with mutation frequencies in individual breeds

	MC5R	IGF1R	STC2	GHR	CHRNB1	MC1R	MSRB3	TYRP1	IGF1	KRT71	FGF5	BMP3
chr:position	1:24430748	3:41849479	4:39182836	4:67040898	5:32382510	5:63694334	10:8037693	11:33326685	15:41220982	27:2539211	32:4509367	32:5231894
phenotypic trait	fur type	body size	body size	body size	hangling lips	coat color	ear type	coat color	body size	curly coat	hair length	snout length
causal?	yes	yes	likely	likely	likely	yes	yes	yes	yes	yes	yes	yes
included on chip?	yes	no	no	no	yes	yes	no	yes	yes	yes	no	no
Border Collie	0.02	0.01	0.31	0.62	0.43	0.01	0.95	0.28	0.49	0.09	0.68	0.00
Boxer	1.00	0.00	0.77	0.05	0.39	0.00	0.14	0.00	0.02	0.00	0.00	0.00
Cavalier King Charles Spaniel	1.00	0.00	0.99	1.00	0.78	0.81	0.01	0.00	1.00	0.00	0.79	0.00
Cocker Spaniel	0.97	0.04	0.61	0.71	0.12	0.70	0.00	0.04	0.96	0.01	0.42	0.00
Dachshund	0.96	0.88	0.20	0.85	0.33	0.09	0.00	0.03	0.80	0.00	0.28	0.00
English Setter	1.00	0.00	0.04	0.41	0.28	0.80	0.02	0.03	0.42	0.01	0.98	0.00
English Springer Spaniel	0.84	0.00	0.16	0.29	0.60	0.02	0.00	0.53	0.56	0.00	0.52	0.01
French Bulldog	1.00	0.02	0.19	0.24	1.00	0.26	0.65	0.00	0.94	0.00	0.02	0.98
German Shepherd	0.02	0.06	0.08	0.09	0.11	0.11	1.00	0.00	0.01	0.08	0.44	0.00
Golden Retriever	0.60	0.00	0.00	0.09	0.53	1.00	0.15	0.00	0.12	0.05	0.99	0.00
Havanese	0.51	0.22	0.66	0.74	0.16	0.52	0.32	0.25	0.94	0.32	0.73	0.00
Irish Wolfhound	0.99	0.00	0.00	0.00	0.72	0.00	1.00	0.00	0.00	0.00	0.14	0.00
Jack Russell Terrier	0.76	0.12	0.22	0.23	0.46	0.07	0.04	0.07	0.96	0.24	0.04	0.42
Labrador Retriever	0.31	0.00	0.14	0.20	0.10	0.61	0.28	0.31	0.42	0.01	0.09	0.00
Maltese	0.62	0.22	0.74	0.98	0.24	0.99	0.26	0.01	0.97	0.08	0.94	0.02
Miniature Schnauzer	0.69	0.14	0.23	0.94	0.42	0.10	0.89	0.01	1.00	0.01	0.03	0.97
Newfoundland	0.06	0.00	0.01	0.51	0.97	0.00	0.03	0.00	0.00	0.00	0.89	0.00
Papillon	0.33	0.36	0.86	0.31	0.83	0.03	0.47	0.00	0.97	0.00	0.89	0.02
Poodle	0.38	0.05	0.30	0.03	0.12	0.61	0.26	0.12	0.51	0.95	0.74	0.01
Rottweiler	0.99	0.00	0.00	0.01	0.02	0.01	0.73	0.00	0.89	0.00	0.11	0.00
Saint Bernard	0.20	0.11	0.00	0.02	0.96	0.00	0.00	0.00	0.04	0.00	0.48	0.00
Shetland Sheepdog	0.00	0.40	0.00	0.50	0.02	0.00	0.62	0.00	0.35	0.00	0.17	0.06
Shih Tzu	0.80	0.09	0.98	0.98	0.54	0.06	0.06	0.13	1.00	0.02	0.85	0.35
Vizsla	1.00	0.35	0.06	0.01	0.04	1.00	0.02	0.21	0.54	0.00	0.01	0.00
Yorkshire Terrier	0.70	0.73	0.79	0.68	0.24	0.00	0.00	0.02	1.00	0.00	0.36	0.00

The twelve QTLs included in our analysis span a wide range of phenotypic traits that likely experienced positive selection in particular subsets of breeds during the domestication of dogs. For nine of the 12 loci, at least one causal mutation for the phenotypic trait has been identified and for the remaining three loci (STC2, GHR, CHRNB1) we have promising candidate mutations. We focused on one such mutation for each locus (positions are specified in the first row of the table). For six of the 12 loci, these mutations are included on the genotyping chip. We studied 25 dog breeds in our analysis. Numbers in the cells specify the frequency of the known/likely causal mutations in each particular breed (**Materials and Methods**). Higher frequencies are indicated by darker shaded cells.

We analyzed the population genetic signatures we observed around these 12 loci in population samples from 25 dog breeds, spanning a broad range of morphological variation (**Table 2.1**). For each of the 25 breeds, we genotyped a random sample of 25

dogs at ~180 000 SNP markers, using a semi-custom SNP array (**Materials and Methods**). For six of the 12 loci, the known/likely causal mutations are included on the chip. Genotypes were then phased and imputed on the whole set, yielding 50 haploid genomes for each of the 25 breeds (1250 genomes over the whole data set, see **Materials and Methods**). We polarized SNPs using allele information from Culpeo Foxes for SNPs where such information was available (99.46%) and assumed the minor allele to be the derived allele otherwise. To assess whether a particular mutation was likely under positive selection in a particular breed, we estimated population frequencies for the focal mutation at each of the 12 loci in each of the 25 breeds (**Table 2.1**, **Materials and Methods**).

Genomewide selection scans in 25 dog breeds

We first used the hapFLK statistic (FARIELLO *et al.* 2013) to confirm that our 12 positive controls indeed show signatures of positive selection in cross-population comparisons. hapFLK was developed to detect differences in haplotype frequencies across many populations, using an FST-based framework that also incorporates information about the hierarchical structure of the populations. **Figure 2.1** shows the results from our genomewide hapFLK scan including all 25 breeds (only the chromosomes that contain positive controls are shown). Each of our 12 controls is associated with a significant peak ($P < 0.05$) in the hapFLK scan, with 7 of the 12

detected as extreme outliers (P < 0.001). **Figure S2.1** shows the underlying hierarchical breed structure inferred by hapFLK.

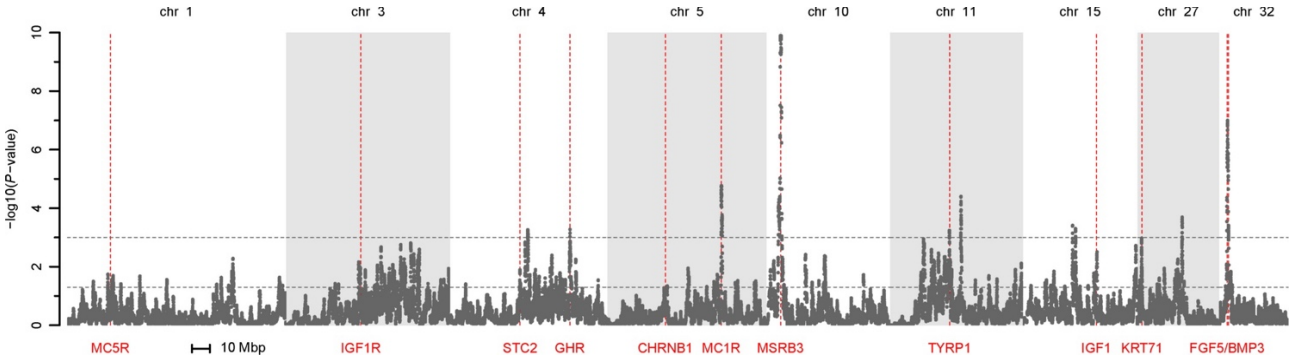


Figure 2.1. HapFLK results. The figure shows the results from the hapFLK scan performed over all 25 breeds. Results are shown only for those chromosomes that contain at least one of our control loci. The genomewide thresholds corresponding to $P < 0.05$ and $P < 0.001$ are shown as horizontal dashed lines. The locations of the control loci are indicated by vertical red lines.

To test whether positive selection at our control loci has also left detectable signatures in the patterns of genetic variation in individual breeds, we ran genomewide scans using seven popular statistics for identifying sweep signatures from a single population sample. We studied both SNP frequency-based and haplotype-based statistics.

Tajima's D is a popular frequency-based statistics that compares the number of segregating sites (s) in a population sample with levels of heterozygosity (π) to detect genomic regions with an excess of low or high frequency SNPs compared to neutral expectations (TAJIMA 1989). Another widely used statistic is CLR, which underlies the programs Sweepfinder (NIELSEN *et al.* 2005) and SweeD (PAVLIDIS *et al.* 2013). We

included both Tajima's D and CLR as two classic representatives of frequency-based statistics in our study. We also included pairwise heterozygosity per nucleotide (π).

Haplotype-based statistics search for elevated levels of haplotype homozygosity expected around a sweep locus. One of the most popular approaches in this category is integrated haplotype score (iHS), which searches for loci where the derived allele resides on a longer haplotype than the ancestral allele (VOIGHT *et al.* 2006). In addition to iHS, we also included the nSL statistic, a recent modification of iHS that has improved power in detecting soft sweeps (FERRER-ADMETLLA *et al.* 2014). Note that iHS and nSL are both targeted at the identification of incomplete sweeps, where the selected allele is not fixed in the sample. We further included the H12 statistic that has been developed for the detection of both hard and soft sweeps (GARUD *et al.* 2015). Finally, we included a simple haplotype statistic (H) that measures the average length of pairwise haplotype homozygosity tracts around each SNP in base pairs (**Materials and Methods**).

All of the above statistics, except H require specification of analysis parameters. For iHS and nSL, minimum haplotype homozygosity levels need to be specified below which haplotypes are no longer extended. To improve sensitivity, iHS and nSL are also combined over neighboring data points, which introduces a window-size parameter (VOIGHT *et al.* 2006). CLR requires the specification of the number of grid points along the chromosome. H12, Tajima's D and π require specification of the

length of an analysis window over which their values are estimated. These windows are typically defined in terms of a fixed number of SNPs. Given that SNPs in our data were estimated from all 25 breeds, we can either define such windows using all SNP on our chip, or only those SNPs that are actually segregating in the particular breed of interest. We decided to define windows using all SNPs on our chip to make results comparable between breeds. Note that this may be considered an unfair advantage to the π and H12 statistics, as it incorporates cross-population information: Consider, for example, a window of 25 neighboring SNPs identified using information from all breeds, for which diversity is depleted entirely in a particular breed. In that case, $\pi = 0$ and H12 will yield a value of one, as only a single haplotype will be present in the window. However, it turns out that in practice, the performance of these statistics is not strongly affected by whether we define window sizes using all SNPs on our chip, or just the segregating sites in the particular breed for which the given statistic is estimated, as we will show below.

Figure 2.2 shows the results of the seven statistics (iHS, nSL, H, H12, CLR, Tajima's D and π) for the example of French Bulldogs. Different statistics vary markedly in appearance and statistical properties, although some statistics are more similar than others. As expected, iHS and nSL identify largely overlapping candidate regions. Likewise, H12 and H behave similar to each other, consistent with the fact that both statistics measure local levels of haplotype homozygosity (although H12 measures homozygosity over a window of fixed size, whereas H measures the average

length of pairwise homozygosity tracts around a SNP in the sample). Tajima's D and π yield similar results as well, suggesting that the signal in Tajima's D in regions with negative values is driven primarily by local reductions in π. Increasing window sizes generally tends to smoothen results for the window-based statistics, reducing noise at the price of lowered sensitivity. **Figure S2.2** shows the results of the scans in French Bulldogs when defining windows using only those SNPs that are actually segregating in our sample. Results are almost indistinguishable between the two approaches, suggesting that our choice of defining window sizes using all SNP on the chip does not have a large effect on the analysis. Selection scans for all 25 breeds are presented in **Figure S2.3**.

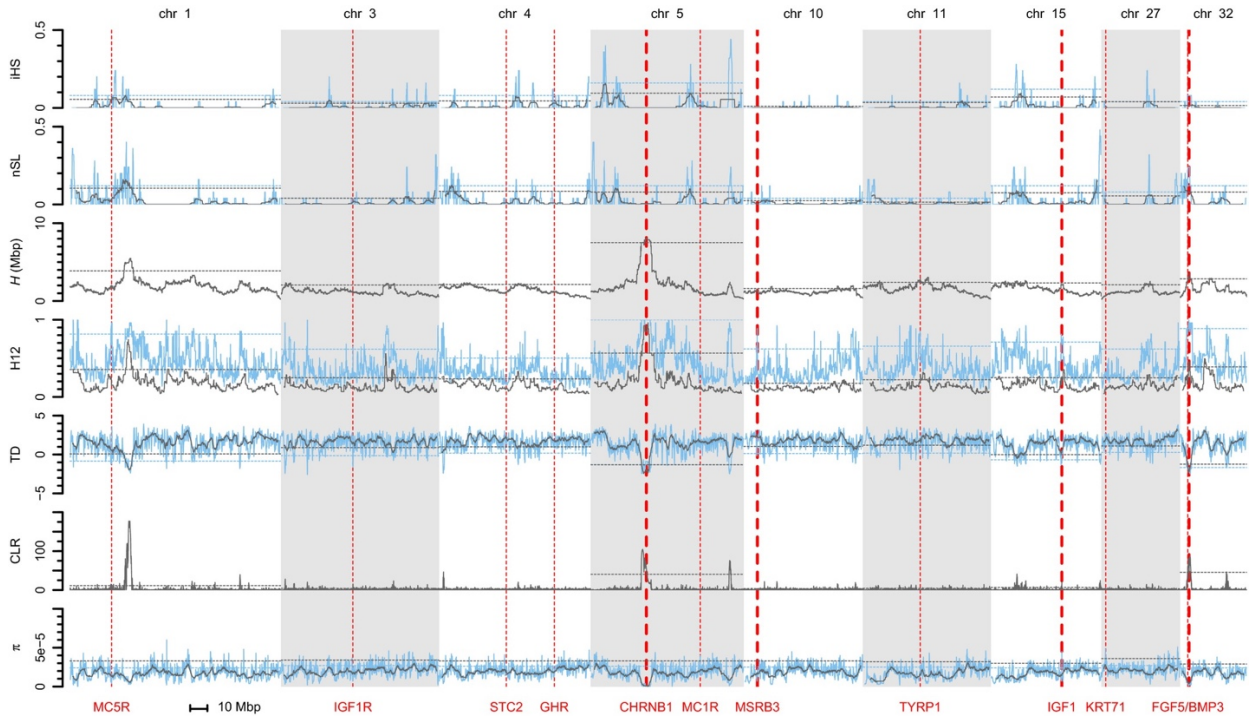


Figure 2.2. Single population selection statistics in French Bulldogs. Results for iHS, nSL, H, H12, Tajima's D, CLR and π along those chromosomes that harbor at least one of our positive controls. For iHS, nSL, H12, Tajima's D and π, the blue lines show results for a window size of 25 SNPs, grey lines show results for a larger windows of 201 SNPs. Note that signals of positive selection correspond to higher values of iHS, nSL, H, H12 and CLR, but lower values of π and more negative values of Tajima's D. Horizontal dashed lines indicate the 95% quantile cut-offs for the given statistic and window size, which we estimated for each chromosome separately. The positions of the controls are indicated by vertical red lines. The width of these lines corresponds to the frequency at which the causal mutation was observed in the breed in our sample (thin lines: low frequency; thick lines: high frequency). Scans for all 25 breeds are presented in **Figure S2.3**.

Note that our SNP data were obtained from a genotyping chip, rather than direct sequencing (**Materials and Methods**). Low-frequency SNPs are therefore underrepresented. This should systematically bias Tajima's D values towards more positive values and may also affect the CLR statistic. However, as we expect these biases to be present genomewide, relative comparisons between different regions along the genome should remain informative. Note also that levels of nucleotide and

haplotype diversity vary widely between breeds and that our data set covers a wide range of these values (**Figure 2.3**).

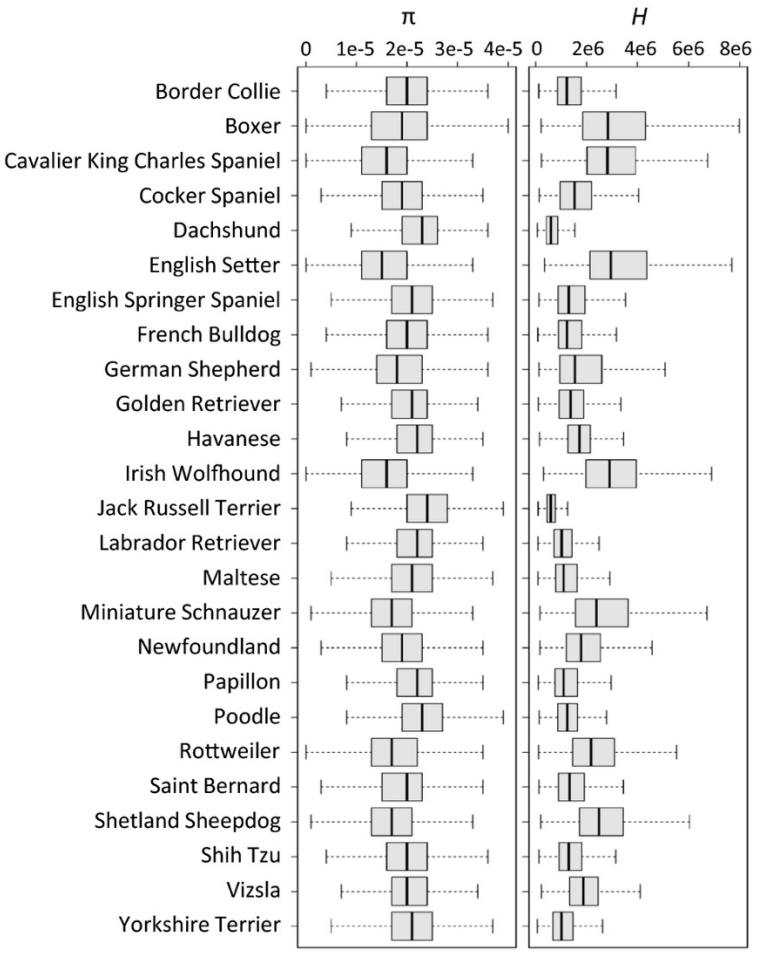


Figure 2.3. Strong differences in nucleotide and haplotype diversity between breeds. The figure shows the average genomewide levels of nucleotide heterozygosity (π) and length of pairwise haplotype homozygosity tracts (H) in each breed. Values were estimated across all genomewide SNP positions for the given breed. Values of π were estimated using a window size of 51 SNPs. Box plots show medians with first and third quantiles. Note that these values were obtained from our genotyping chip, which comprises only a subset of polymorphic sites. The true diversity levels will be higher and homozygosity stretches will be shorter.

Genomewide outlier characteristics

Our genomewide scans reveal characteristic differences in the number, sizes and distributions of 'peaks' identified by the seven selection statistics. To quantify these differences, we assigned peaks across the genome using an outlier criterion: We considered all data points with value above a given chromosome-wide quantile threshold (σ) as candidates for positive selection. For each such data point, we then defined a peak as the window of radius d base pairs around its genomic position. Overlapping peaks were combined into a single peak.

We employed a simple outlier approach, rather than using an explicit neutral null model, as such a model would require information about the particular demographic history of each individual breed. Unfortunately, we do not generally know much about these demographic histories, except that they can be complicated and differ profoundly between breeds. Our outlier criterion does not require knowledge of demography, but it cannot provide us with information about false-discovery rates. However, in our study, we focus on assessing the performance of selection scans at known positive controls, which is conceptually different from the discovery of novel targets in that we are not generally worried about the detection of false positives. Instead, we want to study whether scans correctly place the controls among the top signals genomewide. Our rationale is that our control loci should be located in or near the regions with the strongest signals. The simple outlier approach

allows us to draw general conclusions about the number and distribution of such regions identified by each statistic under a given threshold criteria.

Table 2.2 shows the average number of peaks identified genomewide per breed and the average fraction of the genome covered by these peaks, using two different quantile thresholds ($\sigma = 0.95$ and $\sigma = 0.99$) and three window radii (d = 10, 50 and 250 kbp) for each statistic tested. As expected, lower thresholds and larger peak radii both tend to produce more peaks and larger fractions of the genome covered than higher thresholds and smaller radii. Values range from ~600 peaks identified genomewide per breed by CLR under the 0.95 criterion with d = 10 kbp, to only ~10 peaks identified genomewide per breed for iHS under the 0.99 criterion with d = 250 kbp. Note that nucleotide heterozygosity is very low in our data set: on average, $\pi \sim 10-5$ per site for the breeds in our data set (**Figure 2.3**). Thus, neighboring SNPs tend to be several kbp apart, which is why we chose rather large windows.

Table 2.2. Genomewide peak statistics

	cutoff σ=0.95			cutoff σ=0.99		
	d=10000	d=50000	d=250000	d=10000	d=50000	d=250000
iHS.25	98.1 (13.3%)	87.7 (14.4%)	70.6 (19.0%)	40.2 (3.2%)	34.3 (3.67%)	28.8 (5.5%)
iHS.201	20.4 (6.0%)	18.7 (6.3%)	16.2 (7.3%)	12.2 (2.7%)	11.4 (2.8%)	9.6 (3.4%)
nSL.25	124.8 (8.4%)	110.5 (9.8%)	86.4 (15.5%)	45.2 (1.9%)	40.1 (2.4%)	33.4 (4.5%)
nSL.201	24.8 (4.9%)	22.4 (5.2%)	18.4 (6.4%)	15.4 (1.2%)	13.7 (1.3%)	11.6 (2.1%)
H	115.2 (5.9%)	68.5 (7.1%)	36.5 (9.9%)	64.1 (1.4%)	42.0 (2.0%)	23.6 (3.8%)
H12.25	170.7 (5.9%)	154.0 (7.8%)	115.9 (15.7%)	64.4 (1.7%)	59.8 (2.4%)	50.3 (5.6%)
H12.201	43.0 (5.4%)	40.5 (5.9%)	33.6 (8.0%)	20.8 (1.4%)	19.7 (1.7%)	16.9 (2.7%)
TD.25	207.1 (5.3%)	162.1 (7.5%)	131.8 (16.0%)	57.2 (1.1%)	47.3 (1.7%)	42.7 (4.3%)
TD.201	65.6 (4.9%)	44.1 (5.5%)	30.8 (7.6%)	31.5 (1.0%)	21.5 (1.3%)	15.5 (2.4%)
CLR	589.5 (6.4%)	430.4 (12.3%)	258.1 (31.7%)	88.6 (1.2%)	70.6 (2.1%)	57.8 (5.8%)
π.25	345.0 (4.7%)	251.3 (8.1%)	197.1 (21.3%)	110.5 (1.0%)	87.4 (2.1%)	79.8 (7.0%)
π.201	92.4 (5.7%)	55.9 (6.6%)	35.8 (9.1%)	53.2 (1.8%)	30.9 (2.2%)	20.0 (3.6%)

The table shows the number of peaks identified by each statistic for a given quantile threshold (σ) and window radius (d) along those chromosomes that harbor at least one of our positive controls, averaged across all breeds. The numbers in parentheses specify the average percentage of the genome that is covered by the peaks in the particular scenario.

Performance of selection scans at positive controls

We next assessed the performance of each statistics in identifying signals of positive selection at each positive control locus. This was performed by measuring the distance between the causal sweep mutation and the next data point with a value above the 0.95 chromosome-wide quantile of the given statistic. If the statistic yielded a value above the 0.95 quantile at the actual causal mutation, we set the distance to zero. We used chromosome-wide quantiles, rather than genomewide quantiles, because levels of nucleotide and haplotype diversity vary systematically between the different chromosomes within a breed (**Figure S2.3**).

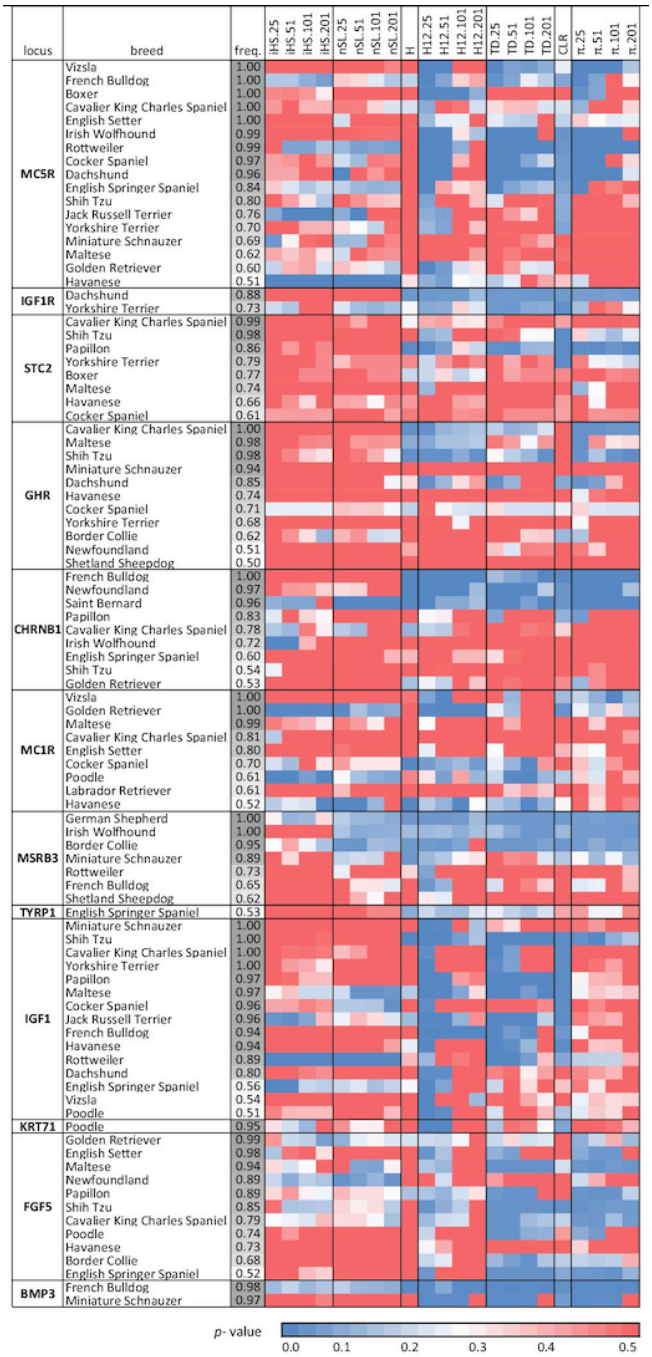
A close distance between a causal mutation and an outlier data point is not itself a clear indication that the given statistic has high power in detecting the locus. The close distance could simply be due to chance if values of the statistic fluctuate fast along the chromosome, so that any random genomic position would typically be close to a data point with value above the 0.95 threshold. To assess the significance of a measured closest distance, we therefore calculated empirical P-values for observing the given or a shorter distance by chance, based on the distribution of closest distances at random genomic locations in the particular chromosome and breed. Note that these empirical P-values are not P-values in the regular sense obtained from a neutral null model, but simply indicate the extent to which the observed distance is an empirical outlier regarding the chromosome-wide distribution.

The resulting P-values for all locus/breed combinations in which the causal allele has a frequency of at least 50% are shown in **Table 2.3**. For the window-based statistics, we show result for window sizes 25, 51, 101 and 201 SNPs. The actual distances between the causal mutation and the closest outlier are provided in **Table S2.1**.

Table 2.3 shows that there is substantial variation in the ability to detect signatures of positive selection among different statistics, loci and breeds. As expected, iHS and nSL produce rather similar results. Interestingly, H12, Tajima's D and π also appear to be more similar to each other than to the other statistics. H12

identifies the largest number of locus/breed combinations, at least when using the small windows size of 25 SNPs (**Table 2.4**). iHS and nSL identify only one or two (depending on window size) of the 15 fixed sweeps under a 0.05 significance level. They fail to identify any fixed sweep when using a stricter 0.001 significance level. These particular results for iHS and nSL are not surprising, given that both statistics were designed to detect incomplete sweeps. CLR does identify several sweeps under the 0.05 significance level but also does not detect any sweep under the 0.001 significance level. H and π have lower performance than H12 and Tajima's D but better performance than CLR, iHS and nSL, especially under the stricter 0.001 significance level.

Table 2.3. Performance of selection scans at individual QTLs



The table shows for each locus the breeds ordered by the frequency of the causal allele in the particular breed (only breeds with frequency above 50% are shown). The coloring of the cells specifies the *P* -value of the measured distance between the causal mutation and the closest data point that lies above the 95% threshold for the given statistic. Our empirical *P* -values were calculated from the empirical distributions of closest distances for random genetic loci. Different statistics vary widely in whether they detect signatures of positive selection for a given locus/breed combination. In general, signatures of positive selection tend to be detected more frequently, the higher the frequency of the selected mutation in the specific breed.

Table 2.4. Scan performances under different significance thresholds

signif.	freq.	*n*	iHS.25	iHS.51	iHS.101	iHS.201	nSL.25	nSL.51	nSL.101	nSL.201	H	H12.25	H12.51	H12.101	H12.201	TD.25	TD.51	TD.101	TD.201	CLR	π.25	π.51	π.101	π.201
p <0.05	f =1	15	1	1	1	2	1	0	0	0	4	11	8	5	2	6	2	5	4	6	8	5	6	4
	0.2<f <1	126	9	8	4	8	9	12	8	9	16	40	24	11	8	31	29	25	14	25	25	19	19	12
p <0.001	f =1	15	0	0	0	0	0	0	0	0	2	7	6	3	2	2	1	1	2	0	2	3	3	1
	0.2<f <1	126	7	5	3	6	5	5	3	3	5	20	12	5	1	13	9	7	5	0	6	6	5	2

fraction identified

0/n *n/2n* *n/n*

The table shows for different sets of locus/breed combinations the number of combinations in which each statistic identifies a signal of positive selection under a significance threshold of $P < 0.05$ (top) or $P < 0.001$ (bottom). We classified locus/breed combinations into two sets according to whether the selected allele is fixed in our sample ($f = 1$) or polymorphic ($0.2 < f < 1.0$). The 'n' column shows the total number of locus/breed combinations in each set. We did not include locus/breed combinations where the selected allele was below 20%.

The similarity between H12 and π may appear counterintuitive at first glance, given that H12 measures haplotype homozygosity, whereas π is based on nucleotide heterozygosity. A likely reason for this is that we defined window lengths using all SNPs present on our chip. Analysis windows are therefore the same in all breeds. In those breeds where a sweep has occurred, fewer sites will actually be polymorphic, reducing π. However, this is also expected to decrease the number of different haplotypes, as fewer SNPs will be present that can break up haplotypes, yielding higher H12 values.

Haplotype homozygosity levels increase with frequency of selected alleles

Generally, we expect that scans should perform better at detecting sweeps the higher the frequency of the selected allele in a particular breed. This tendency is indeed visible in **Figure 2.3** and **Table 2.3**. We also observed a clear positive correlation between the frequency of the selected allele in a breed and the value of H at a locus for all loci, except TYRP1 and IGF1 (**Figure 2.4**). H simply measures the average haplotype homozygosity lengths among all individuals in the sample. The observation of higher H values for higher frequency alleles is therefore consistent with the selected alleles residing on longer haplotypes than the ancestral alleles, as more individuals carrying these longer haplotypes will increase the average haplotype lengths among all individuals.

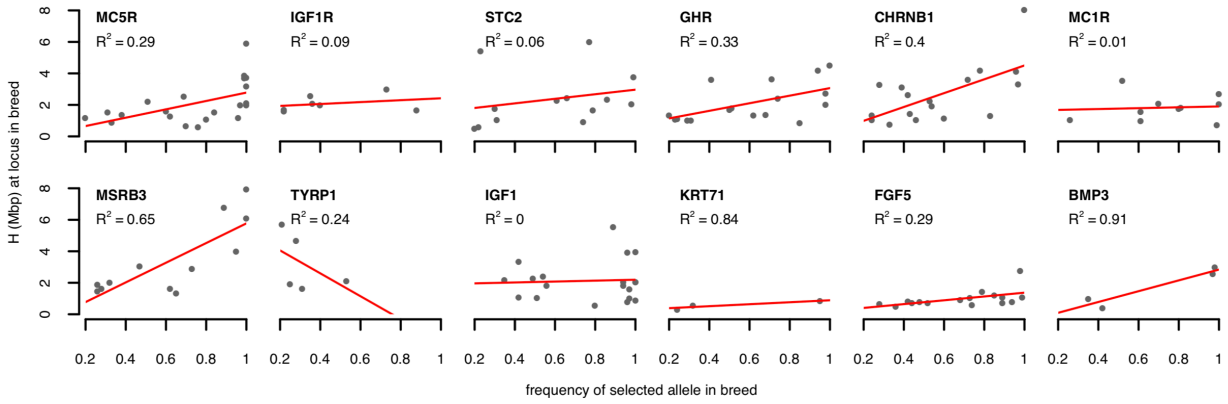


Figure 2.4. Haplotype homozygosity levels increase with frequency of selected allele. Each panel shows for the particular locus the values of H at the causal site as a function of the frequency in the specific breed (only breeds where the selected allele has a frequency >20% are shown). We observed a positive correlation (measured by R^2) between allele frequency and the value of H in the breed for all loci except TYRP1 and IGF1.

Note that iHS and nSL lose power to detect a sweep when the selected allele is fixed in the breed (**Table 2.3**), as has been observed previously (SCHRIDER *et al.* 2015). As mentioned before, this is expected given that both statistics were specifically designed to detect incomplete sweeps, where both the ancestral and derived allele are still segregating in the population and the haplotypes on which they reside can be compared with each other.

Positive selection has produced both hard and soft selective sweeps

We analyzed the haplotype patterns and SNP frequency spectra around individual loci in individual breeds to see whether we can understand why some statistics perform better than others at detecting signatures of positive selection in specific cases.

Figure 2.5a shows the CHRNB1 locus in French Bulldogs, which produced the strongest signal of positive selection in H, H12, Tajima's D and π. The haplotype and SNP patterns around this locus provide a showcase example of a hard selective sweep. Diversity is depleted over >10 Mbp around the locus (**Figure 2.2**). On average, we would expect around 40 sites to be polymorphic over a window of the given size in this breed. However, we do not observe a single polymorphic site at this locus in our sample of 25 French Bulldogs. On average, we would also expect several haplotypes to be present, with the most common haplotype at around 40% frequency. As no site is polymorphic, we only observe a single haplotype. In contrast to the clear

signal identified by H, H12, CLR, Tajima's D and π at this locus, both iHS and nSL are unable to identify the sweep, consistent with the causal allele being fixed (**Figure 2.2**, **Table 2.3**).

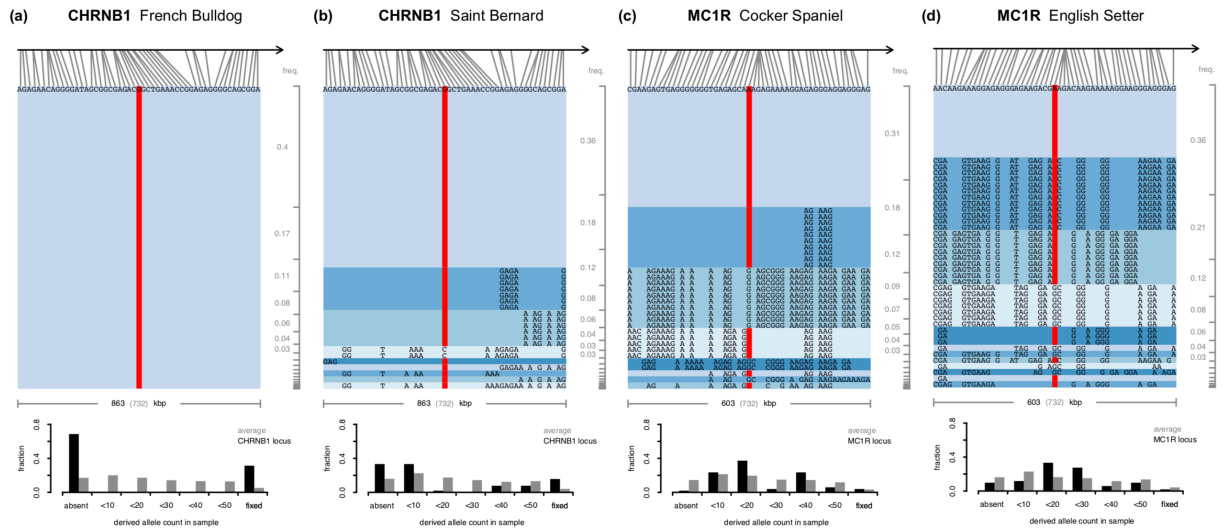


Figure 2.5. Positive selection produced both hard and soft selective sweeps. Haplotypes and SNP frequency distributions at specific loci in specific breeds. The top part of each panel shows the haplotypes in our sample from the particular breed over a window of 51 sites on our genotyping chip, centered on the causal mutation. The grey brackets on the right show the expected haplotype frequencies ordered by their prevalence in an average window of that size in the chromosome. Bar plots on the bottom of each panel show distributions of SNP frequencies in the window (black), compared with the chromosomal average (grey). The red bars indicate presence of the causal allele. (**a**) The CHRNB1 locus in French Bulldogs is a hard selective sweep that is fixed in our sample. None of the 51 sites is polymorphic at this locus and only a single haplotype is present. (**b**) In Saint Bernards, the causal mutation is not fixed in our sample. The most frequent haplotype is at higher-than-expected frequency, but several other haplotypes carrying the mutation are also present that may be variants of the major haplotype from recombination and/or mutation events. The SNP frequency spectrum shows the characteristic distortions towards high and low frequencies expected under a hard selective sweeps. (**c**) At the MC1R locus in Cocker Spaniel the causal mutation is present in 37 of the 50 genomes in our sample. The frequency of the most common haplotype, however, is not much higher than expected by chance and the SNP frequency spectrum is skewed towards intermediate frequencies, compatible with a soft selective sweep. (**d**) In English Setters, the MC1R locus shows even more pronounced signatures of a soft selective sweep.

In Saint Bernards, for comparison, the mutation at CHRNB1 is at high frequency but two genomes in our sample do not carry it (**Figure 2.5b**). Several haplotypes with the causal allele are present at the locus that may be variants of the major haplotype from recombination and/or mutation events early during its sweep (MESSER AND NEHER 2012). The SNP frequency spectrum shows the characteristic distortions of a hard selective sweeps and all scans detect signatures of positive selection at this locus (**Table 2.3**).

Figure 2.5c, d shows the MC1R locus in Cocker Spaniels and English Setters. Both breeds show signatures strongly suggestive of soft selective sweeps: The frequencies of the most common haplotypes are similar or lower to expectations in an average window, and in both breeds, several haplotypes carry the selected mutation. Importantly, some of these haplotypes differ at many sites from each other, including positions right next to the causal site, making it very unlikely that these haplotypes are in fact variants of the same haplotype that arose from mutation or recombination events during the sweep (MESSER AND NEHER 2012). Given that most pure dog breeds are <200 years old (PARKER *et al.* 2004; LARSON *et al.* 2012), yet some of these haplotype variants are quite common in the sample, it is also unlikely that they arose from recombination events after the sweep. Furthermore, the SNP frequency spectra are atypical for a hard selective sweep as they are skewed towards intermediate frequencies. All of these observations are more consistent with soft selective sweeps where positive selection has driven several haplotypes simultaneously, possibly

because selection acted on SGV. Both H12 and H detect signatures of positive selection at MC1R in Cocker Spaniel, other statistics are inconsistent and results strongly depend on window size. All statistics lack power in identifying signatures of positive selection at this locus in English Setters. Only H12, Tajima's D and π show some signal and only when using short analysis windows (**Table 2.3**).

Figure S2.4 shows haplotype patterns and SNP frequency spectra around the IGF1 locus in 12 different breeds. The selected mutation at this locus has been identified as a SINE element insertion in intron 2 of the IGF1 gene (RIMBAULT *et al.* 2013) that appears to be absent in grey wolves, most large dog breeds and all wild canids (GRAY *et al.* 2010). Hence, we do not expect that positive selection has acted on SGV at this locus, but rather that the selected SINE was a de novo mutation that arose during the domestication process. This is largely consistent with the haplotype and SNP frequency pattern in different breeds at this locus, which tend to show signatures of hard selective sweeps.

DISCUSSION

In our study, we examined the population genomic signatures observed around a set of 12 positive control loci known to have experienced positive selection in specific dog breeds due to their association with desirable morphological phenotypes. The dog system is extraordinary in that it provides a very large number of individual populations (breeds) for which we often know the specific selective pressures

experienced. In such a system, the most powerful selection scans should be those that can utilize the information provided by cross-population comparisons, for example FST-and XP-EHH-based methods (VITTI *et al.* 2013). We confirmed this intuition by showing that hapFLK, a powerful cross-population scan that uses haplotype information in an FST-based framework and incorporates information on the hierarchical structure between breeds, indeed identified all of our controls. However, for many other systems we may not have such cross-population information and will thus rely on scans that can detect signatures of selective sweeps from a single population sample.

Our approach of using positive controls in a real system is conceptually different from previous studies that evaluated the performance of selection scans based on computer simulations (TESHIMA *et al.* 2006; HUFF *et al.* 2010; POH *et al.* 2014; LOTTERHOS AND WHITLOCK 2015; SCHRIDER *et al.* 2015). These studies generally assume idealized evolutionary scenarios, such as panmixia, simplified demographic models and constant parameters over time and space, while interactions between selected sites such as background selection, Hill-Robertson interference and epistasis tend to be ignored. Unfortunately, we still lack a clear understanding of the importance of these effects and the extent to which they can obscure footprints of positive selection (BANK *et al.* 2014). In addition, many simulation studies assume that adaptation follows the classic selective sweep model. Whether this is an appropriate

model for describing adaptation in most biological systems is increasingly being questioned (PRITCHARD *et al.* 2010; CUTTER AND PAYSEUR 2013; MESSER AND PETROV 2013).

We found that artificial selection has indeed left detectable signatures in the polymorphism pattern around our positive control loci in purebred dogs. However, whether such signatures were detected varied widely between loci, individual breeds, the particular statistic used and the choice of analysis parameters. Interestingly, one of the most popular haplotype-based statistics, iHS, proved to be less accurate in identifying signatures of positive selection at our controls than the other statistics, including simpler haplotype-statistics such as H12 and H, as well as the frequency-based statistics CLR, Tajima's D and π. This could be due to a number of reasons: It is well known that iHS has difficulties identifying fixed sweeps because it requires the ancestral allele to be segregating in the population (SCHRIDER *et al.* 2015). We indeed observed that both iHS and nSL had particularly low power at those locus/breed combinations where the causal allele was fixed in our sample (**Table 2.4**). Furthermore, the generally high levels of LD in purebred dogs (SUTTER *et al.* 2004; LINDBLAD-TOH *et al.* 2005; BOYKO *et al.* 2010) could limit the sensitivity of haplotype-based statistics, as only extremely strong sweeps may be able to generate haplotypes that are even longer than those already present. Note, however, that two other haplotype-based statistics, H and H12, identified many positive controls.

The H12 statistic estimated over short windows of 25 segregating sites identified the largest number of positive controls in our study, followed by π and Tajima's D. This finding suggests that the signals of positive selection identified by these three statistics may be largely driven by the difference between the local density of SNPs on our genotyping chip (which we used for defining the window length for estimation of H12, π and Tajima's D) and the number of SNPs that are actually polymorphic in a particular breed in the given window.

Purebred dogs are clearly an exceptional system, characterized by strong artificial selection that is sometimes even repeatable between breeds (BOYKO *et al.* 2010). In addition, phenotypic variance for breed-defining morphological traits is often explained by surprisingly few mutations (RIMBAULT *et al.* 2013). As such, purebred dogs provide an excellent system for mapping the genetic basis of positively selected variants.

However, some aspects of our data set could confound the results in our study. First, because SNPs were obtained from a genotyping chip, rather than direct sequencing, they should be biased towards common variants, which might compromise the performance of frequency-based methods such as CLR and Tajima's D. In addition, the high levels of LD in dogs due to increased inbreeding could limit the power of haplotype-based methods. Dog breeds also vary in effective population size by several orders of magnitude (LEROY *et al.* 2013), overlapping the range

observed in smaller natural populations. In many ways, detection of selective sweeps in smaller populations is more difficult than in large populations as extensive drift can obscure and weaken the signatures of sweeps.

The severe bottlenecks during the breeding process could have systematically affected the patterns generated by positive selection, such as whether hard or soft sweeps should be more common. For example, recurring bottlenecks can have 'hardened' sweeps from SGV that were initially soft (WILSON *et al.* 2014). The mode and signatures of adaptation in large natural populations may therefore be quite different from those observed in purebred dogs and additional work is needed to evaluate the performance of methods for detecting selective sweeps in such populations.

ACKNOWLEDGEMENTS

The authors thank Juan Felipe Beltran for programming support and three anonymous reviewers for comments and suggestions. F.S. was supported by a Presidential Life Science Fellowship from Cornell University. P.W.M. was supported by start-up funds from Cornell University. A.R.B. was supported by the National Institute of General Medical Sciences and the National Institute of Aging of the National Institutes of Health under awards R01GM103961 and AG044284-01.

AUTHOR CONTRIBUTIONS

P.W.M. conceived the study. A.R.B. collected the samples and generated the genotyping data. F.S., P.W.M., J.M., R.S. and L.C. performed the selection scans. F.S. and P.W.M. performed the computational and statistical analyses. F.S., A.R.B. and P.W.M. wrote the manuscript. All authors approved the final version of the manuscript.

DATA ACCESSABILITY

Genotyping data were retrieved from (SHANNON *et al.* 2015) at Dryad http://dx.doi.org.proxy.library.cornell.edu/10.5061/dryad.v9t5h. Imputed data and complete results files have been archived at Dryad doi: 10.5061/dryad.hf46s.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Figure S2.1 Population tree of the 25 breeds inferred by hapFLK.

Figure S2.2 French bulldog scans using only SNPs that are segregating in the breed.

Figure S2.3 Selection scans in all breeds (similar to **Figure 2.2**).

Figure S2.4 Haplotype and SNP frequency pattern around IGF1 in 12 breeds.

Table S2.1 Distances (kb) to closest data point above 95% quantile.

CHAPTER 3

DENSE TIME-COURSE GENE EXPRESSION PROFILING OF THE *DROSOPHILA MELANOGASTER* INNATE IMMUNE RESPONSE[2]

INTRODUCTION

Upon microbial infection, *Drosophila* launch rapid and efficient immune responses that are crucial to survival. However, a spurious or over-activated immune response can be harmful to the organism. An excessive or overly prolonged immune response can lead to metabolic dysregulation, causing wasting in mammals and flies (FITZPATRICK AND YOUNG 2013). Furthermore, immune responses are energetically costly (LAZZARO AND GALAC 2006) because they draw resources from other physiological processes (ZEROFSKY *et al.* 2005; DIANGELO *et al.* 2009) such as metabolism, reproduction, and environmental stress responses. It has been shown that allocating resources to the immune system reduces resources for reproduction (MCKEAN *et al.* 2008; HOWICK AND LAZZARO 2014), and the opposite is also true, where mating reduces survivorship after infection and decreases resistance to infection (FEDORKA *et al.* 2007; SHORT AND LAZZARO 2010; SHORT *et al.* 2012). This represents a type of trade-off where both immune response and reproduction are

[2] Manuscript in preparation: Schlamp F, Early A, Wells MT, Basu S, and Clark AG. Dense time-course gene expression profiling of the *Drosophila melanogaster* innate immune response.

costly, and limited resources need to be allocated to either one or the other (SCHWENKE *et al.* 2016). Therefore, we expect that natural selection will act to tune their immune response to strike a balance between the advantage of a rapid and robust ability to fight infection, and the costly side-effects of an over-prolonged or unnecessary immune response. This tuning is likely to be mediated through a series of regulatory and feedback properties of the immune system of the fly.

While gene expression has been examined at several time points after infection in *Drosophila* (DE GREGORIO *et al.* 2001; BOUTROS *et al.* 2002; SACKTON *et al.* 2010), the dynamics of this immune response have not yet been studied with high temporal resolution. Such a high-resolution time-course analysis can help profile with more certainty the types of expression dynamics that different genes and pathways undergo after infection. Dense and extended time-course sampling of gene expression of the immune response can allow us to distinguish between transient and sustained expression patterns, where expression of genes with a transient response to perturbation will return back to normal after a certain period of time, while expression of genes with a sustained response will remain at a different level of expression compared to pre-perturbation levels. This kind of temporal profiling of the immune response can also suggest candidates to examine for possible interactions and trade-offs between the immune response and other physiological processes in the form of regulatory networks.

Inference of gene regulatory networks from time-course gene expression data arises in many contexts in functional genomics and bioinformatics. While several currently existing methods to analyze static RNA-seq data - such as edgeR (ROBINSON *et al.* 2010) or DESeq2 (LOVE *et al.* 2014) - have been utilized to analyze time-course data, they are not ideal for dealing with time-course RNA-seq data for many reasons, as reviewed in (BAR-JOSEPH *et al.* 2012) and (SPIES AND CIAUDO 2015). For example, most methods do not take into consideration the correlation of genes in adjacent time points, which leads to many temporal patterns in expression to not be taken into account for normalization and differential expression analysis (SPIES AND CIAUDO 2015). New approaches for analyzing time-course data, like those introduced in this paper, are essential to reveal dynamic behaviors in organisms and discover regulatory interactions among genes.

In this study, we performed a dense time-course RNA-seq analysis of the *Drosophila* transcriptional response to immune challenge to better understand the dynamics of activation and resolution of the innate immune response. The goal of this RNA-seq experiment was to stimulate a full but transient immune response in *Drosophila* and follow the dynamics in gene expression through time. Flies were sampled over 5 days generating a total of 20 time points post-infection with an additional time point pre-injection as a control. We analyzed the resulting longitudinal RNA-seq dataset using a broad range of statistical methods. We use gene-wise linear models to fit cubic splines with time, and standard empirical Bayes *F*-tests to select

genes whose expression levels were significantly altered across the time course. Additionally, we find strong temporal patterns of transient and sustained responses to infection that occur over different time scales using clustering analysis, and we further identify non-immune expression dynamics of *Drosophila* reproducing the well-characterized cyclic patterns in gene expression of the circadian rhythm. We also performed gene set analysis to detect pathway-specific expression patterns and constructed networks of multivariate Granger causality (GC) relationships (GRANGER 1969) among subsets of DE genes. Our analyses provide several novel functional annotations for previously uncharacterized genes, identify different types of transcriptional dynamics, and suggest new interactions governing temporal gene regulation of the immune response. Throughout all of these analyses we see a continued theme of interplay and trade-off between the immune response and other canonically separate pathways.

MATERIALS AND METHODS

Fly lines, injections, and sample collection

Male adult *Drosophila* of about 4 days old from an F1 cross from two *Drosophila melanogaster* Genetic Reference Panel (DGRP) lines: line 379, which has shown to have low bacterial resistance, and line 360, which has high bacterial resistance (EARLY *et al.* 2017b). Flies were kept on a 12:12 dark-light cycle.

Flies were injected in the abdomen with 9.2 µl of commercial lipopolysaccharide (LPS) (*Escherichia coli* 055:B5 Sigma) derived from the outer membrane of Gram-negative bacteria. LPS is a known non-pathogenic elicitor used to stimulate a full but transient immune response in *Drosophila* (IMLER *et al.* 2000; LEULIER *et al.* 2003). Using commercial LPS instead of living bacteria also gives the advantage of avoiding the confounding effects from the mechanisms the bacteria uses to circumvent immune responses (GRAHAM *et al.* 2011). While it is now argued that purified LPS by itself does not induce an immune response in *Drosophila*, it has been shown that commercial 'crude' LPS preparations do (IMLER *et al.* 2000; LEULIER *et al.* 2003; KANEKO *et al.* 2004; HANDU *et al.* 2015), most probably due to contaminating peptidoglycan in the latter (KANEKO *ET AL.* 2004). For this reason, commercial LPS was chosen for this study, and its ability to induce an immune response was confirmed using qPCR, as explained in the next section.

Flies were injected using a Nanoinjector (Nanoject II, catalog #3-000-204, Drummond), which allows high-throughput fly injections with a constant injection volume. Injections were performed in the abdomen, as it has been shown to be less detrimental to the fly compared to thorax injury (CHAMBERS *et al.* 2014).

Flies were sampled for a total of 21 time points throughout the course of five days, which includes an uninfected un-injected sample as control at time zero, and 20 time points after infection (1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 20, 24, 30, 36, 42, 48, 72, 96, 120 h). This sampling was performed in two blocks, using flies from the same

stock, in two consecutive days. Therefore, all samples have two replicates, giving a total of 42 samples. During collection, a group of ~10 pooled flies corresponding to the sampled time point were flash frozen in dry ice and stored at -80 C for later RNA extraction.

Experimental validation using qPCR

The immune inducibility of commercial LPS was confirmed using qPCR. Adult male *Drosophila* were injected with 9.2 µl or 40 µl of 1 mg/mL LPS and flash frozen at 8 and 24 h for RNA extraction. Uninfected un-injected flies were used as control. Each sampled time point consisted of a group of ~10 pooled flies. Each sample had two replicates. Genes *AttA* and *DptB* were measured to confirm immune inducibility. Gene *Rp49* was used as a baseline for expression normalization. Results showed a significant up-regulation of *AttA* and *DptB* at both volumes (9.2 µl and 40 µl) for both time points (8 and 24 h). We decided to use 9.2 µl so as to cause the least amount of disruption to flies during infections, while still eliciting an immune response.

RNA extraction, RNA sequencing, and quality control filtering

RNA extraction was performed using Trizol (Life Technologies) following the manufacturer's instructions. cDNA libraries were prepared using the TruSeq RNA Sample Preparation Kit (Illumina). RNA purity was assessed using a Nanodrop

instrument. RNA concentration was determined using a Qubit (Life Technologies) instrument. Sequencing was performed on an Illumina Hi-Seq 2500, single-end, and a read length of 75 bp, at Cornell Biotechnology Resource Center Genomics Facility.

Samples had an average of 24.8 M raw reads. Samples went through quality control using FastQC (version 0.11.5) (ANDREWS 2010). Truseq adapter sequences were removed from any sample that showed any level of adapter contamination using cutadapt (version 1.14) (MARTIN 2011). Low quality bases in the beginning and end of the reads were trimmed using fastx_trimmer (version 0.0.13, http://hannonlab.cshl.edu/fastx_toolkit/). Reads were mapped to the *Drosophila melanogaster* genome (r6.17) using STAR (version 2.5.2b) (DOBIN *et al.* 2012). BAM files were generated using SAMtools (version: 1.3.2) (LI *et al.* 2009). Only one sample (4B, at 3 h) out of the original 42 failed to pass the quality thresholds, and all subsequent analysis used the remaining 41 samples. An average of 92.97% reads per library mapped uniquely to the *Drosophila melanogaster* genome. We ended up with an average of 23.4 million uniquely mapped reads per library.

Reads mapping to genes were counted using the R package *GenomicAlignments* (LAWRENCE *et al.* 2013). Genes with zero counts across all samples were removed (923 genes out of 17,736). Samples were normalized to library size. A "+1" count number was added to all genes before performing $\log_2$ transformation, to make sure values after transformation are finite, and stabilize the variance at low expression end. After normalization and $\log_2$ transformation, only genes with more than 5 counts in at

least 2 samples were kept (removing 4,156 genes). We ended up with 12,657 genes for downstream analysis.

Principal components analysis

Principal components analysis (PCA) was performed using function *plotPCA* from the R package *DESeq2* (LOVE *et al.* 2014) after regularized-logarithm transformation of raw counts, using the design ~time+time:time to create the DEseqDataSet. Genes with zero counts across all samples were first removed. The default number of 500 top genes with highest row variance was used to calculate the principal components.

Differential expression analysis

In order to identify genes that had differential expression over the time course, we adopted the linear model-based methodology proposed in (LAW *et al.* 2014) and available in the R package *limma*. We first transformed the normalized RNA-seq read counts (before $\log_2$ transformation) using the *voom* transformation, which estimates the heteroscedastic mean variance relationships of log-counts and adds a precision weight to each observation to make them amenable to the usual linear modeling pipelines that rely on normality. We used gene-wise linear models to fit cubic splines (with 3 degrees of freedom) with time, TMM normalization method (ROBINSON AND OSHLACK 2010), and standard empirical Bayes *F*-tests to select genes whose

expression levels were significantly altered across the time course in both replicates. Next, we checked for differential expression of every gene between time point 0 (control) and time point *t*, for $t = 1, 2, \ldots, 48$ h. For each test, a multiple testing correction at 5% False Discovery Rate (FDR) using the Benjamini-Hochberg method (BENJAMINI AND HOCHBERG 1995) was adopted. Venn diagrams to compare results were adapted from those generated using web tool Venny (http://bioinfogp.cnb.csic.es/tools/venny/) (Oliveros 2007).

GO enrichment

Gene Ontology (GO) enrichment analysis was performed using PANTHER Statistical Overrepresentation Test (http://pantherdb.org/, version 14.1, released 2019/04/29) (MI *et al.* 2018) using default settings (GO-Slim Biological Process annotation data set, Fisher's Exact test, FDR < 0.05).

Detecting cyclic gene patterns

Cyclic gene patterns were identified using the JTK_Cycle algorithm (HUGHES *et al.* 2010) available in R package *JTK_Cycle*. Nine regularly distributed time points were subset from both replicates every 6 hours (0, 6, 12, 18, 24, 30, 36, 42, 48 h). The time point corresponding to 18 h was approximated by averaging normalized gene counts between time points 16 and 20 h. We looked for rhythms between 18-30 h (4 to 6 time points) at an adjusted *P*-value < 0.01.

Temporal clustering

Temporal clustering was performed using the R package *TSclust* (MONTERO AND VILAR 2014). Normalized counts of both replicates were clustered using dissimilarity measures from Autocorrelation-based method (ACF), which computes the dissimilarity between two time series as the distance between their estimated simple autocorrelation coefficients (GALEANO AND PEÑA 2000). This method was used with a *P*-value cutoff of 0.05.

Gene set analysis

Gene set analysis was done using the R package *GSA*, which uses a Gene Set Analysis algorithm (EFRON AND TIBSHIRANI 2007) that improves the GSEA algorithm (SUBRAMANIAN *et al.* 2005) by allowing testing for associations between gene sets and time-dependent variables (EFRON AND TIBSHIRANI 2007; MULLIGHAN *et al.* 2009). Gene set membership was assigned from GO data downloaded from FlyBase.org in January 2019. Normalized counts for both replicates at each time point from 1 to 120 h were compared against both control replicates (0 h), using a two-class paired vector (-1, 1, -2, 2) which corresponds to (control_replicateA, timepointX_replicateA, control_replicateB, timepointX_replicateB). We used 100,000 permutations to estimate false discovery rates. Only pathways with *P*-values below 0.05 and with 5 or more genes from our full dataset were kept. A subset of most relevant pathways was compiled by selecting pathways that had at least one gene from

the subset of 551 most predominant time-dependent genes, and had a score of 2.5 of more in at least one time point from 1 to 48 h. This gave us 41 unique pathways as shown in **Figure 3.9**.

Network inference

Granger causality-based methods (GRANGER 1969) were used to construct putative interaction networks among genes in the form of directed graphs with individual genes as nodes. A directed edge from gene *A* to gene *B* is added if the time course of gene *A* Granger-causes the time course of gene *B*. The notion of 'Granger causality' is popular in learning lead-lag relationships among two or more time series. Formally, if the time series of gene *A*, given by x_t, has some power in predicting the expression of gene *B* at time $t+1$, called y_{t+1}, over and above y_t and conditioned on an information set I_t, then gene *A* is said to exert a *Granger causal* effect on gene *B*. Bivariate Granger causality uses a small information set $I_t = \{x_{1:t}, y_{1:t}\}$ and captures Granger causal relationship from gene *A* to gene *B* by testing whether the regression coefficient in the following bivariate regression is different from zero:

$$y_{t+1} = \alpha y_t + \beta x_t + error_{t+1}$$

A master set of 258 genes was constructed from the 551 predominant time-dependent genes by picking those that had available functional annotation and that had differential expression of at least absolute log fold change of 1. Using linear

regression (function lm() in R), we conducted bivariate (pairwise) Granger causality tests for every pair of genes among this set of 258 genes using data on sliding windows of $t = 6$ consecutive time points and the two replicates (sample size = 12), and ranked them in order of increasing *P*-values (BH method used for calculating FDR), keeping the top resulting edges (BHFDR < 0.05%).

A well-known critique of bivariate Granger causality is its use of a small information set that does not contain any other factors except genes *A* and *B* (MUKHOPADHYAY AND CHATTERJEE 2006). This failure to account for other potential confounding variables can give rise to many spurious edges in our network (MUKHOPADHYAY AND CHATTERJEE 2006), where Granger causal effects from gene *A* to gene *B* is an artefact of gene *C*, which is causal for one or both genes. To address this, we adopted multivariate (or network) Granger causality (BASU *et al.* 2015), allowing us to avoid such spurious inferences through multiple linear regression. In this framework, we start with p genes, and Granger causal relationship of Gene *A* on Gene *B* is tested by regressing y_{t+1} on y_t, x_t and the time courses of the other $p-2$ genes $z_{1t}, z_{2t}, \ldots, z_{pt}$.

$$y_{t+1} = \alpha y_t + \beta x_t + \gamma_1 z_{1t} + \gamma_2 z_{2t} + \cdots + \gamma_{p-2} z_{p-2,tt} + error_{t+1}$$

For small sample size and large p, the above regression is not possible to run using ordinary least squares (OLS), so we use LASSO (TIBSHIRANI 1996) regression. To test if the regression coefficient β in the above regression is different from zero,

we used two different variants of de-biased LASSO (JAVANMARD AND MONTANARI 2014; DEZEURE *et al.* 2015), each of which corrects the bias of lasso and allows quantifying uncertainty of regression coefficients one at a time. A non-zero coefficient β in the above multivariate regression suggests that gene *A* is Granger causal for gene *B*, even after accounting for the effects of the other p-2 genes. Using this method on the master set of 258 genes, we reconstructed putative directed networks of multivariate Granger causality and ranked the edges in increasing order of *P*-values, following the same parameters used in the bivariate (pairwise) Granger causality method (sliding window of 6 consecutive time points in both replicates, keeping the top resulting edges (BHFDR < 0.05%)).

RESULTS

High-resolution profiling of gene expression after immune challenge

To generate a full transcriptional profile of gene expression dynamics in *Drosophila melanogaster* after immune challenge, we injected adult male flies with commercial lipopolysaccharide (LPS), a known non-pathogenic elicitor that can stimulate a full yet transient immune response (IMLER *et al.* 2000; LEULIER *et al.* 2003), while avoiding the confounding effects from a growing and changing population of pathogens. Flies were sampled for a total of 21 time points throughout the course of five days, which includes an uninfected un-injected sample as control at time zero, and

20 time points after infection. Since this is a perturbation-response experiment, denser sampling occurred at early time points (BAR-JOSEPH *et al.* 2012), with the first 13 time points taken within the first 24 h (1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 20, and 24 h). Sampling is also essential at late time points in order to capture gene expression patterns, to know how long it takes to return to 'normality', and to differentiate between transient and sustained responses (BAR-JOSEPH *et al.* 2012). For this reason, sampling continued until day 5 after infection, although more sparsely (30, 36, 42, 48, 72, 96, 120 h) (**Figure 3.1A**). All samples have two replicates, giving a total of 42 samples. Sampled flies were flash frozen to isolate mRNA for RNA-seq analysis as described in the **Materials and Methods**. This yielded 41 high-quality libraries with an average of 23.5 million mapped reads per sample. After normalization of libraries, only genes with more than 5 counts in at least 2 samples were kept, leaving 12,657 genes for further analysis.

Principal components analysis (PCA) of all time points reveals a horseshoe temporal trend, with the control samples clustering in the middle, and all the post-infection timepoints following a horseshoe-shaped track, consistent with a pattern of many genes displaying a coordinated change over the five-day interval (**Figure 3.1B**). This type of "horse-shoe" or arch temporal trend in PCA has been seen in other time-series experiments (DENG *et al.* 2014; LAW *et al.* 2014; BENDJILALI *et al.* 2017; WHITE *et al.* 2017), and is commonly seen in spatial population genetic variation (NOVEMBRE AND STEPHENS 2008) and in ecological gradient data that varies in a non-linear

manner (PODANI AND MIKLÓS 2002). One possible explanation for this pattern is that PCA ordination gets distorted as it tries to fit a nonlinear relationship to an underlying assumption of linearity (CLAPHAM 2011). PC1, PC2, and PC3 captured 35, 15, and 14.5% of the variance in gene expression respectively, and the first six PCs account for over 80% of the total variance in the data.

Proper normalization of the data was confirmed by confirming the behavior of known *Drosophila* housekeeping genes across time (Qiagen Housekeeping Genes RT^2 Profiler PCR Array and (LÜ *et al.* 2018)). As expected, housekeeping genes showed little change across time (**Figure 3.1C**). The success of the immune challenge was confirmed by the immediate up-regulation of known immune response genes within the first time points (**Figure 3.1D**).

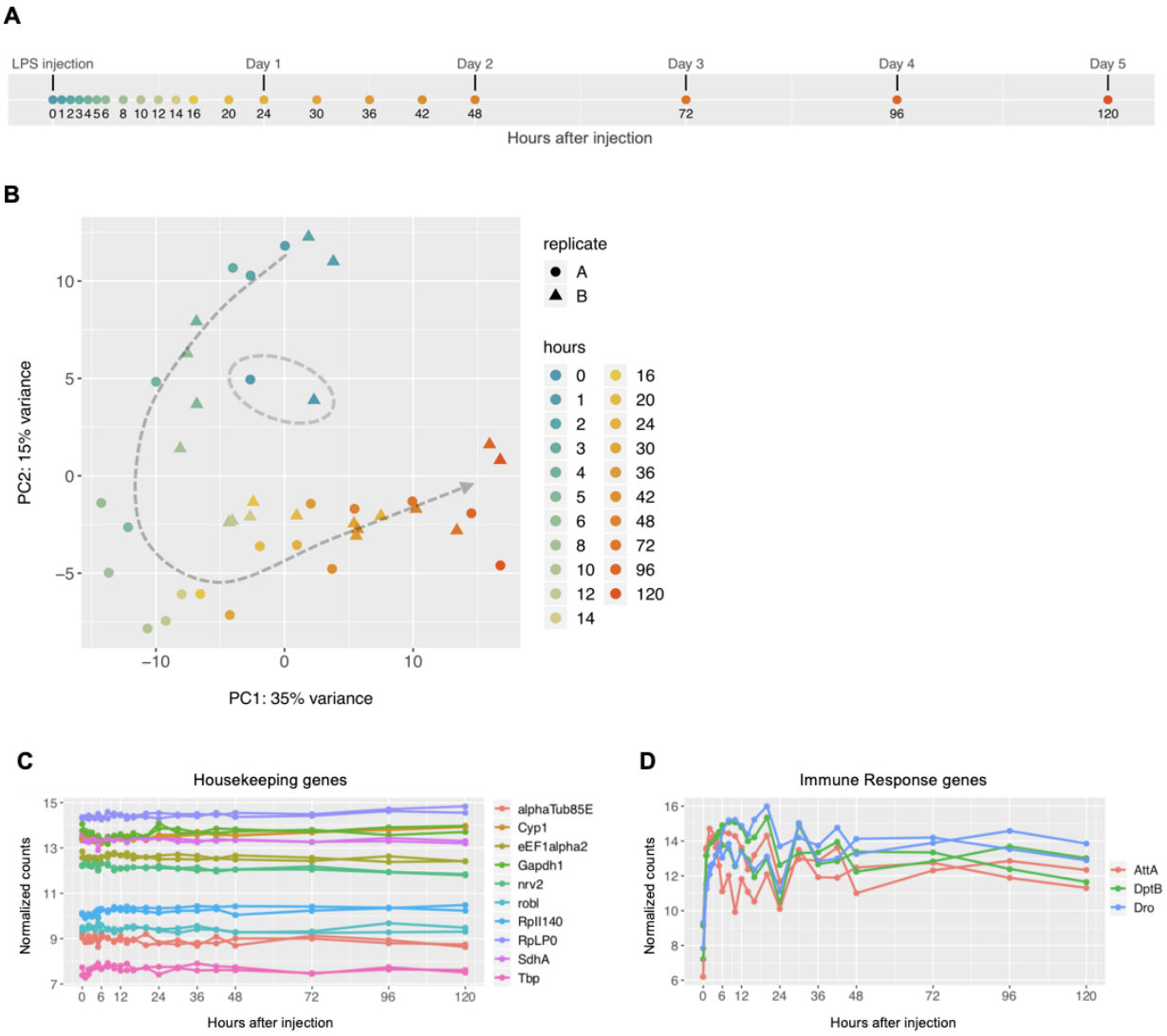


Figure 3.1. **Transcriptional profiling of *Drosophila* immune response.** (**A**) Timeline of 21 time points, including un-infected un-injected sample as control at time 0. Sampling was denser in the first 24 h and continued -although more sparsely- until day 5 (120 h). (**B**) Principal component analysis (PCA) of all time points shows a coordinated change of gene expression over five days. Both replicates are shown for all samples except for the time point at 3 h, where one replicate was excluded from the analysis during RNA-seq data processing. The two samples in blue clustering in the middle (marked with grey dashed circle) correspond to the control time point (0 h). All other time points from 1 to 120 h show a horseshoe temporal pattern around the controls. PC1 and PC2 captured 35 and 15% of the variance in gene expression, respectively. Plots of normalized counts of housekeeping genes (**C**) show little change across time as expected under proper data normalization, while immune response genes (**D**) show up-regulation within the first time points, as expected after a successful immune challenge.

Continuous and pairwise methods identify time dependent genes

First, we wanted to identify genes whose expression levels were significantly altered across the time course. To accomplish this, we applied a precision weight-based transformation 'voom' on the normalized data using R package *limma* (LAW *et al.* 2014), we then used gene-wise linear models to fit cubic splines with time, and standard empirical Bayes *F*-tests to select genes with expression levels that significantly change throughout time, as described in the **Materials and Methods**. We performed a multiple testing correction at 5% False Discovery Rate (FDR) using the Benjamini-Hochberg method (BENJAMINI AND HOCHBERG 1995). The time point at 3 h was removed in the analysis due to lack of replication.

This analysis identified 380 genes with significant changes in expression through the first 48 h, and 48 genes with significant changes in expression through the first 8 h (**Figure 3.2A**). The smoothing applied by continuous frameworks (REINSCH 1967) such as splines or quadratic trends are expected to miss some intricacies of the temporal expression pattern of genes, but they are adequate for modelling general trends. Long time spline fit on the first 48 h can detect gradual changes and 'global' patterns, but misses early impulse patterns, such as those observed in known immune response genes such as *AttA* and *DptB* (**Figure 3.2C**). On the other hand, short spline fit over the first 8 h can accurately identify early impulse patterns, but will not be able to identify patterns of expression that alter later in time, such as the ones shown by genes *Gale* and *Galk* (**Figure 3.2D**).

Next, we characterized the behavior of expression temporal patterns by estimating the differential expression of every gene at each time point, from 1 to 48 h, compared to the un-infected un-injected control samples at time zero. Pairwise comparisons were done on normalized counts of both replicates, as described in **Materials and Methods**. This method identified 729 differentially expressed (DE) genes that were significantly (FDR < 0.05) up- or down-regulated by an absolute $\log_2$-fold change of at least 1 (which corresponds to a 2-fold change in expression) in at least one time point throughout the first 48 h after injections. Within this gene set, there were 214 genes that were up- or down-regulated by an absolute log2-fold change of at least 2 (4-fold change in expression) (**Table S3.1**). **Figure 3.2B** shows the overlap between these sets and those identified by spline modeling.

We combined the set of genes identified by the spline modeling over 48 and 8 h (411 genes) and the 214 DE genes identified by pairwise differential expression to compile a subset of 551 most predominant time-dependent genes (**Table S3.2**).

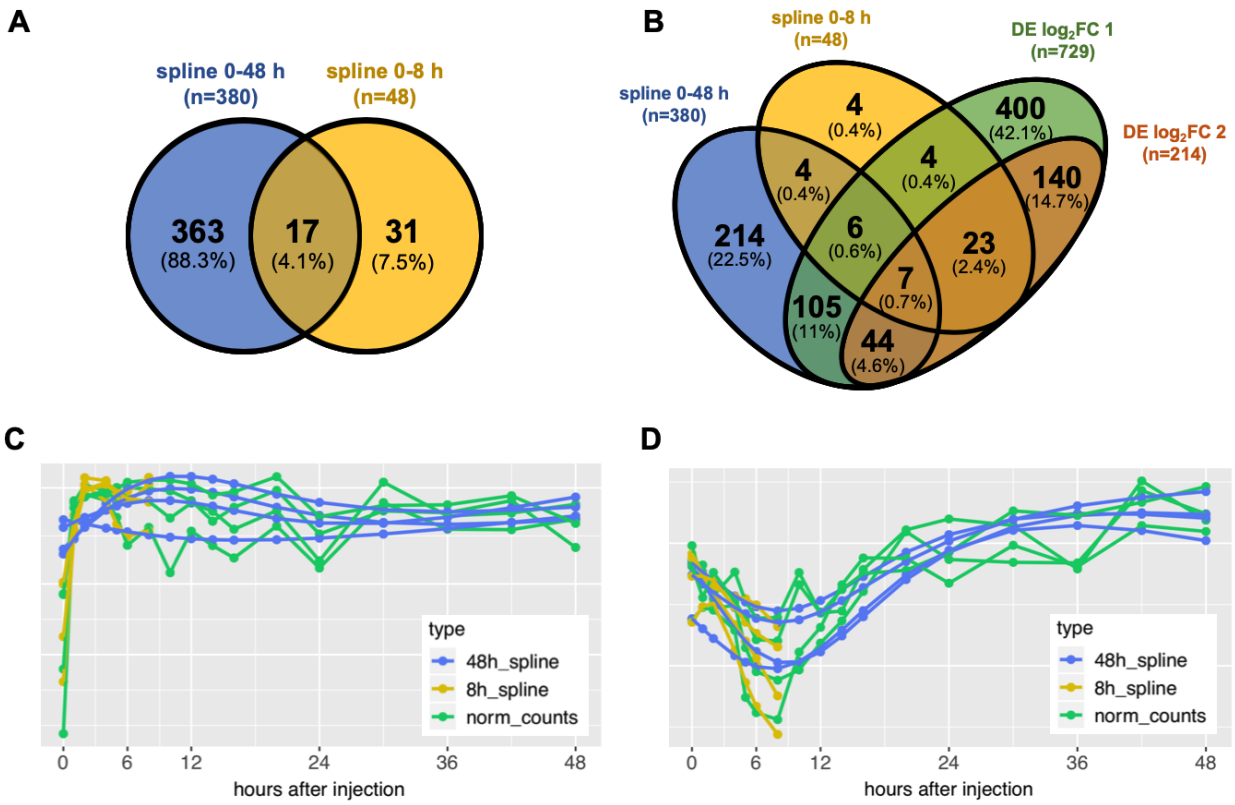


Figure 3.2. Identification of time-dependent genes. (**A**) Genes that significantly change in expression across time according to spline analysis in first 48 h (blue) vs 8 h (yellow). (**B**) Comparing results from spline analysis (over 48 h in blue and over 5 h in yellow) vs. results from differential expression analysis (absolute log_2FC of 1 in green and log_2FC of 2 in orange) at FDR < 0.05. (**C**) Spline modeling of two immune genes (*AttA* and *DptB*) when using first 48 h (blue) and first 8 h (yellow) compared to pattern of normalized counts (green), spline modeling over 48 h smooths out the early impulse signal. (**D**) Spline modeling of genes *Galk* and *Gale* when using first 48 h (blue) and first 8 h (yellow) compared to the pattern of normalized counts (green), spline modeling over 8 h misses the main change in pattern.

Global dynamics show different patterns of expression and biological function

The 551 most predominant time-dependent genes show four main hierarchical clusters of gene expression profiles across time (**Figure 3.3A**). Clusters 1 and 4 are characterized by an initial decrease in expression followed by an increase in expression after 8 and 3 h respectively, out to 5 days after injection (**Figure 3.3B**), with cluster 1

showing a stronger decrease in expression in the early hours after injection. These clusters have a significant enrichment of Gene Ontology (GO) terms for various biosynthetic, catabolic, and metabolic processes (**Figure 3.3C**). Clusters 2 and 3 both have a strong increase in expression after injection (**Figure 3.3B**), and show significant enrichment of GO terms for multiple immune and stress response related processes, and abiotic stimulus response, respectively (**Figure 3.3C**). Cluster 2 has a more immediate increase in expression following injection, reaching a maximum peak within the first 2 h (**Figure 3.3B**) and contains immune response genes Attacins (*AttA*, *AttB*, *AttC*) and Cecropins (*CecB*, *CecC*), as well as Heat Shock protein family genes (*Hsp70Aa*, *Hsp70Ab*, *Hsp70Ba*, *Hsp70Bb*, *Hsp70Bbb*, *Hsp70Bc*) which are known to protect cells from high temperatures and other forms of stress, but also play many roles in the immune system (BINDER 2014). Cluster 3, on the other hand, reaches a maximum expression later at around 9 h (**Figure 3.3B**) and contains the Immune-induced peptide family (*IM1*, *IM2*, *IM3*, *IM4*, *IM14*, *IM23*, *IMPPP*) and other immune response related genes, as well as genes from the Turandot family (*TotA*, *TotB*, *TotC*, *TotM, TotX*) which are involved in humoral stress response and can be induced under several stress conditions, such as bacterial challenge, high temperature, mechanical pressure, among others (EKENGREN *et al.* 2001).

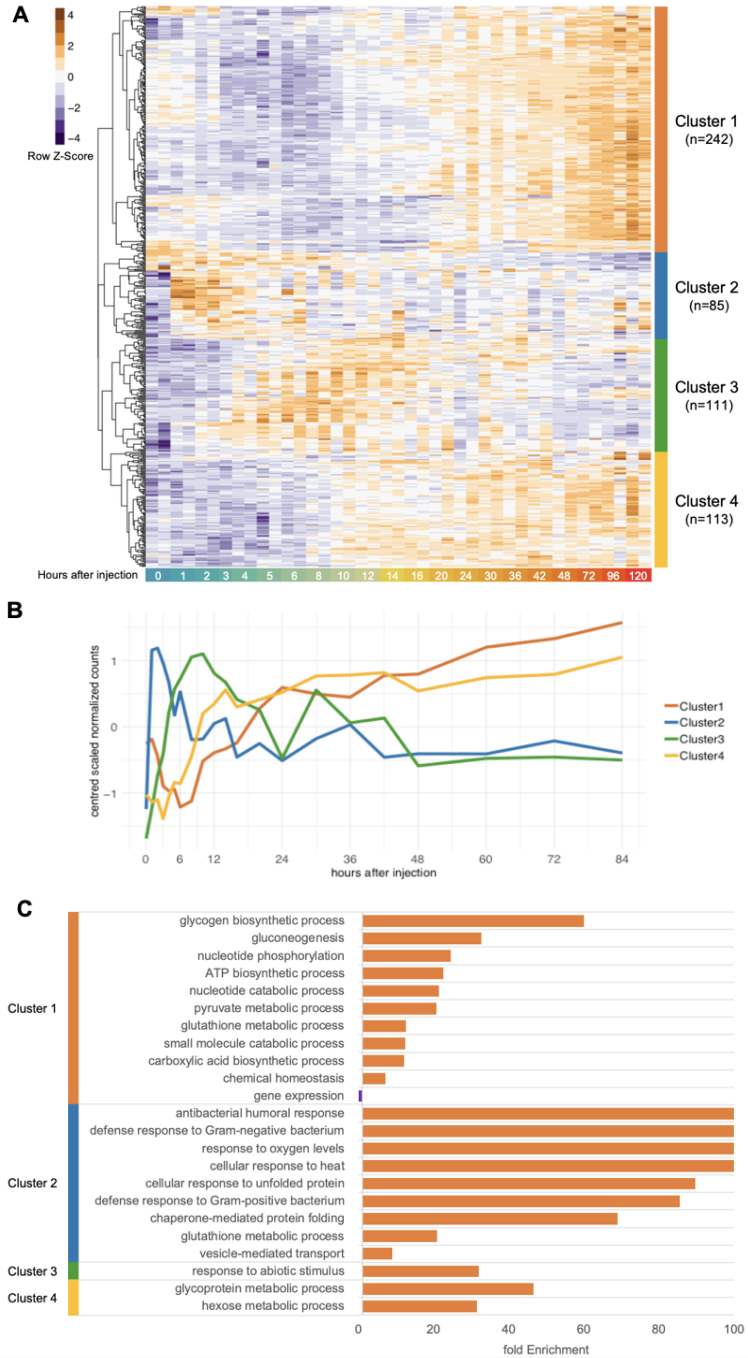


Figure 3.3. Global dynamics of time-dependent genes show divergent patterns of expression. (**A**) Heatmap of 551 most predominant time-dependent genes, identified by spline modeling over 48 and 8 h (FDR < 0.05) and pairwise differential expression (absolute $log_2FC > 2$ and FDR < 0.05). Hierarchical clustering of the genes shows four main clusters characterized by time points in which the genes reach maximum and minimum expression across time. Z-score values of each gene are shown from dark purple (minimum expression across time) to dark orange (maximum expression across time). (**B**) Mean patterns of expression across time for genes within each of the four main clusters, displayed by their centered and scaled normalized counts. (**C**) Significant Gene Ontology terms (FDR < 0.05) for overrepresented Biological Processes at each cluster.

GO analysis on the 214 top DE genes shows a significant overrepresentation of immune response related genes, such as Attacins (*AttA*, *AttB*, *AttC*), Diptericins (*DptA*, *DptB*), Cecropins (*CecB*, *CecC*), Immune-induced peptides (*IM1*, *IM2*, *IM3*, *IM4*, *IM14*, *IM23*, *IMPPP*), *Drosocin* (*Dro*), *Drosomycin* and Drosomycin-like genes (*Drs*, *Drsl1*, *Drsl2*, *Drsl3*), *Metchnikowin* (*Mtk*), Peptidoglycan Recognition Proteins (*PGRP-SB1*, *PGRP-SD*), *Diedel*, *Relish* (*Rel*), *elevated during infection* (*edin*), among others. This confirms that the organism is having an immune response to the commercial LPS injections, as it is consistent with known gene expression profiles of immune response deployment in *Drosophila* (DE GREGORIO *et al.* 2001; BOUTROS *et al.* 2002). Among these 214 top DE genes we also find genes related to other stress response pathways, such as Turandots (*TotA*, *TotC*, *TotM*) and Heat Shock proteins (*Hsp70Aa*, *Hsp70Ab*, *Hsp70Ba*, *Hsp70Bb*, *Hsp70Bbb*, *Hsp70Bc*). A heatmap of the $\log_2$-fold change in expression of all 214 top DE genes can be found in **Figure S3.1**.

Next, we further filtered these 214 top DE genes using more stringent cutoffs to identify core DE genes for additional characterization. This results in a core of 91 genes that are significantly (FDR < 0.01) up- or down-regulated by an absolute $\log_2$-fold change > 2 in at least two time points throughout the first 48 h after injection. A heatmap of the $\log_2$-fold change of these 91 core DE genes from 1 to 48 h can be seen in **Figure 3.4A**. The distribution of all significantly up- and down-regulated genes at each timepoint can be seen **in Figure 3.4B**. Many of the up-regulated genes at each timepoint are also known immune genes, as identified by a list of immune

genes curated in (EARLY *et al.* 2017a). The number of up-regulated genes is much higher than the number of down-regulated genes across all timepoints. The bottom of the heatmap in **Figure 3.4A** shows a cluster of the most up-regulated genes, composed of *DptB*, *AttC*, *Mtk*, *Dro*, *CR45045*, *DptA*, *CG43920*, and *edin*. These are mostly immune-related genes that are strongly up-regulated in early timepoints after infection, and remain elevated by approximately 32-fold 48 h later (**Figure 3.4C**).

Within these 91 core DE genes, we also find circadian rhythm genes *period* (*per*), *timeless* (*tim*), *takeout* (*to*), and *vrille* (*vri*), which when plotted against time exhibit the classic 24 h periodic expression of the circadian rhythm (**Figure 3.4D**). Features like these serve to validate the normalization and differential expression analysis of this dataset, demonstrating that this time-course profiling is accurately identifying previously well characterized temporal patterns.

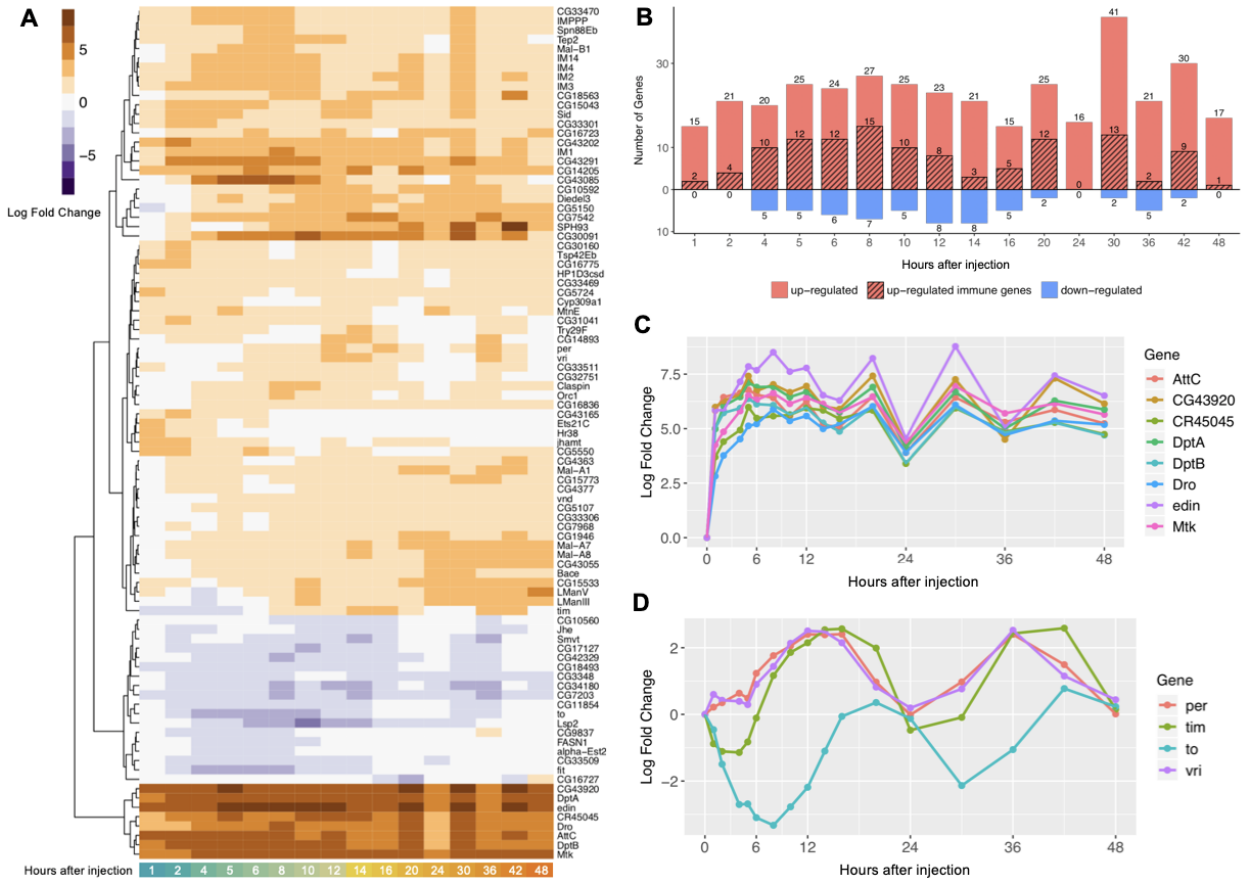


Figure 3.4. Characterization of top DE genes. (**A**) Heatmap of gene expression changes. Up-regulated genes in orange, down-regulated genes in purple. FDR correction of 0.01, absolute $\log_2$-fold change cutoff of 2 in at least two time points. 91 genes total, across 48 h. Genes ordered using Euclidean distance. (**B**) Distribution of genes significantly up- and down-regulated at each timepoint (in red and blue, correspondingly) and how many of those are immune genes (shaded over red), no down-regulated immune genes were observed. (**C**) Temporal dynamics of gene expression of the most up-regulated cluster of genes in heatmap (*DptB*, *AttC*, *Mtk*, *Dro*, *CR45045*, *DptA*, *CG43920*, and *edin*), first 48 h after injection. (**D**) Temporal dynamics of gene expression of circadian rhythm genes (*per*, *tim*, *to*, *vri*) show a classic and well characterized 24 h periodic expression.

Transcription Factors are differentially expressed

Among the top 214 DE genes, we find 6 transcription factors (TFs), identified from FlyTF database. Four well characterized TFs: *Ets at 21C (Ets21C)*, *Hormone receptor-like in 38* (*Hr38*), *Relish* (*Rel*), and *stripe* (*sr*) show a fast impulse of up-regulation immediately following injection, reaching maximum expression in the first hour

(**Figure 3.5A**). *Ets21C* is a stress inducible TF and *Relish* is a downstream component of the immune deficiency (imd) pathway, which regulates antibacterial response (MYLLYMÄKI *et al.* 2014; MUNDORF *et al.* 2019). *Hr38* and *stripe* are the two most robust activity-regulated genes (ARGs, defined as genes that are rapidly induced with neuronal activity, mostly within an hour) in *Drosophila* (CHEN *et al.* 2016b). In mammals, ARGs (also known as immediate-early genes, IEGs) are induced rapidly and transiently upon stimulation in neurons, and are usually enriched for TFs which trigger secondary transcriptional responses (CHEN *et al.* 2016b). On the other hand, candidate TF *Origin recognition complex subunit 1* (*Orc1*) has a later up-regulation, reaching maximum expression at hour 8 (**Figure 3.5A**). *Orc1* codes for a component of ORC, which binds origins of replication and is essential for gene amplification and cell proliferation. TF *vrille* (*vri*) and candidate TF *cryptochrome* (*cry*) are known for their circadian rhythm functions (CYRAN *et al.* 2003; COLLINS *et al.* 2006), and their 24 hour circadian oscillations in RNA level are recapitulated in our analysis (**Figure 3.5B**).

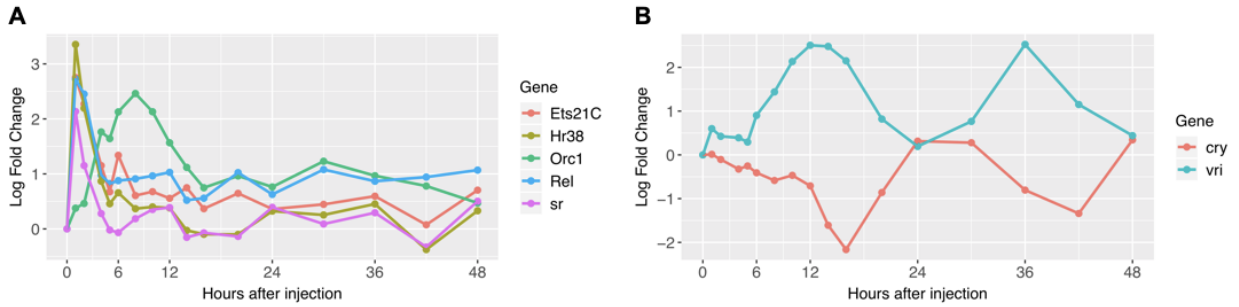


Figure 5. Temporal dynamics of Differentially Expressed Transcription Factors. (**A**) Immediately early (*Ets21C*, *Hr38*, *Rel*, and *sr*) and late (*Orc1*) up-regulation after immune challenge. (**B**) 24 h circadian rhythm patterns (*cry*, *vri*).

Identification of genes with circadian rhythm patterns

The dense sampling scheme of our temporal profiling allowed us to extract well characterized periodic features, such as highly expressed genes (**Figure 3.4D**) and transcription factors (**Figure 3.5B**) with known circadian rhythm functions. As a follow up, we decided to use R package *JTK_Cycle* (HUGHES *et al.* 2010) to identify additional genes with 24 h cycling patterns in our data set. JTK_Cycle is a non-parametric algorithm developed to identify periodic features, while estimating their period length, phase, and amplitude (HUGHES *et al.* 2010). 485 genes were identified to have a 24 h cycle with an adjusted *P*-value < 0.01 (**Table S3.3**). Out of those 485 genes, the top 22 periodic genes were identified using a cutoff of BH *Q*-value < 0.05 and amplitude > 0.5 (**Figure 3.6**). Among them we find 4 well characterized circadian genes: *period* (*per*), *takeout* (*to*), *vrille* (*vri*), and *PAR-domain protein 1* (*Pdp1*), as well as 9 genes which do not have assigned circadian functions but have evidence of cyclic behavior in previous literature (**Table 3.1**), and 8 uncharacterized genes that have not yet been reported to have cyclic expression outside this study (**Table 3.1**).

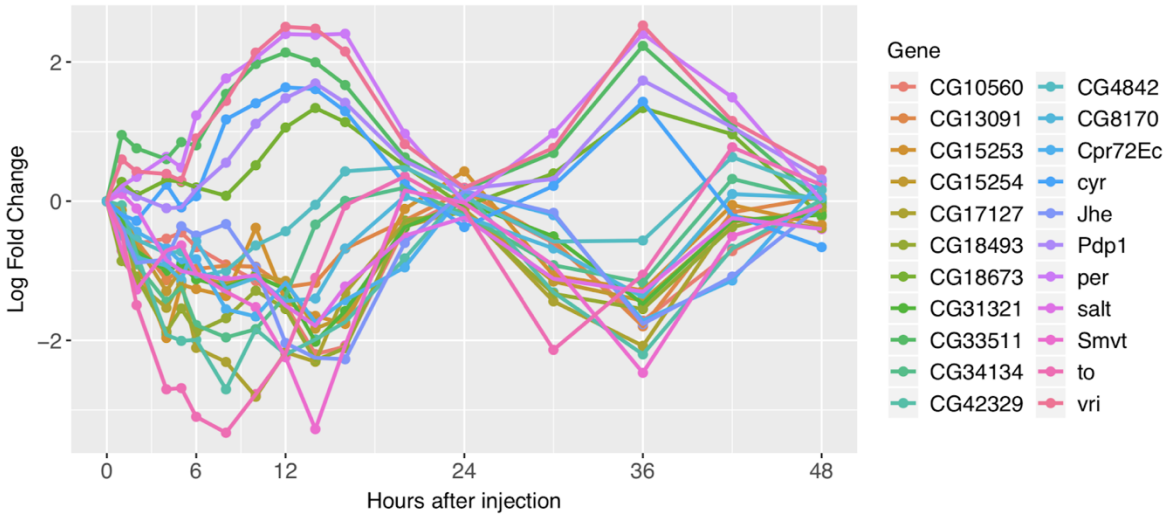


Figure 3.6. Top 22 genes identified by JTK_Cycle show 24 h temporal cycling.

Table 3.1. Evidence of cyclic behavior for top genes identified by JTK_Cycle.

Gene	Evidence of cyclic behavior	Source
CG18673 (*CAH5*, *carbonic anhydrase 5*)	upregulated DE gene in response to light stimulation	Adewoye *et al.* 2015
cypher (*cyr*)	dowregulated DE gene in response to light stimulation	
CG17127	photoperiodic	Pegogaro and Tauber 2018
	rhythmically expressed in constant dark conditions	Ueda *et al.* 2002
Alcohol dehydrogenase *CG4842 and Cuticular protein 72Ec (Cpr72Ec)*	downregulated in the retina after *heme oxygenase* (*ho*) silencing in photoreceptor cells	Damulewicz *et al.* 2019
CG8170	cycling mRNA	Huang *et al.* 2013
Juvenile hormone esterase (*Jhe*)	cyclic hemolymph activity	Zhao and Zera 2004
Sodium-dependent multivitamin transporter (*Smvt*)	modulated by the circadian rhythm in mice	He *et al.* 2016
CG10560, *CG13091* (*Sgroppino*), *CG15253*, *CG15254*, *CG18493*, *CG31321*, *CG33511*, *CG34134*, *CG42329*, and sodium/solute symporter *salty dog* (*salt*)	none reported	

Sources: (UEDA *et al.* 2002; ZHAO AND ZERA 2004; HUANG *et al.* 2013; ADEWOYE *et al.* 2015; HE *et al.* 2016; DAMULEWICZ *et al.* 2018; PEGORARO AND TAUBER 2018)

Temporal clustering identifies distinct patterns of up- and down-regulation of immune processes, and suggests function of uncharacterized genes

Groups of genes that share common functions are often activated and regulated together, which represents a measurable signal in the form of temporal co-

occurrence. We used clustering analysis to discern these temporal patterns of co-expressed genes, which might allow us to characterize their behavior during immune response and identify processes of co-activation and co-regulation. Normalized counts of both replicates were clustered using the autocorrelation-based distance function in R package *TSclust* as explained in **Materials and Methods**. This clustering analysis shows strong temporal patterns that correspond to early and late induction of immune processes, as well as both transient and sustained responses to infection.

Clustering analysis shows temporal patterns of sustained responses to infection. The gene cluster with highest expression after immune response induction includes *AttA*, *AttB*, *AttC*, *DptA*, *DptB*, *Dro*, *edin*, *Mtk*, *PGRP-SB1*, *PGRP-SD*, *CG43236, CG43920*, *CR44404*, and *CR45045*. This cluster is characterized by a strong early induction ~2.5 to 6 log fold change within the first hour, reaching a maximum of 6-8.5 log fold change, and maintaining persistent up-regulation of 2.5 to 5 log fold change throughout 5 days (120 h) (**Figure 3.7**). *AttA*, *AttB*, *AttC*, *DptA*, *DptB*, *Dro*, and *Mtk* are known effector genes of the immune response with antimicrobial peptide (AMP) function (BULET *et al.* 1993; LEVASHINA *et al.* 1995; HEDENGREN *et al.* 2000), *edin* codes for a signaling peptide of the immune response, and *PGRP-SB1* and *PGRP-SD* are peptidoglycan recognition proteins. *CG43236, CG43920*, *CR44404*, and *CR45045* are uncharacterized transcripts that have been shown to be up-regulated after bacterial infections (TROHA *et al.* 2018). This cluster has a pattern of immediate

activation, and most surprisingly genes in this cluster remain strongly up-regulated up to 120 h.

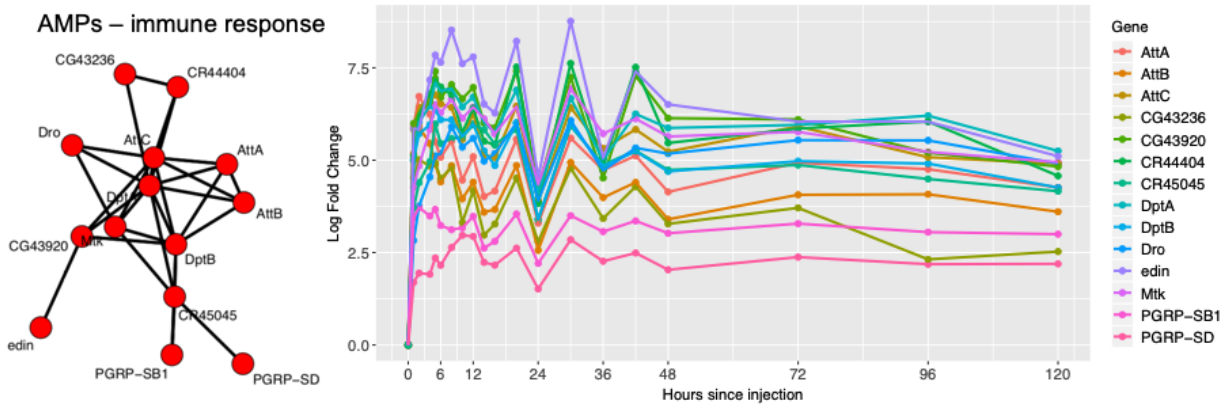


Figure 3.7. Clusters of AMPs show sustained expression after immune inducement throughout 5 days (120 h).

Clustering analysis also identified clusters of genes with transient responses to infection. One of these clusters was composed by a putative class of immune induced peptides: *IM1*, *IM14*, *IM2*, *IM23*, *IM3*, *IM4*, *IMPPP*, and *CG33470*. *CG33470* is an uncharacterized transcript that is located 3.3 kb downstream of *IMPPP* and might belong to the same open reading frame, as both are sometimes referred to as *IM10* (KENMOKU *et al.* 2017), and show nearly identical gene counts in our dataset. This cluster of immune induced molecules is characterized by an early induction (but not as immediate as the AMP cluster) of ~2.5 to 3.5 log fold change within the first two hours, reaching a max of 2.5 to 5 log fold change, and returning to a steady state after 3-5 days (**Figure 3.8A**). This shows that clustering analysis identifies effector immune genes segregating by function: AMPs show an immediate early sustained up-regulation

even after 5 days (**Figure 3.7**), while the IM family has an early up-regulation that eventually returns to steady state levels (**Figure 3.8A**).

Another cluster of transient responses to infection shows later induction and reaches its maximum log fold change after 8-12 h and returns to baseline after 2-3 days (**Figure 3.8B**). Unlike the previous clusters, this cluster is characterized by most genes being first down-regulated immediately after injection for the first 1-2 hours. This cluster was composed of genes from the stress-induced Turandot family (EKENGREN *et al.* 2001) *TotA*, *TotB*, *TotC*, *TotM* and *TotX*, and by *Diedel*, *Grik*, *lectin-24A*, *NimB3*, *CG11459*, *CG16836*, and *CG30287*. *Diedel* is an immunomodulatory cytokine known to down-regulate the imd pathway. *Grik* is a receptor for glutamate, a ubiquitous neurotransmitter that mediates information flow between neurons. *Lectin-24A* is a C-type lectin, which are pattern recognition receptors that mediate pathogen encapsulation by hemocytes (AO *et al.* 2007). *Lectin-24A* has been shown to be down-regulated in the first 2 hours following septic injury and then up-regulated 9 hours after (KEEBAUGH AND SCHLENKE 2012), consistent with the pattern we see in our data. *NimB3* is part of the Nimrod gene family, which are involved in the initial steps of phagocytosis through bacterial binding (ZSÁMBOKI *et al.* 2013). *CG11459* is a predicted cathepsin-like peptidase induced by bacterial infection and injury (KATZENBERGER *et al.* 2016). *CG16836* is located near IM genes *IM1*, *IM2*, *IM3* and *IM23* (expressed in the previous cluster, **Figure 3.8A**), which could explain the similar co-expression patterns. This cluster of genes in the 55C4 region of chromosome 2R

have been recently labeled as "Bomanins" (CLEMMONS *et al.* 2015). *CG30287* is a predicted serine protease, which play many roles in immune response proteolytic cascades (BUCHON *et al.* 2009).

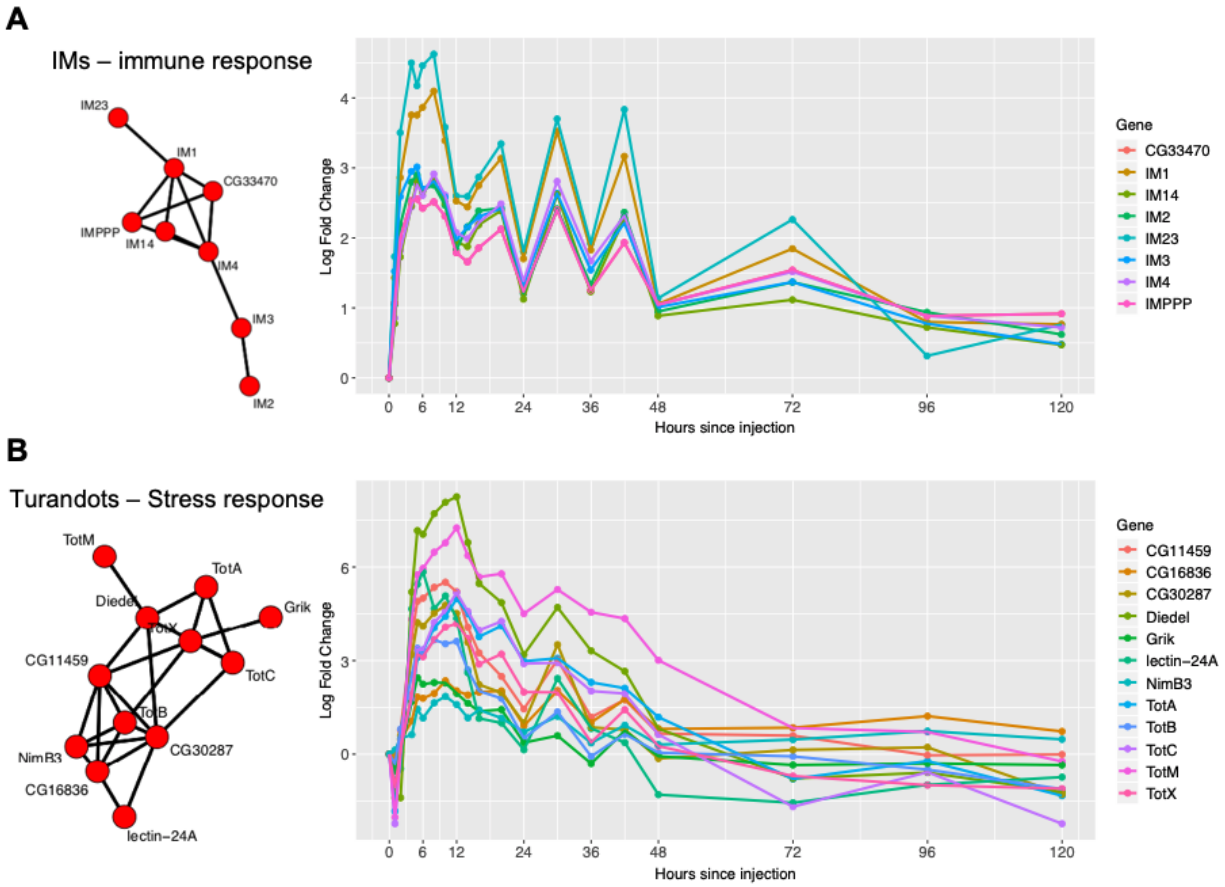


Figure 3.8. Clusters of genes with a transient response, corresponding to (**A**) putative effector immune genes, and (**B**) Turandots (humoral stress response) return to steady state by day 5 (120 h) post immune inducement.

Clustering analysis of this dataset allows us to distinguish between immune response processes with different temporal dynamics (sustained vs. transient, early vs. late induction). Although these clusters show a very robust grouping by function, it is important to note that this grouping was solely driven by temporal co-occurrence in

expression alone, as at no point in the analysis was function or any other gene annotation used as a separating factor. Due to this, clustering analysis can help elucidate the functions of uncharacterized genes.

Temporal gene set pathway analysis shows a divergence in expression between immune and metabolic processes

Functional interpretation of clustering results is limited by those genes that have similar expression patterns across all timepoints. To better elucidate which biological pathways change over time and how, we can integrate prior knowledge of gene affiliation to specific functional categories. This analysis was done using gene set analysis to identify temporal pathway behavior.

Gene set analysis was applied with the R package *GSA*, which uses a Gene Set Analysis algorithm (EFRON AND TIBSHIRANI 2007) that improves the GSEA algorithm (SUBRAMANIAN *et al.* 2005) by allowing testing for associations between gene sets and time-dependent variables (EFRON AND TIBSHIRANI 2007; MULLIGHAN *et al.* 2009). The original GSEA was developed in order to identify relevant pathways and processes being up- or down-regulated in gene expression data. Single-gene methods such as Differential Expression focus only on the top-scoring genes, which can lead to missing biologically significant signals from genes with modest and non-statistically significant expression changes. GSEA is able to detect sets of genes with strong cross-correlation of expression, exposing through their aggregated expression

changes, collections of genes that belong to the same pathway or process (SUBRAMANIAN *et al.* 2005).

Gene set analysis shows that the top up-regulated pathways are all related to immune response, defense response to bacteria, and peptidoglycan functions (**Figure 3.9**). Within these we find pathways related to defense response against both Gram-negative and Gram-positive bacteria. While the commercial LPS used for injections is derived from the outer membrane of Gram-negative bacteria, the injections themselves also result in septic injury, which is known to activate both Gram-positive and Gram-negative immune pathways (Toll and Imd pathways correspondingly) (HOFFMANN AND REICHHART 2002).

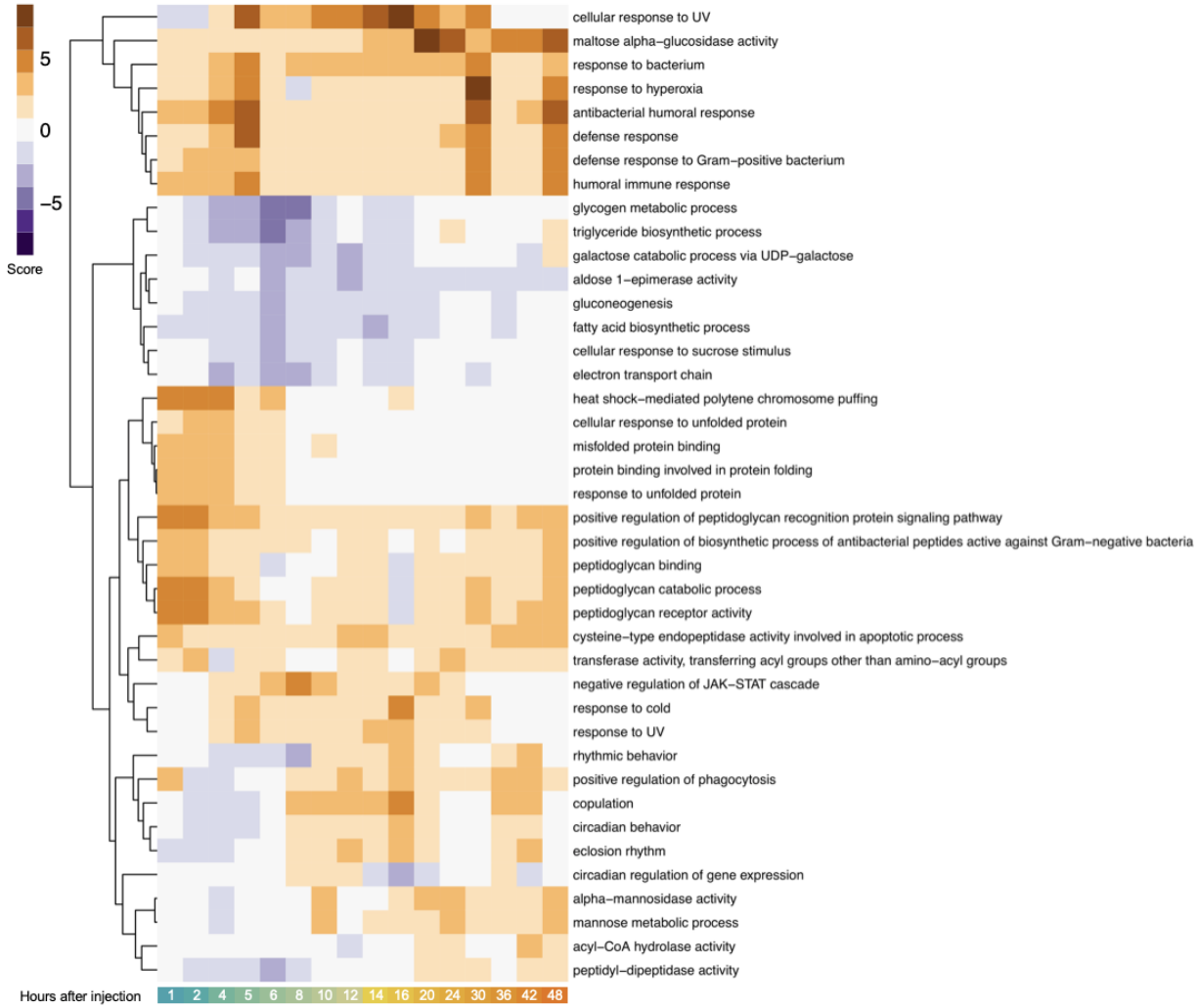


Figure 3.9. Heatmap showing top up- and down-regulated pathways (orange and purple respectively) through the first 48 h post-injections. "Top" pathways had an absolute score > 2.5 and *P*-value < 0.05 in at least one time point, and at least one gene from the pathway was also member of the 551 most predominant time-dependent genes.

Among down-regulated pathways we find many metabolism-related functions, consistent with GO enrichment seen in down-regulated global dynamics Clusters 1 and 4 (**Figure 3.3B**). Three of these pathways (glycogen metabolic process, triglyceride biosynthetic process, and gluconeogenesis) are highlighted in **Figure 3.10**. The glycogen pathway down-regulation pattern seems to be driven by genes *Fatty acid synthase 1* (*FASN1*), and *UGP*, which codes for a UTP--glucose-1-phosphate (**Figure**

3.10A). The triglyceride pathway down-regulation seems to be driven by *FASN1* and *minotaur* (*mino*), a glycerol-3-phosphate 1-O-acyltransferase (**Figure 3.10B**). Finally, the gluconeogenesis pathway down-regulation seems to be driven by *fructose-1,6-bisphosphatase* (*fbp*), a rate limiting enzyme for gluconeogenesis (MIYAMOTO AND AMREIN 2017). These metabolic genes reach their lowest expression within the first 6 hours after injections, and mostly recover by hours 12-24. These metabolic recoveries are much faster than what we observed for transient immune and stress response genes, which take 2-4 days to fully recover (**Figure 3.8**).

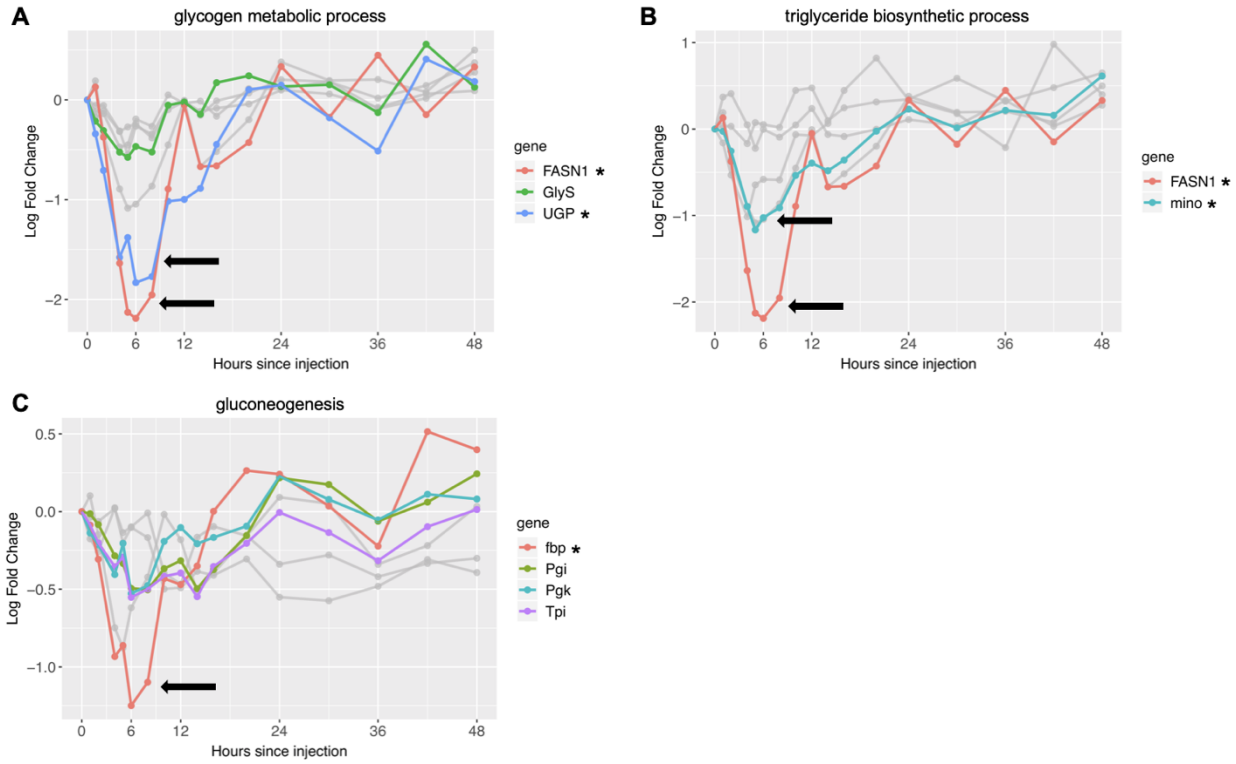


Figure 3.10. Selected significantly down-regulated metabolic pathways with corresponding gene memberships. Genes that have been previously classified as part of the 551 most predominant time-dependent genes are highlighted in color and annotated in the legends, while the rest are in grey. Genes with the strongest expression signals (absolute log Fold Change > 1) are highlighted with arrows and labeled with asterisks (*) in the legends.

Gene interaction modeling of lead-lag patterns using Granger causality

Gene interactions are dynamic, and it follows that temporal gene expression profiles should be able to unveil causal dependencies among genes. If two genes have expression patterns that are correlated but with a certain lag, this lagged correlation, called Granger causality, can help infer a regulatory interaction. This causality concept is based on predictability, if the knowledge of the past of one time series improves the prediction of a second one, the first is said to be Granger causal (GC) for the second. Thus, we constructed directed GC edges and networks of putative interactions among genes (**Figure 3.11**).

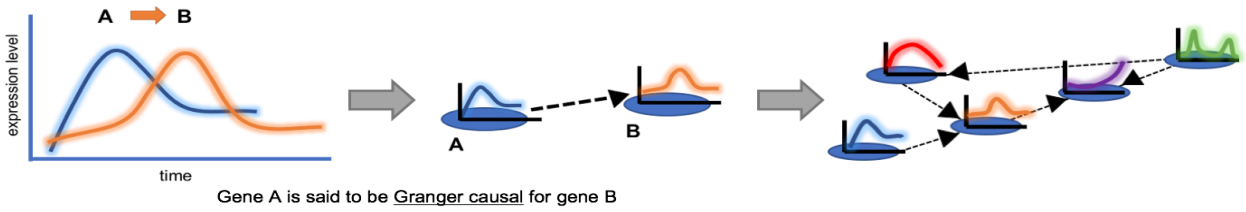


Figure 3.11. Diagram describing the process of constructing directed networks from Granger causality. Lagged correlated expression between two genes (Granger causality) leads to the construction of a directed edge between two genes (nodes), which in turn is used to build directed networks of putative interactions among genes.

First, we compiled a subset of 258 genes by selecting predominant time-dependent genes that had available functional annotation and that had differential expression of at least absolute log fold change of 1. Next, we performed Granger causality analysis on sliding windows of 6 time points on the normalized counts of both replicates (12 data points) using bivariate and multivariate methods as explained

in **Materials and Methods**. Among the GC pairs of genes with the highest number of consecutive windows with significant directed edges we find circadian rhythm genes such as *cryptochrome* and *Smvt* (6 consecutive windows, 6 to 16) (**Figure 3.12A**), *vrille* and *takeout* (4 consecutive windows, 9 to 17) (**Figure 3.12B**), *period* and *takeout* (4 consecutive windows, 9 to 17) (**Figure S3.2**), and *Smvt* and *takeout* (4 consecutive windows, 9 to 17) (**Figure S3.3**). This shows that Granger causality can be used to infer gene dependencies/interactions using global gene expression behavior.

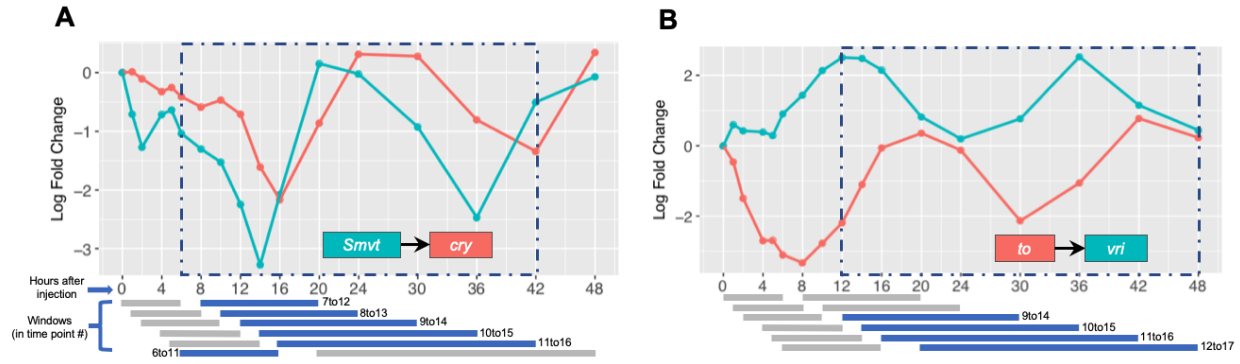


Figure 3.12. GC edges of circadian rhythm genes plotted against time. Significant windows colored in blue, non-significant colored in grey. Resulting overall consecutive windows are labeled in blue dashed rectangles. Individual windows represent 6 consecutive time points, note that time points are not regularly distributed with time, therefore windows have different time ranges, but identical number of samples.

Having found the broadest GC relationships, spanning mostly cyclic genes, we constructed a high-quality set of consistently significant GC edges of divergent expression. To this end we first filtered the subnetwork by (a) removing all nodes corresponding to cyclic genes identified earlier through the JTK_Cycle method, (b) using only pairs of nodes with significant edges (BHFDR < 0.05%) in at least 3

consecutive windows within the first 24 hours of the time course, and (c) trimming the final filtered network by removing all edges with a positive weight, as these edges are more likely to capture spurious causality due to the high correlation between the genes at all timepoints.

Our resulting high-quality GC network contains 51 nodes and 35 edges in 16 connected components (**Figure S3.4**). This network, by design, should include the most interesting examples of divergent expression changes from our full dataset.

The largest connected component in this network (Component #1) is a multifunctional chain of 6 genes annotated in **Figure 3.13A**. GC pairs within this cluster included two metabolic genes, *Sorbitol dehydrogenase 1* (*Sodh-1*) and *UGP*, both negatively directing *Claspin* (both 4 consecutive windows, 2 to 10 and 4 to 12, respectively) (**Figure 3.13C** & **S3.5**). *Claspin* is known to play a role in DNA replication stress (LEE *et al.* 2012). It is known that there is an interplay between host immune systems and replication stress (UBHI AND BROWN 2019). The immune system can detect and respond to replication stress, which is an important feedback loop necessary to remove defective cells (LIU *et al.* 2015). Furthermore, the activation of the immune response generates reactive oxygen species (ROS) and reactive nitrogen species (RNS), and can promote chronic inflammation, all of which can trigger DNA damage (NAKAD AND SCHUMACHER 2016). *UGP* and *fbp* were identified earlier during gene set analysis to drive the down-regulation of metabolic pathways (**Figure 3.10A** and **3.10C**), and in this cluster they are both negatively directed by *LpR2* (3

consecutive windows, 6 to 13) (**Figure 3.13D** and **S3.6**). *LpR2* is a lipophorin receptor, known to regulate the innate immune response by clearing serpin protease complexes from the hemolymph through endocytosis (SOUKUP *et al.* 2009). Lipophorin is a known humoral factor that contributes to clot formation (KARLSSON *et al.* 2004; KRAUTZ *et al.* 2014). Finally, *LpR2* is also shown to negatively direct *juvenile hormone acid methyltransferase* (*jhamt*) (4 consecutive windows, 1 to 9) (**Figure 3.13E**). JHAMT is an enzyme that activates juvenile hormone (JH) precursors at the final step of JH biosynthesis pathway in insects (SHINODA AND ITOYAMA 2003). JH is a known hormonal immunosuppressor in *Drosophila* (ROLFF AND SIVA-JOTHY 2002; FLATT *et al.* 2008; SCHWENKE AND LAZZARO 2017).

Interestingly, *Claspin* was identified to be part of the same pathway as *Orc1* in our previous gene set analysis, showing similar patterns and window of up-regulation (mitotic DNA replication checkpoint pathway, **Figure S3.7**). In our network, *Orc1* is part of an isolated edge with metabolic gene *ABGE* (Component #2, 4 consecutive windows, 4 to 12) (**Figure 3.13A** and **3.13E**).

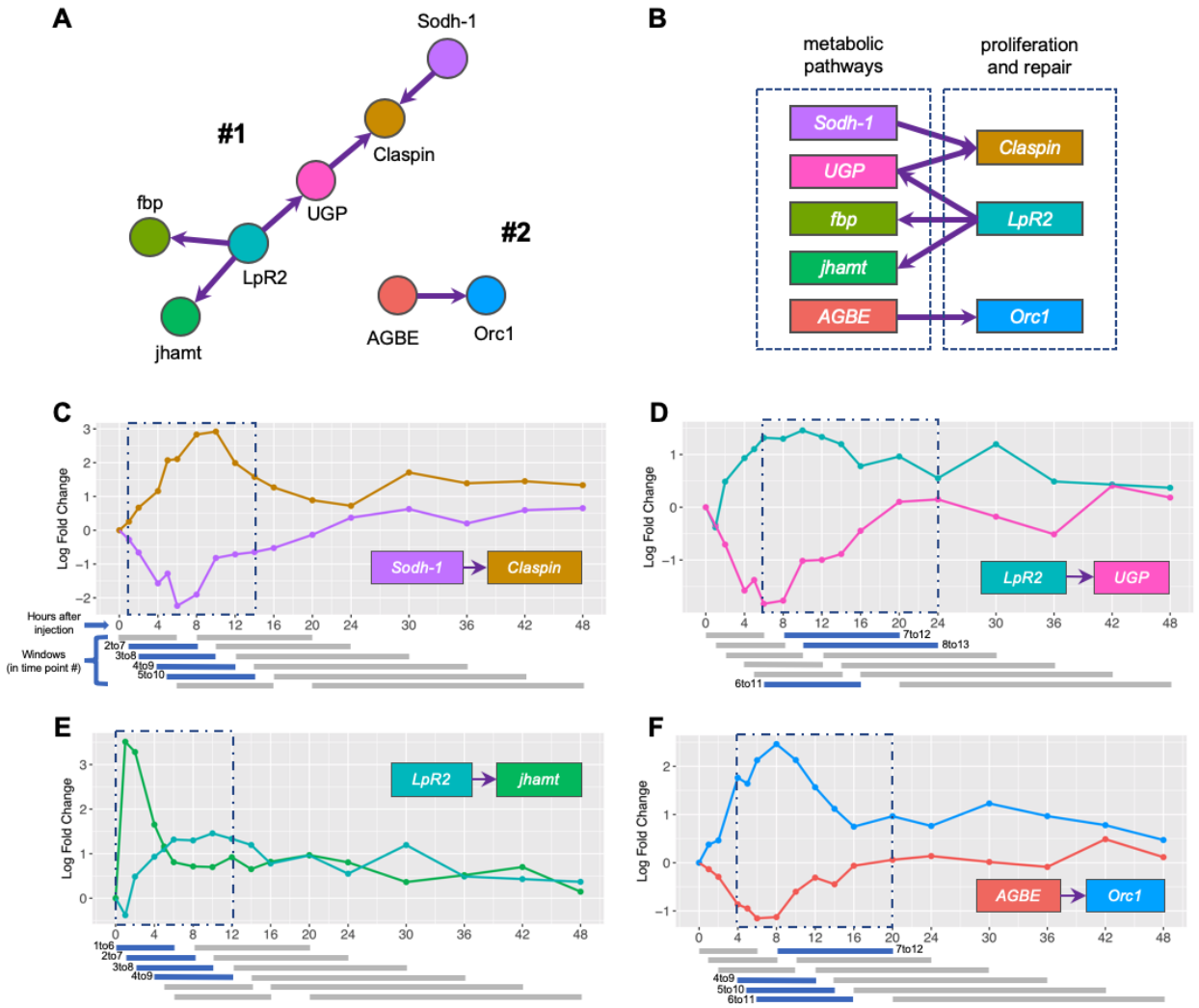


Figure 3.13. High-quality GC network components and their edges. (A) Components #1 and #2 from GC network (Figure S3.4). (B) Diagram summarizes interplay between main represented pathways on the selected components. (C-D) Selected edges from the components plotted against time. Significant windows colored in blue, non-significant colored in grey. Resulting overall consecutive windows are labeled in blue dashed rectangles. Individual windows represent 6 consecutive time points, note that time points are not regularly distributed with time, therefore windows have different time ranges, but identical number of samples.

These prioritized subnetwork components suggest an underlying interplay between metabolic pathways and other pathways such as proliferation and repair (**Figure 3.13B**), motivating follow-up studies to determine which pathways might be regulating and trading off with each other in the hours following an immune

challenge. The overall unfiltered GC network has a multitude of relationships worth exploring further, but limitations in the ability to discern between different types of causality make widespread conclusions from the network challenging.

DISCUSSION

We have produced a dense and high-quality time course profiling of the *Drosophila* transcriptome response to commercial LPS immune challenge and injection using RNA-seq sampling over 20 time points in 5 five days. This profiling provides a high-quality high-dimensional dataset, which we analyzed using a broad range of statistical methods.

First, we identified a subset of most significant time-dependent genes using spline fitting and pairwise differential expression analysis. We grouped and classified these time-dependent genes based on their temporal expression profiles to identify patterns of cyclic behavior, as well as distinct responses to immune challenge with divergent initiation and resolution dynamics, all discussed in detail below. Recurring patterns of expression corresponded to distinct functional categories, allowing us to infer the functional class of previously uncharacterized genes.

Next, we expanded our profiling of the immune response by directly querying the functional pathways of the genes in our dataset. We identified differentially expressed biological pathways using pairwise gene set analysis, which allowed us to

characterize functional pathway transcription dynamics. The resulting global behaviors of immune response and metabolic pathway expression point to potential resource tradeoffs, precipitated by the immune challenge.

Finally, we constructed directed Granger-causal networks of putative interactions to detect the regulation of one gene by another. We filtered this causal network to generate a high-quality set of divergent network connected components in which the interplay between metabolic pathways and separate pathways such as proliferation and repair becomes clear.

Below, we describe and discuss in detail the main insights from these analyses, as well as limitations and future steps.

Cyclic patterns of expression

The dense sampling nature of this time course allows us to discern the clear cycling patterns of differentially expressed genes and transcription factors such as *period*, *timeless*, *takeout*, *vrille*, and *cryptochrome*, all of which have well-characterized circadian rhythm functions (KONOPKA AND BENZER 1971; MYERS *et al.* 1996; SO *et al.* 2000; CYRAN *et al.* 2003; COLLINS *et al.* 2006) (**Figure 3.4D**, **3.5B**). We systematically collected all genes fitting this 24-hour expression pattern using the JTK_Cycle algorithm. This method detected multiple well-known circadian rhythm genes, as well as genes that have been shown to be or indicated as cyclic in previous literature (**Figure 3.6**). More interestingly, this method also detected multiple uncharacterized

genes that - to our knowledge - have not yet been reported to have cyclic expression, such as *CG10560*, *CG13091* (*Sgroppino*), *CG15253*, *CG15254*, *CG18493*, *CG31321*, *CG33511*, *CG34134*, *CG42329,* and *Sodium-dependent multivitamin transporter* (*salt*) (**Table 3.1**).

The identification of canonical circadian rhythm patterns both validates our methods of data normalization and differential expression analysis, and increases the certainty that we are accurately profiling novel temporal dynamics. It is important to note, however, that proper validation of the cycling behavior of our novel cyclic genes should be performed under normal *Drosophila* conditions, as we do not know whether immune challenge affected their expression.

Sustained versus transient responses to immune challenge

We characterized two temporal co-expression clusters presenting a transient response to the immune challenge. This response follows a pattern of up-regulated expression after injection with a return to the pre-challenge state within 2-5 days (**Figure 3.8**). These two clusters were mainly composed of known immune-induced molecules (IMs, or Bomanins (CLEMMONS *et al.* 2015)) and stress-induced Turandot genes. We additionally identified a separate set of metabolic genes (e.g. *FASN1*, *UGP*, *fbp*, and *mino*) present in pathways with the opposite transient response to the immune challenge. This response follows a pattern of down-regulated expression after injection followed by recovery to the pre-challenge state (**Figure 3.10**). Unlike the

immune/stress recovery period (2-4 days), the recovery of these genes was markedly faster (12-24 hours). Finally, we describe a temporal co-expression cluster of a sustained response to immune challenge. This response follows a pattern of up-regulated expression within the first hour post-injection and remains up-regulated during the entire 5-day time course (**Figure 3.7**). This cluster was mainly composed of antimicrobial peptides (AMPs).

Our time series allows us to characterize both transient and sustained responses to immune challenge. Variations in this response include a marked difference in recovery time and, in some cases, a lack of recovery even after 5 days. We observed that metabolic genes recovered faster than immune genes, suggesting that the early stages of infection likely involve the greatest tradeoffs. The striking difference in expression level recovery is highlighted by the sustained expression in AMPs throughout our entire time course. Further work is needed to better define what it means to return to normality, if normality is achieved at all, and how that might be contributing to long term effects either detrimental or beneficial.

Different stages of transcriptional change

Gene sets belonging to different functions show an up- or down-regulation immediately after immune challenge, reaching their highest or lowest point of expression within 4 different time frames. We first observed these phenomena when clustering time-dependent genes based on their global expression profiles (**Figure**

3.3B), and later characterized individual functional categories during temporal clustering analysis. AMPs showed the fastest up-regulation within the first 1-2 hours after immune challenge (**Figure 3.7**), a pattern we also observed in many transcription factors (**Figure 3.5A**). Metabolic genes and IMs reach their lowest and highest point of expression respectively at 5-8 hours after immune challenge (**Figure 3.10** and **3.8A**). Proliferation and repair genes reach their highest point of expression at 8-10 hours (**Figure 3.13**). Finally, stress-related Turandot genes reach their highest point of expression at 10-12 hours (**Figure 3.8B**).

Our analysis shows that the immune response may be orchestrated by the changing expression in multiple functional groups with different initiation and resolution dynamics. Together, the above analyses paint a picture of an immune response that involves both transient and sustained changes in expression that occur over different timescales. This time course allows us to see layers of expression dynamics at a resolution that is unprecedented, allowing us to generate new hypotheses about the interplay and potential tradeoffs between functional pathways.

Predicting function by association

Our analysis identified genes that showcase cyclic behavior but are not canonically circadian-associated genes (**Figure 3.6**). This includes 9 genes which do not have assigned circadian functions but do have some evidence of cyclic behavior in

previous literature. It also includes 8 genes that had not been reported to exhibit any cyclic expression before this study, described in detail in **Table 3.1**.

Temporal clustering analysis identified uncharacterized genes (*CG43236, CG43920*, *CR44404*, and *CR45045*) that share similar expression dynamics with AMP pathway associated genes (**Figure 3.7**). *CR44404* and *CR45045* are uncharacterized long noncoding RNA (lncRNAs) and, while lncRNAs do not encode proteins, they have been associated with transcriptional repression and activation (LONG *et al.* 2017). In mammals, lncRNAs have been implicated in regulating immune gene expression (CARPENTER AND FITZGERALD 2015; MUMTAZ *et al.* 2017), but it is not yet known if they play a role in *Drosophila* immunity. *CG43236* has been shown to be up-regulated after injury, among other innate immune genes (KATZENBERGER *et al.* 2016). *CR44404, CR45045, CG43236,* and *CG43920* were shown to be up-regulated in a *D. melanogaster* transcriptome profiling after immune challenge with 10 bacterial infections (TROHA *et al.* 2018). *CG43236* and *CG43920* have also been shown to encode small proteins predicted to be cationic (IM 2018), properties shared by known AMPs (LEMAITRE AND HOFFMANN 2007); while lncRNAs *CR44404* and *CR45045* were predicted to physically interact with antimicrobial peptide transcripts (IM 2018). These predictions in the literature are consistent with our dynamics-based implication of these uncharacterized genes as AMPs.

We are able to implicate these genes as potential members of these functional pathways due to their strong expression-dynamic similarity. This is impactful both in

the novel functional implication of previously uncharacterized genes, but also in demonstrating the potential this method of function-by-association has to assign function to other uncharacterized genes through RNA expression time course experiments.

Functional interplay and potential trade-offs

Our dataset shows distinct global dynamics pointing to a divergence in the functional responses to immune challenge. We first observed divergence in expression patterns when clustering time-dependent genes based on their global expression profiles, with clusters enriched for immune and stress response functions being up-regulated and clusters enriched for metabolic processes being down-regulated (**Figure 3.3**). The divergence in expression between immune and metabolic processes is strikingly confirmed in gene set analysis, as the most up-regulated pathways related to immune response functions, while the most down-regulated pathways related to metabolic functions (**Figure 3.9**). *FASN*, which shows the strongest down-regulation in both glycogen metabolic process and triglyceride biosynthetic process (**Figure 3.10A-B**), is a lipogenic gene whose down-regulation might indicate a need to have easily accessible nutrients instead of storing them. Indeed, infections in mammals are known to induce adipose tissue lipolysis (WOLOWCZUK *et al.* 2008) and bacterial peptidoglycan is a ligand that stimulates lipolysis as well (CHI *et al.* 2014). The gene with the strongest down-regulation in the gluconeogenesis pathway was *fbp* (**Figure**

3.10C), which codes for fructose-1,6-bisphosphatase, the rate limiting enzyme for gluconeogenesis. This gene was significantly down-regulated in a study showing that *Listeria monocytogenes* infection in *Drosophila* causes a decrease in energy stores, with reduced levels of triglycerides and glycogen (CHAMBERS *et al.* 2012). The divergent dynamics detected in our dataset are thus in agreement with known individual mechanisms characterized in the immune response.

We further see implications of functional interplays with our Granger Causal (GC) network analysis. Main subnetwork components showed significant GC directional edges between down-regulated metabolic genes (such as *Sodh-1*, *UGP*, *fbp*, and *AGBE*) and up-regulated genes with cell proliferation and repair functions (*Claspin*, *LpR2*, and *Orc1*) (**Figure 3.13**). These results further suggest an underlying interplay between metabolic pathways and proliferation and repair mechanisms such as regulation of DNA replication stress, endocytosis, and clot formation.

Overall, the analysis of transcriptional patterns of the *Drosophila* genes in our experiment points to a global tradeoff between the immune response and metabolic processes. GC networks of putative gene interactions further suggest an interplay between metabolic and repair functions. The clear divergent functional responses to immune challenge, along with their distinct initiation and resolution expression dynamics, help us characterize and further understand the orchestration of the immune response.

Limitations and future steps

One of the main limitations of this time-course experimental design is the lack of time-matched controls to account for expression changes associated with phenomena outside the immune challenge, such as aging. It is important to note, however, that it is still highly valuable to develop and improve methods for analyzing time-course transcriptional data lacking time-matched control samples since they are needed to analyze processes such as development, where such controls are not possible. Experimental data and theoretical analysis has shown that under reasonable assumptions, sampling time points at higher resolution is preferred over having more replicates (SEFER *et al.* 2016), an important strategy to consider when having a limited experimental budget. However, even under this consideration our time-course experimental design would have benefited from more replicates, as two replicates per time point led to exclusion of one time point (3 h) when one of the replicates did not pass quality filters.

In our dataset, Granger Causality analysis excelled at showcasing the relationships between divergent gene pairs, but was overly sensitive to the extreme temporal correlation between large groups of genes. To avoid a prohibitively dense network for analysis, we relied on heuristic network trimming criteria, which was effective, but is likely not generalizable to other similar experiments. Developing co-integration methods that take into account the specific bias found in high-dimensional RNA-seq datasets would provide a more robust statistical analysis of the causal

relationships observed in this type of data. Alternatively, a more aggressive, systematic filtering of genes based on prior biological knowledge could help alleviate this sensitivity to correlation, while also decreasing the computational power required to run this analysis. Potential filtering schemes could include binning genes by their common biological pathways, or generating *a priori* assumptions on the Granger causality of any two genes given that they are protein-protein interactors, their chromosomal location, or their transcription factor promoter/target relationships. Finally, it is indispensable that these statistical causal relationships be confirmed with direct experimental disruptions of a system, as only carefully controlled intervention can truly demonstrate biological causality.

Overall, this analysis motivates innovation in computational methods for longitudinal omics data, both to account for their inherent high-dimensionality and the complex underlying architecture that contains both causal and spurious coordination. Further, this should serve as a proof of concept for the future of high-density time-course RNA-seq in other model organisms.

AUTHOR CONTRIBUTIONS

F.S., A. E., and A.G.C. conceived the study. F.S. collected the samples and generated the time course data. F.S., S.B., M.T.W., and A.G.C. conceived the computational and statistical analyses. F.S. and S.B. performed the computational and statistical analyses. F.S., S.B., and A.G.C. wrote the manuscript.

SUPPORTING INFORMATION

Figure S3.1. Heatmap of 214 genes

Figure S3.2. GC edge between *period* and *takeout*

Figure S3.3. GC edge between *Smvt* and *takeout*

Figure S3.4. GC filtered network

Figure S3.5. GC edge between *Claspin* and *UGP*

Figure S3.6. GC edge between *LpR2* and *fbp*

Figure S3.7. Down-regulated pathway corresponding to 'mitotic DNA replication checkpoint'

Table S3.1. Table with all 214 DE genes

Table S3.2. 551 most predominant time dependent genes

Table S3.3. All 485 JTK genes adj pval < 0.01

CHAPTER 4

HIGH-RESOLUTION QTL MAPPING WITH DIVERSITY OUTBRED MOUSE STRAINS IDENTIFIES GENETIC VARIATION THAT IMPACT GUT MICROBIOME COMPOSITION[3]

INTRODUCTION

The gastrointestinal tract of all vertebrates, including humans, harbors a complex ecological community of highly diverse microbes referred to as the gut microbiota. The microbiota colonizes the gut for the first time during the birth of the host and its composition is influenced by many factors during the host's life such as disease, diet, and antibiotics (FRANCINO 2016; BATTAGLIOLI AND KASHYAP 2018; DUDEK-WICHER *et al.* 2018; DASH *et al.* 2019). Variation in the human gut microbiome composition has also already been associated with host immune responses (ROUND AND MAZMANIAN 2009; GARRETT *et al.* 2010; VEIGA *et al.* 2010), metabolic phenotypes (TURNBAUGH *et al.* 2009; RIDAURA *et al.* 2013), and diseases such as obesity (LEY *et al.* 2005), heart disease (FAVA *et al.* 2006), and diabetes (WEN *et*

[3] Manuscript in preparation: Schlamp F, Zhang DY, Cosgrove E, Edwards M, Simecek P, Pack A, Goodrich JK, Ley R, Churchill GA, and Clark AG. High-resolution QTL mapping identifies genetic variation associated with gut microbiome composition in Diversity Outbred mice.

al. 2008). Given the roles of the gut microbiome in complex human diseases, it is important to characterize the factors that impact microbiome composition.

While it is clear that the gut microbiome composition is strongly impacted by environmental exposures (ROTHSCHILD *et al.* 2018), the role of host genetics has only recently been implicated (GOODRICH *et al.* 2014b; BLEKHMAN *et al.* 2015; GOODRICH *et al.* 2016). Studies have identified multiple genetic variants significantly associated with specific bacterial taxa abundances (DAVENPORT *et al.* 2015; BONDER *et al.* 2016; TURPIN *et al.* 2016; WANG *et al.* 2016; GOODRICH *et al.* 2017; IGARTUA *et al.* 2017; ROTHSCHILD *et al.* 2018), despite the observation that generally the primary determinants of microbiome composition are non-genetic (ROTHSCHILD *et al.* 2018). Human genetic studies have significant limitations for accurate assessment of genetic effects on the microbiome, including accessibility to large and diverse sample populations as well as a general lack of control over confounding variables. One major limitation is that there is minimal control of diet and other environmental factors, and so only the strongest genetic effects can be detected.

The mouse model, with the ability to control diet, along with well-defined genetic differences among inbred lines, provides a better opportunity to dissect genetic and environmental factors impacting microbiome composition. Quantitative trait locus (QTL) mapping efforts show that gut microbiota composition is a polygenic trait, with clearly mappable genetic factors influencing the gut microbiome

composition (BENSON *et al.* 2010; MCKNITE *et al.* 2012; SNIJDERS *et al.* 2016). Standard QTL mapping approaches have low mapping resolution, however, and advanced intercross lines provide one excellent means of improving mapping resolution. Belheouane *et al.* (BELHEOUANE *et al.* 2017) performed genetic and 16S rRNA gene analysis of skin microbiomes of a collection of 15-generation advanced intercross lines, and demonstrated that the improved mapping resolution also improved the specificity and significance of genetic associations. It is clear that the mouse model will provide further opportunities to dissect the means by which the host genome can modulate microbiome composition. A logical next step is a mapping experiment to identify portions of the genome that influence functional pathways that modulate the microbiome.

Here we extend the analysis of the link between the host genome and microbiome using the Diversity Outbred mouse model. The Diversity Outbred (DO) population is a heterogeneous mouse stock derived from the same eight progenitor lines (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ) used to establish the Collaborative Cross (CC) (COLLABORATIVE CROSS CONSORTIUM 2012). Mice from the CC lines at early stages of inbreeding were used to establish the DO population, which is maintained by randomized outbreeding among 175 mating pairs. The result is each individual DO mouse represents a unique combination of segregating alleles, whose genome is a

unique mosaic of the original eight progenitor lines. The advantages of this outbreeding include normal levels of heterozygosity — similar to the human genetic condition — and substantially increased genetic mapping resolution (CHURCHILL *et al.* 2012). The CC/DO mice founder progenitor lines have already proven to be successful in identifying genetic associations with intestinal microbiome composition (O'CONNOR *et al.* 2014).

In this study, motivated by the high level of environmental control of the laboratory mouse and the improved mapping resolution of the Diversity Outbred mouse system, we identified genetic underpinnings of the gut microbiota of 247 Diversity Outbred mice. We uncover strong evidence of host genetic factors influencing the composition of many specific attributes of the gut microbiome. These included not only associations between specific host genetic variants and abundances of particular bacterial taxa, but also associations with functional molecular pathways.

MATERIALS AND METHODS

Animal population and sample collection

Male mice from the Diversity Outbred Mouse Panel were obtained from The Jackson Laboratory (Bar Harbor, ME, USA) at 6 weeks of age. Mice were group housed (5 animals per cage) for 2 weeks of post-travel acclimation, and then single housed at identical conditions. All mice were reared on chow diet. Fecal pellets from

249 mice were collected at 3 months old (two samples were later discarded, leaving a final analyzed dataset of 247 mice). Pellets were stored in Eppendorf tubes placed on dry ice and moved to a -80°C freezer until processing.

Microbial DNA extraction, 16S rRNA gene PCR, and sequencing

Microbial community DNA was extracted from one single frozen pellet per sample using the MO BIO PowerSoil-htp DNA Isolation Kit (MO BIO Laboratories, Inc., cat # 12955-4), but instead of vortexing, samples were placed in a BioSpec 1001 Mini-Beadbeater-96 for 2 minutes. We used 10-50 ng of sample DNA in duplicate 50 µl PCR reactions with 5 PRIME HotMasterMix and 0.1 µM forward and reverse primers. We amplified the V4 region of 16S using the universal primers 515F and barcoded 806R and the PCR program previously described (CAPORASO *et al.* 2011), but with 25 cycles. We purified amplicons using the Mag-Bind® E-Z Pure Kit (Omega Bio-tek, cat # M1380) and quantified with Invitrogen Quant-iT™ PicoGreen® dsDNA Reagent, and 100 ng of amplicons from each sample were pooled and paired end sequenced (2x250bp) on an Illumina MiSeq instrument at Cornell Biotechnology Resource Center Genomics Facility.

16S data processing

We performed demultiplexing of the 16S rRNA gene sequences and OTU picking using open source software package Quantitative Insights Into Microbial

Ecology (QIIME) version 1.9.0 with default methods (CAPORASO *et al.* 2010). The total number of sequencing reads was 15,149,384, with an average of 61,334 sequences per sample and ranging from 17,658 to 135,803. Open-reference OTU picking at 97% identity was performed against the Greengenes 8_13 database. 12% of sequences failed to map in the first step of closed-reference OTU picking. The taxonomic assignment of the reference sequence was used as the taxonomy for each OTU. 'NR' within taxa names represents New Reference OTUs defined as those with sequences that failed to match the reference and are clustered *de novo*. Random subsamples were used to create a new reference OTU collection and 'NCR' represents New Clean-up Reference OTUs that failed to match the new reference OTU collection (RIDEOUT *et al.* 2014).

For the non-rarefied data, read count was used as an additional covariate during QTL mapping to reduce the effect of sequencing depth. A rarefied dataset was also used for heritability estimates and QTL mapping, as explained in **Supplemental Material**. Two extreme outliers were omitted from further analysis, yielding a total of 247 samples. To differentiate the non-rarefied taxa from the rarefied taxa, we use 'NonR' to represent the non-rarefied dataset and 'R' to represent the rarefied dataset.

For heritability estimates and QTL mapping, a filter was applied across all 247 samples that removed any taxon that was not present in more than 50% of the

samples. Relative abundance of reads (number of reads clustered to each taxa divided by the total number of reads in a given sample) was used as the tested phenotype.

Stacked bar plots of the most abundant taxa within each taxonomic level were plotted with R-package *ggplot2*. A box-plot was first generated for each taxonomic level depicting the abundances of the taxa within that taxonomic level across the 247 samples (**Figure S4.1**). The top ten taxa with the highest average abundances are selected to be plotted in the stacked bar plot, ordered by the most abundant taxon. A heatmap that correlates similarities between taxa from the non-rarefied and rarefied datasets based on the Pearson correlation coefficient was plotted using the R-package *corrplot* (**Figure S4.2**).

SNP genotyping

SNP genotyping was done at the Jackson Laboratories on each of the 247 mice using The Mega Mouse Universal Genotyping Array (MegaMUGA). A total of 57,973 SNPs passed all the QC metrics and were used in the heritability and mapping analysis reported here.

Heritability calculations

Heritabilities of the various bacterial taxa were quantified and calculated on autosomes using a linear mixed model as implemented in R-package *lme4qtl* via the

relmatLmer() function (ZIYATDINOV *et al.* 2018) (https://github.com/variani/lme4qtl). This linear mixed model enables us to decompose variability into genetic and environmental components. The variance of the genetic component is expected to be $\sigma_g^2 K$, where K is a kinship matrix normalized as proposed in (KANG *et al.* 2010). The kinship matrix in specified via the "relmat" argument in relmatLmer(). To account for the potentially confounding effects of shared cages during acclimation (as noted above in Section 5.1), we also included cage as a random effect in our model. Thus, the model included estimates of variance of the genetic component (σ_g^2) and the cage component (σ_{cage}^2), and the residual variance due to unspecified environmental factors (σ_{rs}^2).

The narrow sense heritability was then estimated as:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{cage}^2 + \sigma_{rs}^2}$$

Sequencing lane was included as a covariate in both non-rarefied and rarefied datasets. For our non-rarefied dataset, narrow sense heritabilities were calculated using the number of read counts as an additional covariate. Significance of heritability estimates was assessed by conducting a restricted likelihood ratio test using the exactRLRT() function in the R-package *RLRsim* (SCHEIPL *et al.* 2008), as applied in Supplementary Note 3 in (ZIYATDINOV *et al.* 2018).

QTL Mapping

For QTL mapping, the relative abundances were rank Z-transformed using R-package *DOQTL* (GATTI *et al.* 2014) and then mapped using a linear mixed model in R-package *lme4qtl::relmatLmer()* (ZIYATDINOV *et al.* 2018) on autosomes with kinship included as random effect to account for genetic relatedness among animals. For the bacterial taxa from the five taxonomic levels, we generated QTL mappings for all taxa that passed the 50% zero cut-off (i.e. those present in at least 50% of the mice), with the taxa designated as the phenotype. Sequencing lane (fixed effect) and cage (random effect) were included in both non-rarefied and rarefied datasets. We included read count as an additional covariate (fixed effect) for our non-rarefied dataset. Significant and suggestive associations were identified in a two-step procedure. First, we applied likelihood ratio tests comparing models with and without genotype. *P*-values derived from these tests were adjusted for multiple testing across SNPs (within a given taxon) using R function p.adjust() with method "BH" (BENJAMINI AND HOCHBERG 1995). In the second step, we conducted permutation tests (1000 permutations) for taxa that had associations with adjusted p-value < 0.1 in the first step.

For every bacterial taxon from the five taxonomic levels with a statistically significant QTL association, we mapped the OTUs belonging to that taxon. We applied a 50% zero cut-off filter to only retain common OTUs. With the OTUs

obtained, we generated QTL mappings and assessed significance just as we had done for the five taxonomic levels.

Gene Set Pathway Analysis

We used the open-source online DAVID annotation tool (HUANG *et al.* 2008) and the Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, CA) software to conduct gene set pathway analysis. We used DAVID v6.8 and their functional annotation tool to reduce large gene sets into smaller groups of functionally related genes. A list of gene names was uploaded onto the website with the identifier parameter set to 'official_gene_symbol' and the species *Mus musculus* selected. DAVID then outputs a list of categories, such as functional, gene ontology, tissue expression, and others, which contained subsets of the inputted gene set. Within each category, DAVID also lists more specific categories and by displaying the genes for each sub-category, we were able to view which of the genes from our gene list were found to be associated with various different classifications. From the association results, a *p*-value filter allowed us to view only the results above a certain EASE *p*-value threshold, a modified Fisher-Exact *p*-value score. We chose the groupings with shown higher significance and reinforced the results outputted by DAVID with KEGG pathway database (KANEHISA *et al.* 2017) by simply confirming the presence of each gene in their organized category, as by DAVID, in KEGG's online database.

Using IPA, a new "core analysis project" was created and then our list of genes was uploaded as a dataset with parameters chosen to fit the formatting of our gene list. Before running the analysis, we set the reference set to be Ingenuity® Knowledge Base and then ran our analysis. IPA uses multiple categories to classify the inputted gene set and we focused on their disease and functions category. Others include expression, regulatory effects, and other canonical pathways. Additionally, IPA generates networks of genes proven to be either directly or indirectly related to each other. We chose the most significant network outputted and identified the intersection of that network with the network relating the genes in our QTL hits with the respective disease.

RESULTS

Variation of gut microbiota

High-throughput sequencing of fecal samples from 247 three month old male mice from the Diversity Outbred Mouse Panel generated 15,149,384 16S rRNA gene sequences that passed the quality filtering criteria after demultiplexing (see **Materials and Methods**). On average, 61,334 sequences were obtained per sample (ranging from 17,658 to 135,803 sequences). Sequences were sorted into 57,014 operational taxonomic units (OTUs) at 97% identity against the Greengenes 8_13 database using open-reference OTU picking. Next, OTUs were summarized at five levels of

taxonomy (phylum, class, order, family, genus). In order to focus on the most abundant microbes, only the taxa present in at least 50% of samples (i.e. present in 124 samples or more) were used for all following analysis, leaving a total of 80 taxa to test at the five levels of taxonomy (7 phyla, 9 classes, 12 orders, 21 families, and 31 genera). The most predominant taxa at the phylum level were Firmicutes (average relative abundance = 48.64%) and Bacteroidetes (46.41%), which is consistent with previous findings (BENSON *et al.* 2010; MCKNITE *et al.* 2012; ORG *et al.* 2015). The relative abundances of these taxa were highly variable, with Firmicutes ranging from 11% to 94%, and Bacteroidetes ranging from 1% to 88% (**Figure 4.1**).

Figure 4.1. Relative abundances of top ten most abundant phyla across the 247 mouse strains. Relative abundances shown, mouse strains sorted by phylum Firmicutes, the most abundant phylum.

The top 8 most abundant genera were present in at least 99% of the samples. The two most abundant genera were an unidentified genus within Bacteroidales family S24-7 (average relative abundance = 43.89%, ranging from 1% to 88%) and another unidentified genus within Clostridiales (32.35%, ranging from 4% to 78%), consistent with previous findings (SHIN *et al.* 2016). Stacked bar plots and box plots depicting relative abundance frequencies for all five taxonomic levels are available in **Figure S4.1**.

When dealing with uneven sequence counts across samples, microbiome studies commonly use rarefaction as a data normalization approach, consisting of randomly selecting from each sample an equal number of sequences (GOODRICH *et al.* 2014a). It has been argued, however, that rarefaction is not an ideal approach due to valuable data being discarded (MCMURDIE AND HOLMES 2014). Therefore, we decided to present our analysis of the non-rarefied data using sequence counts per sample as a covariate, noting also that the rarefied data provided highly concordant results (see **Supplemental Material**).

Heritability estimation

Each of the 247 individual mice used in this study was genetically unique. The unit of inference for phenotypes was the relative abundance of each taxon in each individual, while the units of genetic inference were the SNP genotypes at each of

57,973 sites for each mouse using the mouse array. We estimated narrow-sense "SNP" heritability (h^2) using a linear mixed model in R-package *lme4qtl* (ZIYATDINOV *et al.* 2018). A linear mixed model was used to predict whether the effects of the autosomal genotype on the phenotype is proportional to the genetic similarity between the mice, after adjustment for known factors. Thus, calculations were based on the kinship matrix (genetic similarity), expression of a phenotype (taxon abundance) across all samples, and additional covariates (such as sequencing lane, read counts, and cage effect). Significance was assessed by an exact (restricted) likelihood ratio test using R-package *RLRsim* (SCHEIPL *et al.* 2008). More details can be found in **Materials and Methods**. In total, 27 of the 80 tested taxa were heritable (nominal p-value < 0.05), with 3 additional taxa having statistically suggestive heritabilities of 20% or more (nominal p-value < 0.1) (**Table S4.1A**). Proportion variance estimates for kinship and cage for all taxa are presented in **Figure 4.2.**

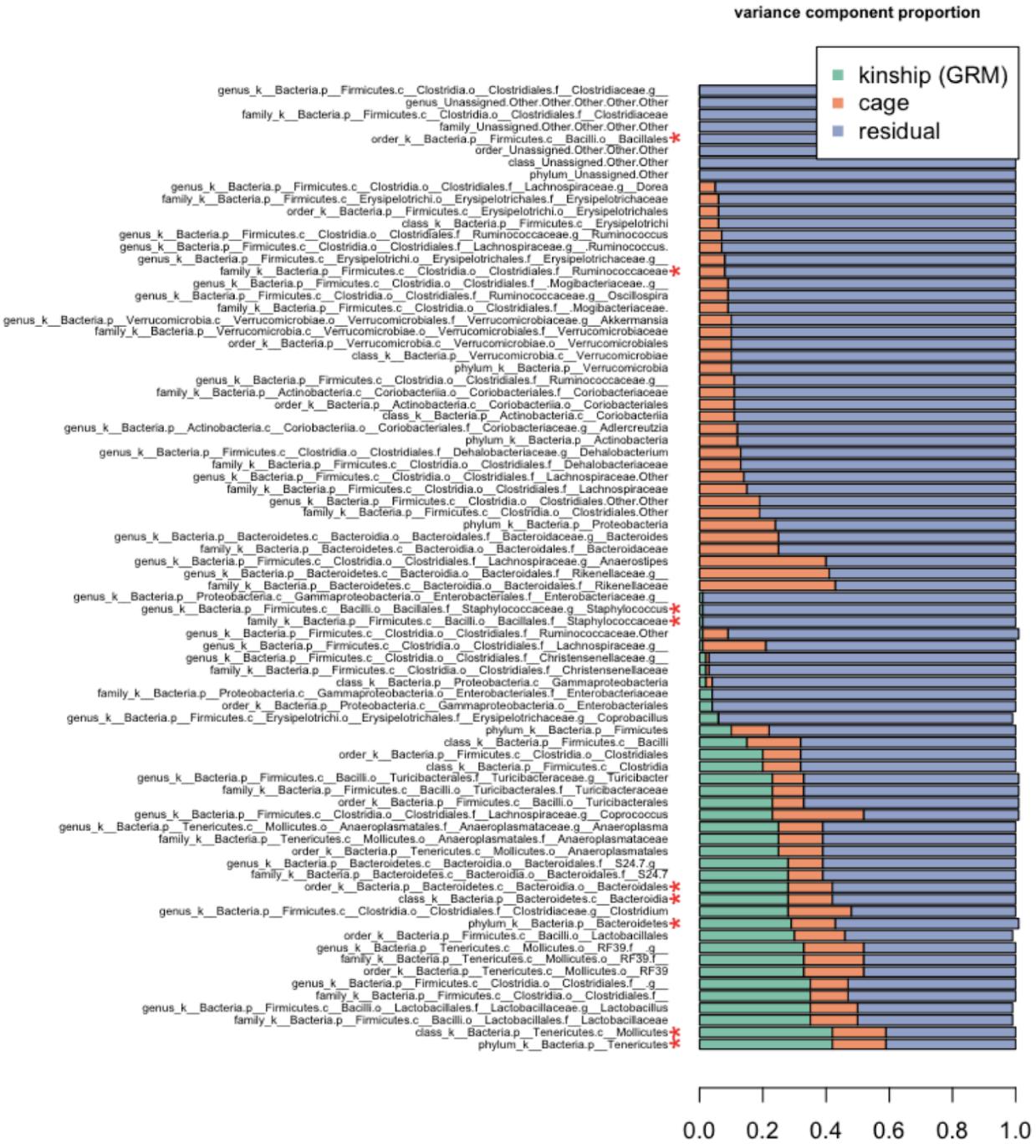


Figure 4.2. Proportion variance estimates for kinship and cage for all taxa. Proportion of variance for each taxon that can be explained by additive effects (heritability) using a kinship or Genomic Relationship Matrix (GRM) (green), cage effects (orange), and unexplained residual effects (blue). Taxa marked with a red asterisk have statistically suggestive QTL hits (adj. *p-value* < 0.1).

The most heritable taxon was the class Mollicutes with a heritability estimate of 39% (*p*-value of 0.002) (**Table 4.1**). Within Mollicutes, an unidentified genus in order RF39 was also found to be highly heritable, with a heritability of 34% (*p*-value 0.010) and the genus *Anaeroplasma* has a heritability of 28% (*p*-value 0.013). Within class Clostridia, an unidentified genus in order Clostridiales showed a heritability of 38% (*p*-value 0.0106). Furthermore, the genus *Lactobacillus* within class Bacilli and the entire Firmicutes phylum were also heritable, at 36% (*p*-value 0.008) and 23% (*p*-value 0.049) respectively. The genus *Turicibacter* within class Bacilli had high heritability estimates as well at 35% (*p*-value 0.0043) and 28% (*p*-value 0.029) respectively. Given the large proportion of the microbiome is composed of either Firmicutes or Bacteroidales, their proportions are strongly negatively correlated. This means that the high heritability of Firmicutes abundance implies also a high heritability of the order Bacteroidales (31%, *p*-value 0.013), as well as an abundant unidentified genus in family S24-7 (32%, *p*-value 0.014).

Table 4.1. Heritability of taxa at five taxonomic levels.

		Heritability %	*p*-value
Taxa	**p_Tenericutes;c_Mollicutes**	**39%**	**0.002**
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Unclassified;g_Unclassified	**38%**	**0.011**
	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;g_*Lactobacillus*	**36%**	**0.008**
	p_Firmicutes;c_Bacilli;o_Turicibacteriales;f_Turicibacteraceae;g_*Turicibacter*	**35%**	**0.043**
	p_Tenericutes;c_Mollicutes;o_RF39;f_Unclassified;g_Unclassified	**34%**	**0.010**
	p_Firmicutes;c_Bacilli;o_Lactobacillales	**32%**	**0.011**
	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_S24.7;g_Unclassified	**32%**	**0.014**
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;g_*Clostridium*	**31%**	**0.022**
	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales	**31%**	**0.013**
	p_Tenericutes;c_Mollicutes;o_Anaeroplasmatales;f_Anaeroplasmataceae;g_*Anaeroplasma*	**28%**	**0.013**
	p_Firmicutes;c_Clostridia;o_Clostridiales	**28%**	**0.029**
	p_Firmicutes;c_Bacilli	**28%**	**0.029**
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_*Coprococcus*	**25%**	**0.019**
	p_Firmicutes	**23%**	**0.049**
	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_*Coprobacillus*	23%	0.063
	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae	20%	0.071

Only showing ranked results with heritability above 20%. Results with *p*-value < 0.05 (statistically significant) are bolded. When results were identical across taxa in the same phylogenetic branch, only the lowest (most specific) taxon was kept. The designations p_, c_, o_, f_, and g_ are for phylum, class, order, family, and genus, respectively. Complete table of heritability results, including rarefied data, can be found in **Tables S4.1A-B**.

QTL Mapping

QTL mapping of the bacterial taxa at the five taxonomic levels revealed significant findings that suggest statistically significant associations between host genotype and abundances of certain taxa. QTL regions on autosomes were found using the R-package *lme4qtl* (ZIYATDINOV *et al.* 2018). Significance was assessed first by comparison of models with and without genotype via a likelihood ratio test,

followed by a genome-wide permutation test. The reported *p*-values were corrected for multiple testing across SNPs (but not across taxa). In total, genetic associations with 3 taxa were found to be statistically significant (adj. *p*-value < 0.05), and genetic associations with 3 additional taxa were statistically suggestive (adj. *p*-value < 0.1) (**Table 4.2**, **Table S4.2A**).

We found statistically significant QTL hits associated with the abundance of family Ruminococcaceae, order Bacillales, and genus *Staphylococcus* (**Table 4.2**). We also found statistically suggestive QTL hits associated with phylum Bacteroidetes, order Bacteroidales, and class Mollicutes. Multiple QTLs for various taxa overlapped with the QTL regions for their parent taxa, such as QTL hit for genus *Staphylococcus* (which is below the taxonomic branch for order Bacillales) overlapping the QTL hit for order Bacillales (**Table 4.2**). These overlaps are a common occurrence in both the significant and non-significant QTLs (**Table S4.2A**).

Table 4.2. QTL regions for taxa at five taxonomic levels.

Taxa	chr[a]	maxlod[b]	pos[c]	from[d]	to[e]	*p*-value	adj.*p*-value	perm.*p*-value
p_Bacteroidetes	5	7.11	118.27	118.01	118.42	2.97E-05	0.089	NA
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales	5	7.20	117.73	117.43	117.76	2.46E-05	0.085	0.105
	5	6.84	117.79	117.79	117.80	5.06E-05	0.085	0.192
	5	7.49	118.58	118.01	118.81	1.38E-05	0.085	0.061
p_Firmicutes;c_Bacilli;o_Bacillales	**19**	**8.37**	**27.02**	**26.55**	**27.42**	**2.38E-06**	**0.042**	**NA**
p_Firmicutes;c_Bacilli;o_Bacillales;f_Staphylococcaceae;g_*Staphylococcus*	**19**	**8.30**	**27.04**	**26.51**	**27.46**	**2.73E-06**	**0.023**	**0.008**
	19	**7.97**	**27.82**	**27.61**	**28.20**	**5.36E-06**	**0.023**	**0.022**
	19	6.68	32.10	31.83	32.28	6.82E-05	0.075	0.243
	19	6.56	32.43	32.43	32.46	8.73E-05	0.078	0.292
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae	**5**	**7.12**	**31.93**	**31.90**	**32.16**	**2.91E-05**	**0.046**	**0.169**
	5	**7.27**	**32.52**	**32.27**	**33.36**	**2.14E-05**	**0.046**	**0.111**
	2	**6.87**	**170.57**	**170.51**	**170.64**	**4.71E-05**	**0.048**	**0.09**
p_Tenericutes;c_Mollicutes	1	7.00	121.32	120.23	125.20	3.66E-05	0.089	0.155

[a] Chromosome in which lies the QTL
[b] Maximum LOD score within the QTL
[c] Position in the chromosome where the maximum LOD score is found
[d] Chromosomal position where the QTL begins
[e] Chromosomal position where the QTL ends

Only showing ranked results with adj. *p*-value < 0.1 (statistically suggestive). Results with adj. *p*-value < 0.05 (statistically significant) are bolded. When results were overlapping across taxa in the same phylogenetic branch (such as p_Bacteroidetes and o_Bacteroidales), permutations were calculated only for the lowest (most specific) taxon. The designations p_, c_, o_, f_, and g_ are for phylum, class, order, family, and genus, respectively. Complete table of QTL results, including rarefied data, can be found in **Tables S4.2A-B**.

Looking at specific QTL peaks, we identified the genes *Insig2* and *Ksr2* on the highest point in the region for the class Mollicutes (chr1:121,315,223, LOD = 7.002) and the order Bacteroidales (chr5:117,733,508, LOD = 7.203) respectively. INSIG2 plays a central role in the pathway by which the circadian clock regulates liver lipid metabolism (ZHANG *et al.* 2017) and *Ksr2* has been implicated in being associated with BMI and severe early-onset obesity through large scale GWAS studies (MILANESCHI *et al.* 2019).

OTU level analysis

Next, we decided to increase the specificity of the taxonomic classifications to operational taxonomic units (OTUs) by compiling all OTUs identified within taxa that had statistically suggestive QTL peaks (**Table 4.2**). We filtered out OTUs that were present in less than 50% of the samples, resulting in 362 OTUs. QTL mapping performed on these selected OTUs resulted in 59 OTUs with at least one statistically suggestive association. Additionally, 99 OTUs were found to be heritable ($h^2 > 20\%$, p-value < 0.05), of which 28 OTUs also had statistically suggestive QTLs (**Tables S4.3 and S4.4**). Proportion variance estimates for kinship and cage for all tested OTUs are presented in **Figure S4.3**.

QTL associations to OTUs varied compared to overlapping QTL regions associated to taxa at higher taxonomic levels, some were sharper and stronger, others were less specific and wider (**Table 4.3**). These results are interesting because a sharper QTL peak associated with an OTU may suggest that the overlapping QTL region associated with the broader taxonomic group is being driven by that specific OTU. On the other hand, if the overlapping QTL region associated with the broader taxonomic group is smaller and more specific than the region seen on an individual OTU, this might suggest a cumulative effect of multiple sub-taxonomies driving a stronger signal at the broader taxonomic level. For example, QTL hits for OTU

338796 and New.CleanUp.ReferenceOTU 170146 within family Ruminococcaceae were both statistically significant and overlapped with the QTL region for Ruminococcaceae, but the QTLs for the OTUs were both wider.

Table 4.3. QTL regions for OTUs.

	chr	maxlod	pos	from	to	*p*-value	adj.*p*-value	perm.*p*-value
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;OTU_421792	5	5.87	118.69	118.63	118.82	3.26E-04	0.076	0.642
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;OTU_460953	5	5.79	118.69	118.58	118.79	3.84E-04	0.078	NA
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;OTU_190835	5	5.67	118.67	118.58	118.74	4.77E-04	0.076	NA
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;OTU_209408	5	7.02	118.67	118.50	118.82	3.53E-05	0.064	NA
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_NR.OTU100	5	8.22	31.93	30.25	32.06	3.23E-06	0.064	0.023
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_338796	**2**	**8.15**	**170.57**	**169.64**	**170.96**	**3.74E-06**	**0.023**	**0.023**
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_NR.OTU95	5	7.65	31.83	31.34	32.00	1.01E-05	0.076	0.088
	5	7.36	32,27	32.11	32.40	1.80E-05	0.076	0.048
	5	7.36	32.27	32.11	32.39	1.80E-05	0.076	0.081
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_336810	2	6.32	170.54	170.48	170.56	1.39E-04	0.100	NA
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_NCR.OTU170146	**5**	**7.63**	**36.01**	**32.27**	**36.22**	**1.05E-05**	**0.030**	**NA**

Taxa

Only showing OTUs with adj. *p*-value < 0.1 (statistically suggestive) and with a QTL region overlapping QTLs from higher-level taxonomies. Results with adj. *p*-value < 0.05 (statistically significant) are bolded. Complete table of QTL results for OTUs can be found in **Tables S4.4**.

Comparison to other studies

Results from other published studies on heritabilities of the various bacterial taxa in the gut microbiome were compiled and compared with our results, both in rarefied and non-rarefied datasets (**Figure 4.3**). A full comparison of heritabilities among all analyzed taxa in our study and other studies can be found in **Table S4.5.**

Family S24-7 within order Bacteroidales had a high heritability in our study (h^2 = 0.32) and it has been reported as heritable and significant in both mice (h^2 = 0.60 in

(ORG *et al.* 2015)) and humans (h^2 = 0.33 in (TURPIN *et al.* 2016)) (**Figure 4.3**). Genus *Lactobacillus* was found to have a high and significant heritability (h^2 = 0.36) and it was also found to be highly heritable in one mouse study (h^2 = 0.74 in (O'CONNOR *et al.* 2014)) and both highly heritable and significant in a pig study (h^2 = 0.34 in (CAMARINHA-SILVA *et al.* 2017)) and multiple human studies (h^2 = 0.36 in (DAVENPORT *et al.* 2015), 0.26 in (TURPIN *et al.* 2016), and 0.15 in (LIM *et al.* 2017)). Genus *Turicibacter*, also within class Bacilli, was found to have a high and significant heritability as well (h^2 = 0.57) and was found to be highly heritable in one mouse study (h^2 = 0.54 in (ORG *et al.* 2015)). *Turicibacter* was also found to be significantly heritable in human studies (h^2 = 0.26 in (TURPIN *et al.* 2016) and 0.36 in (GOODRICH *et al.* 2016)). Under the class Clostridia and still within the phylum Firmicutes, family Christensenellaceae in our rarefied dataset had a high, significant heritability (h^2 = 0.31) that did not appear in our non-rarefied dataset. In human studies, Christensenellaceae has been found to be highly heritable and statistically significant (h^2 = 0.64 in (TURPIN *et al.* 2016), 0.42 in (GOODRICH *et al.* 2016), and 0.31 in (LIM *et al.* 2017)). Additionally, we found genus *Clostridium* to have a high and significant heritability (h^2 = 0.31) that was also seen in other human studies as well (h^2 = 0.24 in (GOODRICH *et al.* 2016) and 0.46 in (DAVENPORT *et al.* 2015)). The genus *Coprococcus* was also significant and highly heritable from our study (h^2 = 0.25) as well as in various human studies (h^2 = 0.46 in (DAVENPORT *et al.* 2015) and 0.16 in (LIM *et al.* 2017)).

The clade within the phylum Tenericutes including genus *Anaeroplasma* gave high and statistically significant heritabilities in both our rarefied and non-rarefied datasets (**Figure 4.3**). Few other studies found similar results either because they did not include these taxa in their study or their results gave weaker heritabilities with non-significant *p*-values. Nonetheless, one mouse study did find a high heritability for genus *Anaeroplasma* (h^2 = 0.48 in (O'CONNOR *et al.* 2014)). In human studies, significant heritabilities were found for phylum Tenericutes (h^2 = 0.34 in (GOODRICH *et al.* 2016) and 0.23 in (LIM *et al.* 2017)), class Mollicutes (h^2 = 0.32 in (GOODRICH *et al.* 2016) and 0.23 in (LIM *et al.* 2017)), and order *RF39* (h^2 = 0.31 in (GOODRICH *et al.* 2016)).

Both our non-rarefied and rarefied datasets gave insignificant heritability estimates of 0.02 for the genus *Akkermansia* and all the way up its taxonomic branch to phylum Verrucomicrobia, yet estimates from (ORG *et al.* 2015) were as high as h^2 = 0.92, and heritability of *Akkermansia* from (O'CONNOR *et al.* 2014) was h^2 = 0.62 in mice. Studies conducted using human microbiome samples show a diversity of heritability estimates for this taxonomic branch: moderately high and significant (h^2 = 0.30 for both in (TURPIN *et al.* 2016)), low and significant (h^2 = 0.15 for Verrucomicrobia and h^2 = 0.14 for *Akkermansia* in (GOODRICH *et al.* 2016)), and close to zero and non-significant (up to h^2 = 0.01 for *Akkermansia* in (DAVENPORT *et al.* 2015), and 0.05 for Verrucomicrobia and 0.06 for *Akkermansia* in (LIM *et al.* 2017)).

	Heritability														
	Mouse								Human						Pigs
	Our		Org					O'Connor	Goodrich	Davenport			Turpin	Lim	Camarinha-
	nonR	R	All	M	F	Avg	One		'16	W	S	C			Silva
p_Bacteroidetes	**0.31**	**0.32**	0.53	0.82	0.73	0.00	0.02		0.08				**0.33**	**0.25**	0.20
c_Bacteroidia	**0.31**	**0.32**	0.53	0.82	0.73	0.00	0.02		0.08				**0.33**		0.20
o_Bacteroidales	**0.31**	**0.32**	0.53	0.82	0.73	0.00	0.02		0.08		0.00	0.00	**0.33**	**0.25**	0.00
f_S24-7	**0.32**	**0.32**	0.60	0.86	0.82	0.00	0.00						**0.33**		
p_Firmicutes	**0.23**	0.22	0.56	0.71	0.77	0.15	0.16		0.00	0.00	0.00	0.00	0.18	0.10	
c_Bacilli	**0.28**	**0.24**	0.68	0.74	0.76	0.01	0.00		0.03				0.19	**0.35**	
o_Lactobacillales	**0.32**	**0.35**	0.77	0.79	0.55	0.00	0.00		0.00				0.10	**0.33**	
f_Lactobacillaceae	**0.36**	**0.37**							0.04		0.13		**0.26**	**0.17**	0.00
g_*Lactobacillus*	**0.36**	**0.37**						0.74	0.04	**0.36**	0.00	0.19	**0.26**	**0.15**	**0.34**
o_Turicibacterales	**0.35**		0.54	0.75	0.82	0.12	0.12		**0.39**				**0.26**		
f_Turicibacteraceae	**0.35**		0.54	0.75	0.82	0.12	0.12		**0.39**				**0.26**		
g_*Turicibacter*	**0.35**		0.54	0.75	0.82	0.12	0.12	0.29	**0.39**	0.00	0.19	0.13	**0.26**		0.00
c_Clostridia	**0.28**	**0.27**	0.58	0.80	0.77	0.00	0.03		0.03				0.22	0.07	0.00
o_Clostridiales	**0.28**	**0.27**	0.58	0.80	0.77	0.00	0.03		0.03	0.00	0.00	0.00	**0.33**	0.07	0.00
f_Christensenellaceae	0.18	**0.31**							**0.42**				**0.64**	**0.31**	
f_Clostridiaceae	0.00	0.00	0.61	0.83	0.80	0.09	0.05		**0.30**	**0.35**		0.00	**0.35**	**0.34**	
g_*Clostridium*	**0.31**								**0.24**	0.10	**0.46**	0.04	0.20	0.14	0.10
f_Lachnospiraceae	0.07	0.08	0.52	0.60	0.69	**0.36**	0.07		0.16	0.13	0.00	**0.29**	0.17	**0.15**	0.09
g_*Coprococcus*	**0.25**	0.29	0.28	0.61	0.55	0.00	0.02	0.19	0.09	**0.46**	0.06	**0.26**	0.04	**0.16**	
p_Tenericutes	**0.39**	**0.39**							**0.34**				0.06	**0.23**	
c_Mollicutes	**0.39**	**0.39**							**0.32**				0.18	**0.23**	
o_Anaeroplasmatales	**0.28**	**0.29**													
f_Anaeroplasmataceae	**0.28**	**0.29**													
g_*Anaeroplasma*	**0.28**	**0.29**						0.48							
o_RF39	**0.34**	**0.32**							**0.31**				0.18		
p_Verrucomicrobia	0.02	0.11	0.54	0.85	0.92	0.13	0.33		**0.15**				**0.30**	0.05	
c_Verrucomicrobiae	0.02	0.11	0.54	0.85	0.92	0.13	0.33		**0.14**				**0.30**		
o_Verrucomicrobiales	0.02	0.11	0.54	0.85	0.92	0.13	0.33		**0.14**				**0.30**		
f_Verrucomicrobiaceae	0.02	0.11	0.54	0.85	0.92	0.13	0.33		**0.14**				**0.30**	0.06	
g_*Akkermansia*	0.02	0.11	0.54	0.85	0.92	0.13	0.33	0.62	**0.14**	0.00	0.01	0.00	**0.30**	0.06	

Figure 4.3. Comparison of taxon heritabilities across mouse, human, and pig studies. The green shading over heritability estimates ranges from lowest heritability estimate (white) to highest heritability estimate (green) in a given study. Statistically significant results are shown in bold font. For our Diversity Outbred study, we report both non-rarefied (nonR) and rarefied (R) results. For (ORG *et al.* 2015) we report results using all mice (All), just males (M), just females (F), an average per strain (Avg), and a single mouse per strain (One). (ORG *et al.* 2015) and (O'CONNOR *et al.* 2014) did not report significances. For (GOODRICH *et al.* 2016) the estimates are calculated by the ACE model, bold values indicate estimates with a 95% confidence interval not overlapping 0. For (DAVENPORT *et al.* 2015) the estimates are the proportion of variance explained (PVE) estimates ("chip heritability"), we report winter (W), summer (S), and combined seasons (C) datasets, and bold values indicate estimates with a standard error not overlapping 0. For (TURPIN *et al.* 2016) and (LIM *et al.* 2017) estimates are polygenic heritability (H2r). For (CAMARINHA-SILVA *et al.* 2017) estimates are narrow-sense heritability (h^2). Grey indicates that the taxon was not observed or excluded in a given study. Figure adapted from (GOODRICH *et al.* 2016). Selected comparisons shown, full comparison found in **Table S4.5**.

In addition to comparing our heritability estimates with other studies, we also contrasted our QTL mapping results of the gut microbiome with those from previous findings (**Figure 4.4**). A full comparison of QTLs among all analyzed taxa in our study and other studies can be found in **Table S4.5**.

We identified statistically significant QTL associations for the order Bacillales as well as for the family Staphylococcaceae and the genus *Staphylococcus* within Bacillales in chromosome 19; another mouse study also found statistically significant QTL associations for all of the same taxa but on chromosome 17 (MCKNITE *et al.* 2012). A human microbiome study found statistically significant QTL regions for the class Bacilli, which comprise the above mentioned order and families (BLEKHMAN *et al.* 2015.

Family Ruminococcaceae has been previously found to have significant QTL associations both in mice (chromosome 12, (BENSON *et al.* 2010)) and humans (BLEKHMAN *et al.* 2015). In our study, Ruminococcaceae was identified to be associated with chromosomes 2 and 5. We also identified a QTL hit for the phylum Bacteroidetes in chromosome 5 while another mouse study identified a significant hit in chromosome 14 (WANG *et al.* 2015). Within Bacteroidetes, even though we did not find any significant QTL results for the genus *Bacteroides*, many other mouse studies did (chromosomes 1 (WANG *et al.* 2015), 4 (MCKNITE *et al.* 2012), 9 (LEAMY *et al.*

2014), 11 (BUBIER *et al.* 2018), 16 (LEAMY *et al.* 2014), and 18 (LEAMY *et al.* 2014)) as well as a human study (BLEKHMAN *et al.* 2015).

Phylum Tenericutes had a significant hit in chromosome 1 in both our non-rarefied and rarefied datasets, and family Lachnospiraceae had a statistically suggestive QTL in chromosome 10 in our rarefied dataset but not in our non-rarefied dataset. Both of these taxa had significant QTL hits in a human study (BLEKHMAN *et al.* 2015).

	QTL/GWAS signals							
	Mouse							Human
	Our		Benson	McKnite	Leamy	Wang '15 H	Bubier	Blekhman
	nonR	R						
p_Bacteroidetes	5	5	no	no		**14**		no
└c_Bacteroidia	5	5						no
└o_Bacteroidales	5	5	no					no
└f_Bacteroidaceae			no	no				no
└g_*Bacteroides*			no	**4**	**9,16,18**	**1**	**11**	**9**
p_Firmicutes			no	no				no
└c_Bacilli			no					**2,10,14**
└o_Bacillales	**19**		no	**17**				
└f_Staphylococcaceae	**19**			**17**				
└g_*Staphylococcus*	**19**			**17**				
└c_Clostridia			no					no
└o_Clostridiales			no			3		no
└f_Lachnospiraceae		**10**	no	no				**1**
└f_Ruminococcaceae	**2,5**	**2,5**	**12**	no				**10**
p_Tenericutes	**1**	**1**						**6**
└c_Mollicutes	**1**	**1**						

Figure 4.4. Comparison of taxa with QTL associations across mouse and human studies. Associations with each taxon are marked in blue if statistically suggestive and bolded if statistically significant, or dark grey if not significant. The chromosome number were the QTLs were found are denoted in each box. Light gray indicates that the taxon was not observed or excluded in a given study. For our Diversity Outbred study, we report both non-rarefied (nonR) and rarefied (R) results. Figure adapted from (GOODRICH *et al.* 2016). Selected comparisons shown, full comparison found in **Table S4.5**.

Gene level analysis

Examining the QTL mapping results from previous studies, it was apparent that although different studies might all have found significant QTL regions for a particular bacteria taxon, they identified different genomic positions as showing associations. In order to identify common pathways shared by different QTL regions, we ran a cumulative geneset pathway analysis on the genes within our identified regions and the genes within the regions indicated in other studies. In total, there were 60 significant QTL hits with an additional 256 suggestive hits across the six taxonomic levels (phylum, class, order, family, genus, and OTU) (**Table S4.2 and S4.4**).

Of the analyzed gene subsets, the collection of genes within QTLs among the taxa and OTUs that fall under the family Ruminococcaceae returned the most significant results. The Ingenuity Pathway Analysis (IPA) software was employed to analyze and categorize our geneset (IPA®, QIAGEN Redwood City, CA). Overall, 372 genes from 58 statistically significant and suggestive Ruminococcaceae QTLs (**Table S4.6**) were submitted to IPA. A core analysis to find associated pathways and diseases generated multiple gene networks that revealed genes strongly associated with ovarian, breast, and colon cancer pathways (**Figure 4.5A-B)**.

Five genes, *Vegfa*, *Kat2b*, *Smad4*, *Fgfr2*, and *Yes1*, from our gene set were found to be highly linked to ovarian cancer (p-value = 5.43 x 10^{-6}) (**Figure 4.5A**). Although

all of these genes have been identified to be related to ovarian cancer pathways, we recognized *Smad4* as a well-known and prevalent tumor suppressor gene. SMAD4 mediates the TGF-beta signaling pathway and occurs frequently in pancreatic and colorectal cancers with malignant progression and appears occasionally in other human cancers (MIYAKI AND KUROKI 2003). Additionally, five genes from our input gene set were found to be significantly associated with breast cancer pathways (*p*-value = 1.17×10^{-5}) (**Figure 4.5B**). Though the results are sparse and under studied, Ruminococcaceae abundance and breast cancer have been linked in previous studies. One study shows that Ruminococcaceae abundance was significantly higher in postmenopausal breast cancer patients when compared to normal healthy patients (YANG *et al.* 2017).

Genes from our significant Ruminococcaceae QTL peaks were also found to be associated with colon carcinoma (*p*-value = 6.67×10^{-5}) and colorectal carcinoma (*p*-value = 1.75×10^{-4}) (**Figure 4.5A**) and associations between these bacteria and colorectal cancer (CRC) have been studied before. Ruminococcaceae was found to be significantly less abundant in cancerous colorectal tissue compared to healthy intestinal lumen (CHEN *et al.* 2012). Furthermore, another study showed findings that suggest Ruminococcaceae provides beneficial effects against risk of colorectal cancer (ERICSSON *et al.* 2015).

In addition to having genes associated with specific cancers, various genes from our Ruminococcaceae gene set were found to have direct interactions with the well-known and prevalent cancer gene *Tp53*. **Figure 4.5C** from IPA depicts a gene network containing 21 genes from the Ingenuity Knowledge Base and 14 genes from our input gene set, 5 of which (*Sorbs1*, *Stau1*, *Cox15*, *Ran*, and *Glb1*) have direct interactions with the widely known tumor suppressor gene *Tp53*. TP53 has been shown to be a critical player in tumor development and how tumor cells avoid apoptosis, and mutations in *Tp53* have been identified in numerous types of cancers (LEVINE *et al.* 1991; GREENBLATT *et al.* 1994; PETITJEAN *et al.* 2007; VOGELSTEIN *et al.* 2010).

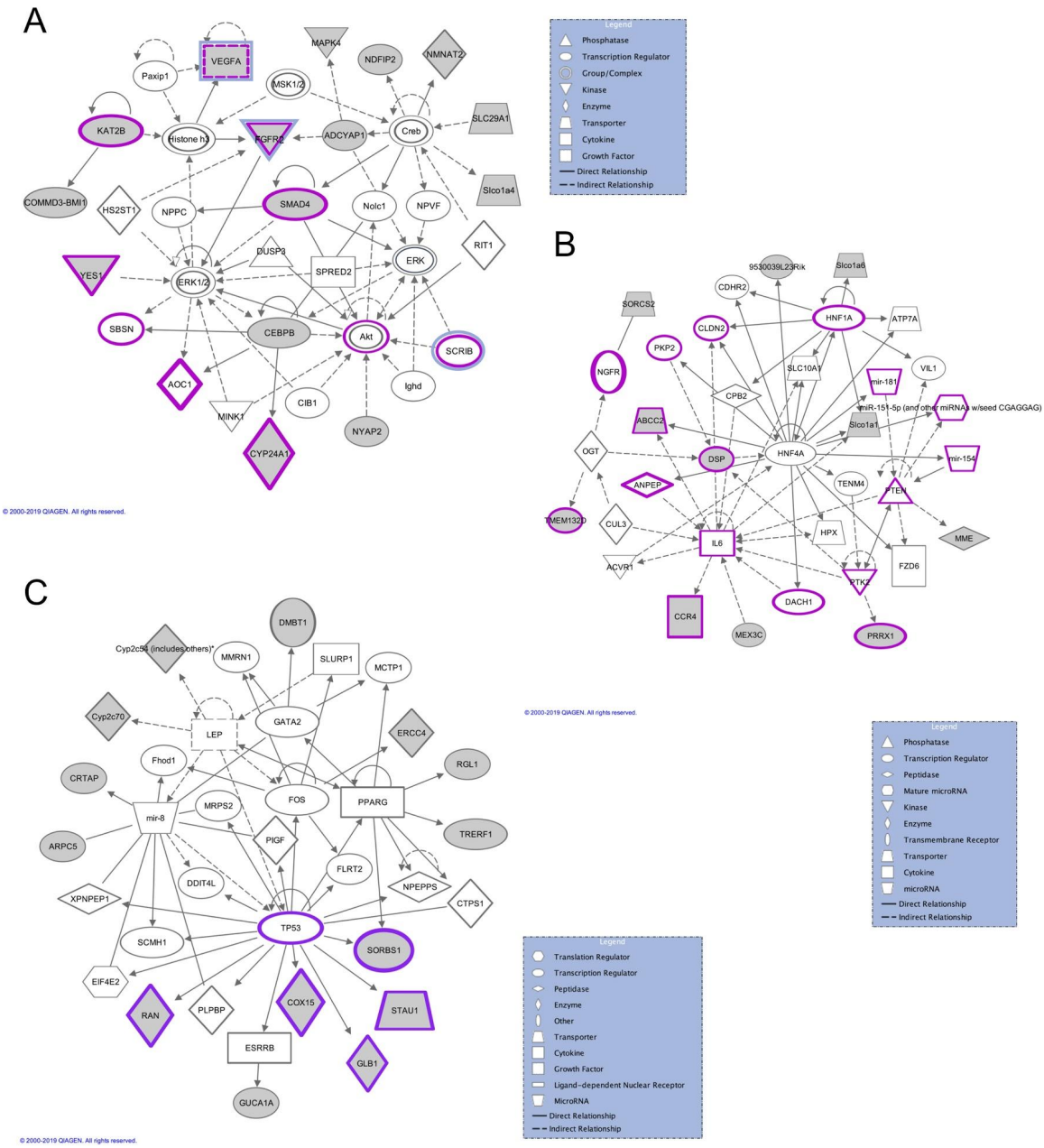


Figure 4.5. Ingenuity Pathway Analysis (IPA) interaction network generated from genes within Ruminococcaceae QTLs. Genes circled in color are all part of specific associated pathways as specified below. Genes colored in gray belong to our dataset whereas un-colored genes are other closely associated genes added by IPA. Refer to **Tables S4.7A-C** for a list of these associated genes from our dataset. (**A**) The network shows genes found within Ruminococcaceae QTLs strongly associate with pathways related to ovarian cancer (circled in pink) and colon carcinoma and colorectal carcinoma (circled in light blue). (**B**) The network shows genes found within Ruminococcaceae QTLs strongly associate with pathways related to breast cancer (circled in pink). (**C**) The network shows genes found within Ruminococcaceae QTLs strongly associated with hallmark cancer gene *Tp53* (circled in purple).

In addition to Ruminococcaceae, we compiled the results from all of our significant QTL peaks under the order Bacillales and used the genes from within the QTL regions to run gene set functional pathway analyses and found these bacteria to be highly associated with pathways involved in lipid and sphingolipid metabolism. IPA identified our input genes *Vldlr* and *Sgms1* to be related to multiple lipid metabolism pathways (p-value = 5.23 x 10^{-8} to 5.14 x 10^{-4}) (**Figure S4.4**, **Table S4.7A**) and DAVID functional annotation tool (HUANG *et al.* 2008) identified our genes *Vldlr*, *Sgms1*, and *Asah2* to be related to lipid metabolism (p-value = 0.0048) as well as genes *Sgms1* and *Asah2* to be related to sphingolipid metabolism (p-value = 0.0063) (**Table S4.7B**). Associations between gut microbiota and host lipid metabolism have been investigated previously, and proof of causality between specific microbial associations with lipid metabolism and sphingolipid production has been demonstrated (GHAZALPOUR *et al.* 2016; HEAVER *et al.* 2018; BROWN *et al.* 2019; JOHNSON *et al.* 2019).

DISCUSSION

There exists a complex and multifaceted relationship between the gut microbiome and its host's genome, where recent studies are beginning to show the true magnitude of these connections. Our results seek to further understand this

relationship by identifying functional and disease pathways that may be associated with specific bacterial abundances in the mouse gut microbiome.

SNPs with the highest LOD in the QTL regions for Mollicutes and Bacteroidales were found to lie within genes *Insig2* and *Ksr2* respectively. *Insig2* encodes a transmembrane protein that releases SREBP proteins to the endoplasmic reticulum where they exert control over lipid metabolism (PASCHOS AND FITZGERALD 2017). Relationships between gut microbiome and lipid metabolism have already been established (LI *et al.* 2008; VELAGAPUDI *et al.* 2010), and our reported association between Mollicutes and *Insig2* further suggest some kind of interaction between Mollicutes abundance and lipid metabolism. Gene *Ksr2* is known to be associated with BMI and early-onset obesity, as *Ksr2* variants impair cellular fatty acid oxidation and glucose oxidation, often leading to hyperphagia, low heart rate, reduced basal metabolic rate, and severe insulin resistance (PEARCE *et al.* 2013). This provides potential pathways by which *Ksr2* may lead to severe cases of obesity. Additionally, Bacteroidetes relative abundance has been shown to be 50% lower in genetically obese *ob/ob* mice compared to lean mice while Firmicutes relative abundance was higher by a corresponding amount (LEY *et al.* 2005). The association we find between Bacteroidales and *Ksr2* may suggests a potential relationship between Bacteroidales abundance and risk for obesity.

Using the total set of genes from within all 58 statistically significant and suggestive QTL regions for taxa within the family Ruminococcaceae, we identified multiple networks, each of 35 functionally interrelated genes, enriched in disease pathways for ovarian, breast, and colon cancer. Evidence of functional associations between ovarian cancer and Ruminococcaceae is lacking, but various studies have confirmed findings showing increased Ruminococcaceae abundance in breast cancer patients compared to normal healthy individuals (FERNÁNDEZ *et al.* 2018; ZHU *et al.* 2018). While these studies did not uncover a directionality to this association, the significant differences in microbiome composition could be used as independent biomarkers of breast cancer (ZHU *et al.* 2018). In addition to specific links between the family Ruminococcaceae and breast cancer, associations between the gut microbiome and breast cancer have been flagged (FERNÁNDEZ *et al.* 2018). This includes associations between perturbations in the gut microbiome and circulating estrogen levels and metabolites, produced by several bacteria including *Ruminococcaceae* and also known as the estrobolome, which can affect the risk for breast cancer (PLOTTEL AND BLASER 2011; FUHRMAN *et al.* 2014). Indeed, the gut microbiome can influence estrogen metabolism through enterohepatic circulation (ADLERCREUTZ *et al.* 1984; FLORES *et al.* 2012), and thus could be implicated in breast cancer by interacting with estrogen metabolism (MINELLI *et al.* 1990; GOEDERT *et al.* 2015). Outside the gut microbiome, a study looked at the relationship between the breast tissue microbiome and breast cancer and also found significantly different microbiome composition and

functions between women with benign and malignant breast disease (HIEKEN *et al.* 2016). In aggregate, these studies support a role for microbes in the risk of breast carcinogenesis and our study extends this relationship by identifying specific genes involved in breast cancer pathways that may mediate this connection.

Our functional gene networks also revealed genes involved in colon cancer pathways. Ruminococcaceae abundance has been shown to be negatively correlated with risk for colorectal cancer (CRC) (CHEN *et al.* 2012; ERICSSON *et al.* 2015). Looking beyond the specificity of Ruminococcaceae, various other studies have shown strong evidence for a link between the gut microbiome and risk for CRC. Microbiota in the colon form biofilms that line the mucosal surface, and a study has shown evidence suggesting that this biofilm structure may impact cellular proliferation and cancer growth by affecting the metabolome and down-regulating or up-regulating the production and release of metabolites favorable for tumor cells (JOHNSON *et al.* 2015). General decreased microbial community diversity has been shown to be significantly correlated with risk for CRC in a study that compared CRC case subjects to control healthy subjects (AHN *et al.* 2013). Additionally, a study identified the enrichment and depletion of several bacterial populations associated with CRC and used this information in addition to known clinical risk factors for CRC to build a predictive model for evaluating risk for CRC. Used as a screening tool, this new predictive model that included microbial abundances improved accuracy by more than

50 folds (ZACKULAR *et al.* 2014). This not only confirms the existence of strong associations between the gut microbiome and CRC, but also raises the possibility that these data may be used as a potential diagnostic tool for clinical purposes.

In addition to revealing potential disease pathways associated with Ruminococcaceae, our geneset pathway analysis also unveiled connections between multiple genes to the well-characterized cancer gene *Tp53*. Genes *Sorbs1* and *Stau1* found in our QTL analysis have been shown to be down-regulated in cells that have undergone p53-mediated immortalization and transformation as a direct or indirect result of Ras signaling activity (BOIKO *et al.* 2006). Furthermore, another study showed through gene ontology analysis that p53 regulates various mitochondrial bioenergetic pathways including the up-regulation of our gene *Cox15* involved in ATP synthesis (MAK *et al.* 2017). The same study also found that p53 regulates various genes involved in cardiac tissue function including the down-regulation of our gene *Ran* involved in major signal transduction pathways (MAK *et al.* 2017). P53 was further found to decrease the activity of mouse SA beta-Galactosidase protein (encoded by our gene *Glb1*) in mouse mesothelial cells as well as in mouse embryonic fibroblast cells (PIETRUSKA AND KANE 2007; WANG *et al.* 2007). With multiple genes from within our significant Ruminococcaceae QTL peaks exhibiting interactions with the popular tumor suppressor gene *Tp53*, it is highly suggestive that Ruminococcaceae

abundance may be in some way linked to cancer development and tumor cell proliferation.

Similar geneset pathway analysis was conducted for the QTLs under the order Bacillales and significant associations were found between various genes and lipid metabolism. Although specific interactions between Bacillales and lipid metabolism have not been thoroughly studied before, previous studies have elucidated a relationship between the gut microbiome and the metabolome. One study discovered increased energy metabolites in conventionally raised mice compared to germ free mice and further found microbiome composition to influence levels of various lipid classes, most significantly on triglyceride and phosphatidylcholine molecular species (VELAGAPUDI *et al.* 2010). Furthermore, systems biology analysis comparing human baby microbiota to normal microbiota in mice found that metabolism of dietary lipids was specifically influenced by the microbiome (LI *et al.* 2008). In mouse, a study confirmed the microbiome to exert a strong impact on the metabolism of bile acids with increased bile acid levels in various gut compartments in germ free mice, suggesting that gut microbiome composition may affect host lipid metabolism through bile acid metabolism (CLAUS *et al.* 2008).

A concern with performing microbiome analysis is that the standard data processing method of rarefaction of counts causes notable losses of data and loss of power leading to missed associations (MCMURDIE AND HOLMES 2014). We evaluated

the impact of rarefaction on microbial abundances by clustering rarefied and non-rarefied taxa together by correlation of frequency of counts within each taxonomic level. The majority of the rarefied taxa correlated with their non-rarefied counterparts (**Figure S4.2**). Regardless of this similarity, we conducted all analysis in parallel for non-rarefied and rarefied datasets. Looking at the significantly associated QTLs within various taxa from non-rarefied and rarefied datasets, we notice some differences in the significance of the QTLs and the chromosome in which they reside (**Table S4.2A-B**). While several microbial taxa associations with QTLs were consistent across non-rarefied and rarefied datasets, there were some instances of statistically significant associations being found in only one of the datasets.

Comparing our results with other studies, we found little overlap in the specific bacterial taxa studied as well as the calculated heritabilities and QTL results. This is most likely due to the limited number of existing studies discussing heritabilities and QTL mappings of bacteria within the gut microbiome. Additionally, the absence of a standardized methodology for performing these studies leads to use of different procedures and analytical methods, making it increasingly difficult to compare results across studies (GOODRICH *et al.* 2017). Ultimately, the current state of the field for profiling different characteristics of the gut microbiome is still rapidly evolving and as it matures and more studies are undertaken, it will become easier to compare and validate results.

Although our results support the claim that host genetics can impact the gut microbiome composition in ways that are relevant to the health of the host, our study has some limitations. The biggest limitation to the power of the study is its relatively small sample size (n = 247 DO mice). Conducting QTL mapping with small sample sizes may lead to the 'Beavis effect' which is a failure to detect QTL of small effect sizes as well as an overestimation of effect size of the QTLs that are discovered (MILES AND WAYNE 2008). Our study also shares all the weaknesses common to the Diversity Outbred design: since the genome of each mouse is a unique mosaic of the 8 strains from the CC population, the genotype of each DO mouse is irreproducible. This limits the amount and manner of phenotyping that can be done, and it makes replicating results within the DO population difficult. However, this limitation could be partially circumvented by using the CC lines as a form of validation, since they can provide reproducible genotypes (SVENSON *et al.* 2012). Another limitation is the current lack of experimental validations of associations between disease pathways (such as those for ovarian, breast, and colon cancer) and specific taxa within gut microbiome composition, making it difficult to confirm any associations we find between genes and bacterial abundances.

Our results provide insight into the complex interplay between host genetics and the gut microbiome, and isolate potential associations between microbial taxa and QTLs that may be involved in pathological disease phenotypes. Additional studies are required to verify associations between specific genes and taxon abundance in the gut microbiome, such as performing gene knockouts and observing the effects on microbiome composition. While most of the variation in the gut microbiome composition is not due to genetics but rather environmental factors (ROTHSCHILD *et al.* 2018), attributes of the gut microbiome that are clearly heritable may provide important insights about host-microbiome interactions and mechanisms that impact microbiome composition. The direct genotype-phenotype association approach in this study could be applied to illuminate novel associations between genetic variants and their effects on microbial abundances involved in the microbiome through the mechanism of a complex disease of interest. Understanding the interactions between a host's genome and its microbiome composition may also aid in our understanding of complex diseases and their mechanisms and potentially aid in developing medical treatments.

ACKNOWLEDGEMENTS

The authors want to thank Noah Clark, Jessica L. Sutter, Qiaojuan Shi, Emily Davenport, Angela C. Poole, and Afrah Shafquat for all the help and advice provided. FS was supported by a Presidential Life Science Fellowship (PLSF) from Cornell University.

AUTHOR CONTRIBUTIONS

F.S., A.G.C, G.A.C, and R.E.L. conceived the study. A.P. provided the samples. F.S. extracted and generated the 16S rRNA gene sequencing data. F.S., E.C., P.S., J.K.G., R.E.L., G.A.C, and A.G.C. conceived the computational and statistical analyses. F.S., D.Y.Z., E.C., M.E., and P.S. performed the computational and statistical analyses. F.S., D.Y.Z., E.C, and A.G.C. wrote the manuscript.

SUPPORTING INFORMATION

Figure S4.1. Taxa relative abundance frequencies

Figure S4.2. Correlation plot between non-rarefied and rarefied taxa

Figure S4.3. Proportion variance estimates for kinship and cage for all OTUs

Figure S4.4. IPA network for Bacillales QTL

Table S4.1. Heritability results at 5 taxonomic levels

Table S4.2. QTL results at 5 taxonomic levels

Table S4.3. Heritability results at OTU level in non-rarefied dataset

Table S4.4. QTL results at OTU level in non-rarefied dataset

Table S4.5. Comparison of heritabilities and QTLs with other studies

Table S4.6. Ruminococcaceae genes used for gene-set analysis

Table S4.7. Genes included in networks from **Figure 4.5**

Table S4.8. Genes from gene set analysis using QTLs under Bacillales

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

In this thesis, I presented three orthogonal approaches for surveying genetic variation and its consequences. I used a combination of data collected through three different sequencing methods: In **Chapter 2** I study population genomic data using genotyping, in **Chapter 3** I profile global transcriptome dynamics using RNA-sequencing, and in **Chapter 4** I investigate microbiome composition using 16S rRNA sequencing. Below I synthesize the insights from these studies and discuss general conclusions and future directions for all three projects.

Inferring regions of positive selection in population genomic data

The availability and collection of population genomic data allow us to establish links between genetic variation and phenotypic adaptation by measuring patterns of variation among individuals. Motivated by the need to properly evaluate and benchmark methods for finding selective sweeps, in **Chapter 2** I evaluated the performance of eight selection scans to detect selective sweeps in 25 breeds of dogs. The domestic dog was an extremely useful system for this work, as it provided multiple distinct breeds that have experienced specific selective pressures through artificial selection. Thanks to this, we were able to select a set of 12 positive control

loci known to have experienced positive selection in specific dog breeds due to their association with desirable morphological phenotypes. These positive control loci were then used to assess the performance of popular statistics to detect selection.

This work successfully detected signature patterns of haplotype and nucleotide polymorphism left by artificial selection during dog domestication and demonstrated the power and limitations of different selection scans as well as the importance of choices in parameters used in the analysis. Cross-population comparison made by hapFLK in all 25 dog breeds was the most sensitive, as it identified all 12 of our control loci. This demonstrates the added power of comparing data across breeds. However, other systems that do not have such cross-population information can only rely on scans that detect signatures of selection from single population samples. Our work showed that single population scans varied widely in their ability to detect signatures of selection in our control loci, not only due to the nature of the statistic and the parameters used, but also the nature of the dog population data itself. For example, haplotype-based statistics that required a segregating ancestral allele in the population, such as iHS and nSL, had particularly low power at locus/breed combinations where the causal allele was fixed in our sample. The best performing single population statistic was H12 followed by π and Tajima's D, all of which were calculated over short windows of 25 segregating sites. The window length for these scans was defined based on the density of SNPs in the genotyping chip. Thus, differences between local densities of SNPs in the chip and the number of SNPs that

are polymorphic in a particular breed in a given window may have driven the signals of positive selection identified by these three statistics.

As mentioned above, purebred dog populations provide an excellent system for mapping the genetic basis of positively selected variants. However, the population history of dog domestication and inbreeding (including severe bottlenecks and high levels of Linkage Disequilibrium), as well as the data collection methods (genotyping chip instead of direct sequencing), could confound the results in this study. Additional work is needed to evaluate the performance of selection scans to detect selective sweeps in large natural populations.

As mentioned earlier, the choice of parameters can also affect the performance of selection scans. Often, parameter choice is either arbitrary or empirical, and thus there is a need in the field to develop methods for deciding which parameters are most appropriate. Some possible approaches might be to select parameters based on real and simulated data and using Machine learning, while taking into account population evolutionary history.

Selection scan methods that identify signatures of selective sweeps are still being developed and improved, however one of the main limitations is that most do not actually identify the specific mutation favored by selection. Regions identified as positively selected by selection scans can be very large (up to a few megabases). Current options to try and narrow down the regions under selection include composite of multiple signals from overlapping regions identified by different

statistics (GROSSMAN *et al.* 2010). Genes present within these regions can be tested for functional and pathway enrichment (such as Gene Ontology) and overlap with known QTLs, as done in (SADEGHI *et al.* 2018). Although a more rigorous way of doing this would be to compare results against neutral loci (WEIGAND AND LEESE 2018). Another option is to rank SNPs on the basis of their functional annotations (AKBARI *et al.* 2018). Techniques to identify a smaller region of interest in a larger stretch of the genome are useful, but methods to find causal SNPs need to be developed. For example, the recently developed iSAFE algorithm ranks all mutations within regions under selection based on their contribution to the selection signal (AKBARI *et al.* 2018). This is a promising step in the direction of finding causal SNPs before any molecular experiments are conducted.

Profiling transcription dynamics using RNA sequencing time series

The state-of-the-art of published RNA-sequencing time course experiments rarely provides the resolution necessary to thoroughly characterize expression dynamics, let alone understand the weaknesses of classical time course analysis in this biological domain. An open-source, thoroughly analyzed dataset of this type, can incentivize innovation in computational methods beyond the scope of its individual study, and lay the groundwork for best practices for future analyses. With this in mind, in **Chapter 3** I presented the transcriptome dynamics profiling of the *Drosophila melanogaster* innate immune response to commercial LPS immune challenge. This

resulting dataset of RNA-seq sampling over 20 time point in 5 days represent the most dense and high-quality time course of the immune response in *Drosophila* to date. Given the resolution and high-dimensionality of this dataset, a broad range of statistical methods were required to extract signals such as clustered expression patterns, Granger Causality relationships between genes, and overall function-specific expression dynamics.

Clustering and classification of time-dependent genes based on their temporal expression profiles unveiled distinct responses to the immune challenge with divergent initiation and resolution dynamics. Clusters of metabolic, immune-induced, and stress-induced genes presented transient responses to immune challenge, resolving as early as 12 hours post-injection. In sharp contrast, clusters of antimicrobial peptides (AMPs) presented a strong sustained response to immune challenge, as they remained up-regulated during the entire 5-day time course. Notable among the detected temporal dynamics, well-characterized circadian rhythm patterns could be observed oscillating in a 24-hour cycle. The identification of these canonical rhythm patterns both validates our methods of data normalization and differential expression analysis, and increases the certainty that we are accurately profiling novel temporal dynamics. Additionally, synchronized expression patterns between genes generally corresponded to their membership in distinct functional categories, allowing us to infer the functional class for previously uncharacterized genes. This method of function-by-association shows promising potential in the large-scale assignment of

functional annotation to other uncharacterized genes through RNA expression time-course experiments.

The surprising quality of the correlation between gene expression dynamics and their functional pathway annotation strongly motivated a function-first approach to immune response expression profiling. We identified differentially expressed biological pathways using pairwise gene set analysis, which allowed us to characterize functional pathway transcription dynamics. The resulting global behaviors of up-regulation of immune response pathways and down-regulation of metabolic pathways point to potential resource tradeoffs precipitated by the immune challenge. I also show that the implication of functional interplay between immune response and metabolic pathways is further supported by the construction of directed Granger-causal networks of putative interactions. Main subnetwork components showed significant GC directional edges between down-regulated metabolic genes and up-regulated genes with cell proliferation and repair functions (such as *Claspin*, *LpR2*, and *Orc1*). These results further suggest an underlying interplay between metabolic pathways and proliferation and repair mechanisms such as regulation of DNA replication stress, endocytosis, and clot formation.

Overall, this analysis motivates innovation in computational methods for longitudinal omics data, both to account for their inherent high-dimensionality and the complex underlying architecture that contains both causal and spurious

coordination. Further, this should serve as a proof of concept for the future of high-density time-course RNA-seq in other model organisms.

RNA-seq time course datasets have a limitations on the resolution of transcription timing as they only quantify the abundance of transcripts at given, discrete, points in time. Alternative methods like precision nuclear run-on sequencing (PRO-seq) can measure active transcription by mapping the distribution of all transcriptionally engaged polymerases in the genome and their nascent RNAs (KWAK *et al.* 2013). This assay allows the identification of instantaneous transcriptional dynamics without confounding secondary effects and pre-existing transcript levels in cells.

Unfortunately, current protocols of PRO-seq using whole fly require ~1000 flies per sample in order to get enough nuclei, a significantly higher material requirement than RNA-seq, which can use 10 or fewer flies per sample. This starting material requirement can prove prohibitive in the case of experiments where individual flies need to be manually injected one by one, as was the case in **Chapter 3**. However, this challenge can be alleviated by using tissue culture cells or when studying perturbations that are easier to induce, such as heat shock. Additionally, an alternative strategy recently developed is the use of chromatin run-on sequencing (ChRO-seq), which is an application of PRO-seq that uses chromatin as starting material instead of nuclei (CHU *et al.* 2018). This reduces the number of flies needed per sample to ~400 flies and significantly reduces preparation complexity and time.

Regardless of the data gathering resolution, transcriptomic time-course datasets are inherently high-dimensional, and thus very challenging to analyze. Non-interacting genes can be incorrectly detected to be Granger Causal when they follow the same global expression pattern, generating an overabundance of extremely significant hits that make interpretation of the analysis as a whole very challenging. In **Chapter 3**, I dealt with this challenge by excluding positively-causal edges in the GC network to focus on divergent behaviors, which constituted clearer and rarer signals in our study. Methods like cointegration analysis better contextualize the similarity of the patterns in Granger Causality by detecting latent factors that would explain the relationship between two transcription patterns better than a simple causal relationship. However, methods of cointegration for biological time courses still need to be developed.

An additional way to improve the signal-to-noise ratio in these types of analyses would be the implementation of dimensionality-reduction techniques based on prior biological knowledge. The gene set that is used for Granger Causality analysis can be filtered using pathway annotation, protein-protein interactions, or known causal relationships like transcription factor promoter/target roles. This type of filtering would greatly increase the biological signal detectable through this type of analysis while greatly reducing the computational burden that such a high-dimensional analysis generates.

Finally, it is important to consider that gene expression is a single modality of the larger biological system under study. Incorporating information from other omics

data gathering tools, like parallel proteomic and metabolomic time series could paint a fuller picture of the system at large. Transcriptional analysis is valuable in its own right, but we require a variety of biological signal to explore the system-level dynamics of post-transcriptional modification, translational efficiency and hormone and nutrient dynamics. Admittedly, such datasets will require the continuous innovation of all the approaches used in **Chapter 3** to accommodate the rapidly escalating level of dimensionality. Alternative data gathering methods can also alleviate the computational challenges described above. Tissue-specific and single-cell RNA-seq both can improve the accuracy of causality analysis by coupling gene pairs in time while also providing a spatial resolution to different transcriptional dynamics.

Studying the influence of host genetics on gut microbiome composition

Complex arrays of microbial communities thrive within their hosts, affecting fitness, disease, and their interaction with their environment through the modulation of various biological processes. Characterizing the modulation of a microbiome by their host and vice versa is an increasingly popular pursuit, as new sequencing technologies and innovation in computational methods have made possible the simultaneous study of microbes and hosts. There is still, however, no well-understood model of the interactions between an organism's genetic code and the types of microbial communities that it can host. This motivated the work in **Chapter 4**, where I presented the gut microbiota profiling of 247 Diversity Outbred mice using 16S

rRNA gene sequencing, and showed the influence of host genetics on gut microbiome composition by performing a high-resolution QTL mapping in the Diversity Outbred mouse panel using microbiome abundances as a response variable.

These functional associations isolate potential targets for microbiome drivers in host genes which may be involved in pathological disease phenotypes. However, experimental validation of these interactions are vital to confirm the associations between disease pathways and specific taxa within the gut microbiome composition. This is particularly challenging to do given the source of these associations come from Diversity Outbred mice, where each genome is a unique mosaic of the 8 strains from the CC population, and thus irreproducible. A knock-out experiment using CC as a form of validation would be a reasonable next step towards systematically identifying these host-genetic microbiome drivers.

This type of association is statistically challenging as it requires a large number of mice to have adequate power given the potential for variation in both DO mice and any given microbiome. Additionally, the limited number of existing studies discussing heritabilities and QTL mappings of bacteria within the gut microbiome makes it hard to compare results in our experiment with others. The growing interest in the study of the microbiome makes these limitations largely temporary, as the number and quality of these types of studies is bound to explode in the coming years.

Although 16S rRNA gene sequencing is currently the most common method for surveying bacterial communities, whole metagenome sequencing will surpass it in

the near future, as it provides richer data at an increasingly low price. Repeating this type of study using whole metagenome sequencing would allow us to expand our analysis beyond microbial taxa abundances and their associated functional pathways and instead zero-in directly on the functional units responsible for the underlying relationship between host and microbe. Likewise, the power of microbiome studies could be greatly enhanced if they integrate other levels of data, such as transcriptomics, proteomics, and metabolomics.

REFERENCES

Adewoye, A. B., C. P. Kyriacou and E. Tauber, 2015 Identification and functional analysis of early gene expression induced by circadian light-resetting in *Drosophila*. BMC Genomics 16**:** 570-570.

Adlercreutz, H., M. O. Pulkkinen, E. K. Hämäläinen and J. T. Korpela, 1984 Studies on the role of intestinal bacteria in metabolism of synthetic and natural steroid hormones. Journal of Steroid Biochemistry 20**:** 217-229.

Aggarwal, K., and N. Silverman, 2008 Positive and negative regulation of the *Drosophila* immune response. BMB Reports 41**:** 267-277.

Ahn, J., R. Sinha, Z. Pei, C. Dominianni, J. Wu *et al.*, 2013 Human gut microbiome and risk for colorectal cancer. Journal of the National Cancer Institute 105**:** 1907-1911.

Akbari, A., J. J. Vitti, A. Iranmehr, M. Bakhtiari, P. C. Sabeti *et al.*, 2018 Identifying the favored mutation in a positive selective sweep. Nature Methods 15**:** 279-282.

Akey, J. M., A. L. Ruhe, D. T. Akey, A. K. Wong, C. F. Connelly *et al.*, 2010 Tracking footprints of artificial selection in the dog genome. Proceedings of the National Academy of Sciences 107**:** 1160.

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts *et al.*, 2002 Innate Immunity in *Molecular Biology of the Cell*. Garland Science, New York.

Andrews, S., 2010 FastQC: a quality control tool for high throughput sequence data., pp.

Ao, J., E. Ling and X.-Q. Yu, 2007 *Drosophila* C-type lectins enhance cellular encapsulation. Molecular Immunology 44**:** 2541-2548.

Auton, A., Y. Rui Li, J. Kidd, K. Oliveira, J. Nadel *et al.*, 2013 Genetic recombination is targeted towards gene promoter regions in dogs. PLOS Genetics 9**:** e1003984-e1003984.

Axelsson, E., A. Ratnakumar, M.-L. Arendt, K. Maqbool, M. T. Webster *et al.*, 2013 The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495**:** 360.

Bajaj, J. S., D. M. Heuman, P. B. Hylemon, A. J. Sanyal, M. B. White *et al.*, 2014 Altered profile of human gut microbiome is associated with cirrhosis and its complications. Journal of Hepatology 60**:** 940-947.

Bajaj, J. S., J. M. Ridlon, P. B. Hylemon, L. R. Thacker, D. M. Heuman *et al.*, 2012 Linkage of gut microbiome with cognition in hepatic encephalopathy. American Journal of Physiology. Gastrointestinal and Liver Physiology 302**:** G168-G175.

Bank, C., G. B. Ewing, A. Ferrer-Admettla, M. Foll and J. D. Jensen, 2014 Thinking too positive? Revisiting current methods of population genetic selection inference. Trends in Genetics 30**:** 540-546.

Bar-Joseph, Z., A. Gitter and I. Simon, 2012 Studying and modelling dynamic biological processes using time-series gene expression data. Nature Reviews Genetics 13**:** 552.

Barrett, R. D. H., and D. Schluter, 2008 Adaptation from standing genetic variation. Trends in Ecology & Evolution 23**:** 38-44.

Barton, N. H., 2000 Genetic hitchhiking. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 355**:** 1553-1562.

Basu, S., S. Das, G. Michailidis and A. K. Purnanandam, 2017 A system-wide approach to measure connectivity in the financial sector. Social Science Research Network.

Basu, S., A. Shojaie and G. Michailidis, 2015 Network Granger Causality with inherent grouping structure. Journal of Machine Learning Research 16**:** 417-453.

Battaglioli, E. J., and P. C. Kashyap, 2018 Chapter 33 - Diet Effects on Gut Microbiome Composition, Function, and Host Physiology, pp. 755-766 in *Physiology of the Gastrointestinal Tract (Sixth Edition)*, edited by H. M. Said. Academic Press.

Batut, P. J., and T. R. Gingeras, 2017 Conserved noncoding transcription and core promoter regulatory code in early *Drosophila* development. eLife 6**:** e29005.

Belheouane, M., Y. Gupta, S. Künzel, S. Ibrahim and J. F. Baines, 2017 Improved detection of gene-microbe interactions in the mouse skin microbiota using high-resolution QTL mapping of 16S rRNA transcripts. Microbiome 5**:** 59.

Bendjilali, N., S. MacLeon, G. Kalra, S. D. Willis, A. K. M. N. Hossian *et al.*, 2017 Time-course analysis of gene expression during the *Saccharomyces cerevisiae* hypoxic response. G3: Genes|Genomes|Genetics 7**:** 221-231.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57**:** 289-300.

Benson, A. K., S. A. Kelly, R. Legge, F. Ma, S. J. Low *et al.*, 2010 Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proceedings of the National Academy of Sciences 107**:** 18933-18938.

Billio, M., M. Getmansky, A. W. Lo and L. Pelizzon, 2012 Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics 104**:** 535-559.

Binder, R. J., 2014 Functions of Heat Shock Proteins in pathways of the innate and adaptive immune system. Journal of Immunology 193**:** 5765-5771.

Blekhman, R., J. K. Goodrich, K. Huang, Q. Sun, R. Bukowski *et al.*, 2015 Host genetic variation impacts microbiome composition across human body sites. Genome Biology 16**:** 191.

Boiko, A. D., S. Porteous, O. V. Razorenova, V. I. Krivokrysenko, B. R. Williams *et al.*, 2006 A systematic search for downstream mediators of tumor suppressor function of p53 reveals a major role of BTG2 in suppression of Ras-induced transformation. Genes & Development 20**:** 236-252.

Bonder, M. J., A. Kurilshikov, E. F. Tigchelaar, Z. Mujagic, F. Imhann *et al.*, 2016 The effect of host genetics on the gut microbiome. Nature Genetics 48**:** 1407.

Boutros, M., H. Agaisse and N. Perrimon, 2002 Sequential activation of signaling pathways during innate immune responses in *Drosophila*. Developmental Cell 3**:** 711-722.

Boyko, A. R., P. Quignon, L. Li, J. J. Schoenebeck, J. D. Degenhardt *et al.*, 2010 A simple genetic architecture underlies morphological variation in dogs. PLOS Biology 8**:** e1000451.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140**:** 783.

Brown, E. M., X. Ke, D. Hitchcock, S. Jeanfavre, J. Avila-Pacheco *et al.*, 2019 *Bacteroides*-derived sphingolipids are critical for maintaining intestinal homeostasis and symbiosis. Cell Host & Microbe 25**:** 668-680.e667.

Bubier, J., V. Philip, C. Quince, J. Campbell, Y. Zhou *et al.*, 2018 Systems genetic discovery of host-microbiome interactions reveals mechanisms of microbial involvement in disease. bioRxiv.

Buchon, N., M. Poidevin, H.-M. Kwon, A. Guillou, V. Sottas *et al.*, 2009 A single modular serine protease integrates signals from pattern-recognition receptors upstream of the *Drosophila* Toll pathway. Proceedings of the National Academy of Sciences 106**:** 12442.

Bulet, P., J. L. Dimarcq, C. Hetru, M. Lagueux, M. Charlet *et al.*, 1993 A novel inducible antibacterial peptide of *Drosophila* carries an O-glycosylated substitution. Journal of Biological Chemistry 268**:** 14893-14897.

Cadieu, E., M. W. Neff, P. Quignon, K. Walsh, K. Chase *et al.*, 2009 Coat variation in the domestic dog is governed by variants in three genes. Science 326**:** 150.

Cai, Z., N. J. Camp, L. Cannon-Albright and A. Thomas, 2011 Identification of regions of positive selection using Shared Genomic Segment analysis. European Journal of Human Genetics 19**:** 667-671.

Camarinha-Silva, A., M. Maushammer, R. Wellmann, M. Vital, S. Preuss *et al.*, 2017 Host genome influence on gut microbial composition and microbial prediction of complex traits in pigs. Genetics 206**:** 1637-1644.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman *et al.*, 2010 QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7**:** 335-336.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone *et al.*, 2011 Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings

of the National Academy of Sciences of the United States of America 108 Suppl 1**:** 4516-4522.

Carpenter, S., and K. A. Fitzgerald, 2015 Transcription of inflammatory genes: Long Noncoding RNA and beyond. Journal of Interferon & Cytokine Research 35**:** 79-88.

Chambers, M. C., E. Jacobson, S. Khalil and B. P. Lazzaro, 2014 Thorax injury lowers resistance to infection in *Drosophila melanogaster*. Infection and Immunity 82**:** 4380-4389.

Chambers, M. C., K. H. Song and D. S. Schneider, 2012 *Listeria monocytogenes* infection causes metabolic shifts in *Drosophila melanogaster*. PLOS One 7**:** e50679.

Chen, G., N. Han, G. Li, X. Li, G. Li *et al.*, 2016a Time course analysis based on gene expression profile and identification of target molecules for colorectal cancer. Cancer Cell International 16**:** 22.

Chen, H., N. Patterson and D. Reich, 2010 Population differentiation as a test for selective sweeps. Genome Research 20**:** 393-402.

Chen, W., F. Liu, Z. Ling, X. Tong and C. Xiang, 2012 Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. PLOS One 7**:** e39743.

Chen, X., R. Rahman, F. Guo and M. Rosbash, 2016b Genome-wide identification of neuronal activity-regulated genes in *Drosophila*. eLife 5**:** e19942.

Chi, W., D. Dao, T. C. Lau, B. D. Henriksbo, J. F. Cavallari *et al.*, 2014 Bacterial peptidoglycan stimulates adipocyte lipolysis via NOD1. PLOS One 9**:** e97675-e97675.

Cho, W. K., S. Lian, S.-M. Kim, B. Y. Seo, J. K. Jung *et al.*, 2015 Time-course RNA-Seq analysis reveals transcriptional changes in rice plants triggered by rice stripe virus infection. PLOS One 10**:** e0136736-e0136736.

Chu, T., E. J. Rice, G. T. Booth, H. H. Salamanca, Z. Wang *et al.*, 2018 Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. Nature Genetics 50**:** 1553-1564.

Churchill, G. A., D. M. Gatti, S. C. Munger and K. L. Svenson, 2012 The Diversity Outbred mouse population. Mammalian Genome 23**:** 713-718.

Clapham, M. E., 2011 Ordination methods and the evaluation of Ediacaran communities, pp. 3-21 in *Quantifying the Evolution of Early Life: Numerical Approaches to the Evaluation of Fossils and Ancient Ecosystems*, edited by M. Laflamme, J. D. Schiffbauer and S. Q. Dornbos. Springer Netherlands, Dordrecht.

Claus, S. P., T. M. Tsang, Y. Wang, O. Cloarec, E. Skordi *et al.*, 2008 Systemic multicompartmental effects of the gut microbiome on mouse metabolic phenotypes. Molecular Systems Biology 4**:** 219-219.

Clemmons, A. W., S. A. Lindsay and S. A. Wasserman, 2015 An effector peptide family required for *Drosophila* toll-mediated immunity. PLOS Pathogens 11**:** e1004876.

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190**:** 389-401.

Collins, B., E. O. Mazzoni, R. Stanewsky and J. Blau, 2006 *Drosophila* CRYPTOCHROME is a circadian transcriptional repressor. Current Biology 16**:** 441-449.

Croze, M., D. Živković, W. Stephan and S. Hutter, 2016 Balancing selection on immunity genes: review of the current literature and new analysis in *Drosophila melanogaster*. Zoology 119**:** 322-329.

Cutter, A. D., and B. A. Payseur, 2013 Genomic signatures of selection at linked sites: unifying the disparity among species. Nature Reviews Genetics 14**:** 262-274.

Cyran, S. A., A. M. Buchsbaum, K. L. Reddy, M.-C. Lin, N. R. J. Glossop *et al.*, 2003 *vrille*, *Pdp1*, and *dClock* form a second feedback loop in the *Drosophila* circadian clock. Cell 112**:** 329-341.

Damulewicz, M., M. Świątek, A. Łoboda, J. Dulak, B. Bilska *et al.*, 2018 Daily regulation of phototransduction, circadian clock, DNA repair, and Immune gene expression by Heme Oxygenase in the retina of *Drosophila*. Genes 10.

Dash, N. R., G. Khoder, A. M. Nada and M. T. Al Bataineh, 2019 Exploring the impact of *Helicobacter pylori* on gut microbiome composition. PLOS One 14**:** e0218274-e0218274.

Davenport, E. R., D. A. Cusanovich, K. Michelini, L. B. Barreiro, C. Ober *et al.*, 2015 Genome-wide association studies of the human gut microbiota. PLOS One 10**:** e0140301.

De Gregorio, E., P. T. Spellman, G. M. Rubin and B. Lemaitre, 2001 Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. Proceedings of the National Academy of Sciences of the United States of America 98**:** 12590-12595.

Delaneau, O., B. Howie, Anthony J. Cox, J.-F. Zagury and J. Marchini, 2013 Haplotype estimation using sequencing reads. The American Journal of Human Genetics 93**:** 687-696.

Deng, Q., D. Ramsköld, B. Reinius and R. Sandberg, 2014 Single-cell RNA-Seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343**:** 193-196.

Dezeure, R., P. Buhlmann, L. Meier and N. Meinshausen, 2015 High-dimensional inference: confidence intervals, *p*-values and R-software hdi. Statistical Science 30**:** 533-558.

DiAngelo, J. R., M. L. Bland, S. Bambina, S. Cherry and M. J. Birnbaum, 2009 The immune response attenuates growth and nutrient storage in *Drosophila* by reducing insulin signaling. Proceedings of the National Academy of Sciences of the United States of America 106**:** 20853-20858.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2012 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29**:** 15-21.

Dudek-Wicher, R. K., A. Junka and M. Bartoszewicz, 2018 The influence of antibiotics and dietary components on gut microbiota. Przeglad Gastroenterologiczny 13**:** 85-92.

Early, A. M., J. R. Arguello, M. Cardoso-Moreira, S. Gottipati, J. K. Grenier *et al.*, 2017a Survey of global genetic diversity within the *Drosophila* immune system. Genetics 205**:** 353.

Early, A. M., N. Shanmugarajah, N. Buchon and A. G. Clark, 2017b *Drosophila* genotype influences commensal bacterial levels. PLOS One 12**:** e0170332.

Efron, B., and R. Tibshirani, 2007 On testing the significance of sets of genes. Annals of Applied Statistics 1**:** 107-129.

Ekengren, S., Y. Tryselius, M. S. Dushay, G. Liu, H. Steiner *et al.*, 2001 A humoral stress response in *Drosophila*. Current Biology 11**:** 714-718.

Ericsson, A. C., S. Akter, M. M. Hanson, S. B. Busi, T. W. Parker *et al.*, 2015 Differential susceptibility to colorectal cancer due to naturally occurring gut microbiota. Oncotarget 6**:** 33689-33704.

Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal and B. Servin, 2013 Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 193**:** 929.

Fava, F., J. Lovegrove, R. Gitau, K. Jackson and K. Tuohy, 2006 The gut microbiota and lipid metabolism: implications for human health and coronary heart disease. Current Medicinal Chemistry 13**:** 3005-3021.

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155**:** 1405.

Fedorka, K. M., J. E. Linder, W. Winterhalter and D. Promislow, 2007 Post-mating disparity between potential and realized immune response in *Drosophila melanogaster*. Proceedings of the Royal Society B: Biological Sciences 274**:** 1211-1217.

Fernández, M. F., I. Reina-Pérez, J. M. Astorga, A. Rodríguez-Carrillo, J. Plaza-Díaz *et al.*, 2018 Breast cancer and its relationship with the microbiota. International Journal of Environmental Research and Public Health 15**:** 1747.

Ferrer-Admetlla, A., M. Liang, T. Korneliussen and R. Nielsen, 2014 On detecting incomplete soft or hard selective sweeps using haplotype structure. Molecular Biology and Evolution 31**:** 1275-1291.

Fitzpatrick, M., and S. P. Young, 2013 Metabolomics – A novel window into inflammatory disease. Swiss Medical Weekly 143**:** w13743-w13743.

Flatt, T., A. Heyland, F. Rus, E. Porpiglia, C. Sherlock *et al.*, 2008 Hormonal regulation of the humoral innate immune response in *Drosophila melanogaster*. The Journal of Experimental Biology 211**:** 2712-2724.

Flores, R., J. Shi, B. Fuhrman, X. Xu, T. D. Veenstra *et al.*, 2012 Fecal microbial determinants of fecal and systemic estrogens and estrogen metabolites: a cross-sectional study. Journal of Translational Medicine 10**:** 253.

Foll, M., H. Shim and J. D. Jensen, 2015 WFABC: a Wright–Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Molecular Ecology Resources 15**:** 87-98.

Francino, M. P., 2016 Antibiotics and the human gut microbiome: dysbioses and accumulation of resistances. Frontiers in Microbiology 6**:** 1543-1543.

Freedman, A. H., K. E. Lohmueller and R. K. Wayne, 2016 Evolutionary History, Selective Sweeps, and Deleterious Variation in the Dog. Annual Review of Ecology, Evolution, and Systematics 47**:** 73-96.

Fuhrman, B. J., H. S. Feigelson, R. Flores, M. H. Gail, X. Xu *et al.*, 2014 Associations of the fecal microbiome with urinary estrogens and estrogen metabolites in postmenopausal women. The Journal of Clinical Endocrinology and Metabolism 99**:** 4632-4640.

Galeano, P., and D. Peña, 2000 Multivariate analysis in vector time series. Resenhas Do Instituto De Matemática E Estatística Da Universidade De São Paulo 4**:** 383-404.

Garrett, W. S., J. I. Gordon and L. H. Glimcher, 2010 Homeostasis and inflammation in the intestine. Cell 140**:** 859-870.

Garud, N. R., P. W. Messer, E. O. Buzbas and D. A. Petrov, 2015 Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLOS Genetics 11**:** e1005004.

Gatti, D. M., K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar *et al.*, 2014 Quantitative trait locus mapping methods for diversity outbred mice. G3: Genes|Genomes|Genetics 4**:** 1623-1633.

Geijer, C., I. Pirkov, W. Vongsangnak, A. Ericsson, J. Nielsen *et al.*, 2012 Time course gene expression profiling of yeast spore germination reveals a network of transcription factors orchestrating the global response. BMC Genomics 13**:** 554-554.

Ghazalpour, A., I. Cespedes, B. J. Bennett and H. Allayee, 2016 Expanding role of gut microbiota in lipid metabolism. Current Opinion in Lipidology 27**:** 141-147.

Goedert, J. J., G. Jones, X. Hua, X. Xu, G. Yu *et al.*, 2015 Investigation of the association between the fecal microbiota and breast cancer in postmenopausal women: A population-based case-control pilot study. Journal of the National Cancer Institute 107**:** djv147-djv147.

Good, B. H., M. J. McDonald, J. E. Barrick, R. E. Lenski and M. M. Desai, 2017 The dynamics of molecular evolution over 60,000 generations. Nature 551**:** 45.

Goodrich, J. K., E. R. Davenport, M. Beaumont, M. A. Jackson, R. Knight *et al.*, 2016 Genetic determinants of the gut microbiome in UK Twins. Cell Host & Microbe 19**:** 731-743.

Goodrich, J. K., E. R. Davenport, A. G. Clark and R. E. Ley, 2017 The relationship between the human genome and microbiome comes into view. Annual Review of Genetics 51**:** 413-433.

Goodrich, J. K., S. C. Di Rienzi, A. C. Poole, O. Koren, W. A. Walters *et al.*, 2014a Conducting a microbiome study. Cell 158**:** 250-262.

Goodrich, Julia K., Jillian L. Waters, Angela C. Poole, Jessica L. Sutter, O. Koren *et al.*, 2014b Human genetics shape the gut microbiome. Cell 159**:** 789-799.

Graham, A. L., D. M. Shuker, L. C. Pollitt, S. K. J. R. Auld, A. J. Wilson *et al.*, 2011 Fitness consequences of immune responses: strengthening the empirical framework for ecoimmunology. Functional Ecology 25**:** 5-17.

Granger, C. W. J., 1969 Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37**:** 424-438.

Gray, M. M., N. B. Sutter, E. A. Ostrander and R. K. Wayne, 2010 The IGF1 small dog haplotype is derived from Middle Eastern grey wolves. BMC Biology 8**:** 16-16.

Greenblatt, M. S., W. P. Bennett, M. Hollstein and C. C. Harris, 1994 Mutations in the *p53* tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer Research 54**:** 4855.

Grossman, S. R., K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki *et al.*, 2013 Identifying recent adaptations in large-scale genomic data. Cell 152**:** 703-713.

Grossman, S. R., I. Shylakhter, E. K. Karlsson, E. H. Byrne, S. Morales *et al.*, 2010 A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327**:** 883.

Han, L., and M. Abney, 2013 Using identity by descent estimation with dense genotype data to detect positive selection. European Journal of Human Genetics 21**:** 205-211.

Handu, M., B. Kaduskar, R. Ravindranathan, A. Soory, R. Giri *et al.*, 2015 SUMO-enriched proteome for *Drosophila* innate immune response. G3: Genes|Genomes|Genetics 5**:** 2137-2154.

He, L., J. A. Hamm, A. Reddy, D. Sams, R. A. Peliciari-Garcia *et al.*, 2016 Biotinylation: a novel posttranslational modification linking cell autonomous circadian clocks with metabolism. American Journal of Physiology. Heart and Circulatory Physiology 310**:** H1520-H1532.

Heaver, S. L., E. L. Johnson and R. E. Ley, 2018 Sphingolipids in host–microbial interactions. Current Opinion in Microbiology 43**:** 92-99.

Hedengren, M., K. Borge and D. Hultmark, 2000 Expression and evolution of the *Drosophila Attacin/Diptericin* gene family. Biochemical and Biophysical Research Communications 279**:** 574-581.

Hermisson, J., and P. S. Pennings, 2005 Soft Sweeps. Genetics 169**:** 2335.

Hieken, T. J., J. Chen, T. L. Hoskin, M. Walther-Antonio, S. Johnson *et al.*, 2016 The microbiome of aseptically collected human breast tissue in benign and malignant disease. Scientific Reports 6**:** 30751-30751.

Hoffmann, J. A., and J.-M. Reichhart, 2002 *Drosophila* innate immunity: an evolutionary perspective. Nature Immunology 3**:** 121.

Holsinger, K. E., and B. S. Weir, 2009 Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nature Reviews Genetics 10**:** 639-650.

Houle, D., G. H. Bolstad, K. van der Linde and T. F. Hansen, 2017 Mutation predicts 40 million years of fly wing evolution. Nature 548**:** 447.

Howick, V. M., and B. P. Lazzaro, 2014 Genotype and diet shape resistance and tolerance across distinct phases of bacterial infection. BMC Evolutionary Biology 14**:** 56.

Huang, D. W., B. T. Sherman and R. A. Lempicki, 2008 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4**:** 44-57.

Huang, Y., J. A. Ainsley, L. G. Reijmers and F. R. Jackson, 2013 Translational profiling of Clock cells reveals circadianly synchronized protein synthesis. PLOS Biology 11**:** e1001703.

Huff, C. D., H. C. Harpending and A. R. Rogers, 2010 Detecting positive selection from genome scans of linkage disequilibrium. BMC Genomics 11**:** 8-8.

Hughes, M. E., J. B. Hogenesch and K. Kornacker, 2010 JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. Journal of Biological Rhythms 25**:** 372-380.

Igartua, C., E. R. Davenport, Y. Gilad, D. L. Nicolae, J. Pinto *et al.*, 2017 Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. Microbiome 5**:** 16.

Im, J. H., 2018 Functional and population genetics of *Drosophila* innate immunity, pp. Cornell University.

Imler, J.-L., S. Tauszig, E. Jouanguy, C. Forestier and J. A. Hoffmann, 2000 LPS-induced immune response in *Drosophila*. Journal of Endotoxin Research 6**:** 459-462.

Innan, H., and Y. Kim, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America 101**:** 10667.

Javanmard, A., and A. Montanari, 2014 Confidence intervals and hypothesis testing for high-dimensional regression. Journal of Machine Learning Research 15**:** 2869-2909.

Johnson, Caroline H., Christine M. Dejea, D. Edler, Linh T. Hoang, Antonio F. Santidrian *et al.*, 2015 Metabolism links bacterial biofilms and colon carcinogenesis. Cell Metabolism 21**:** 891-897.

Johnson, E. L., S. L. Heaver, J. L. Waters, B. I. Kim, A. Bretin *et al.*, 2019 Sphingolipid production by gut Bacteroidetes regulates glucose homeostasis. bioRxiv**:** 632877.

Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato and K. Morishima, 2017 KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research 45**:** D353-D361.

Kaneko, T., W. E. Goldman, P. Mellroth, H. Steiner, K. Fukase *et al.*, 2004 Monomeric and polymeric gram-negative peptidoglycan but not purified LPS stimulate the *Drosophila* IMD pathway. Immunity 20**:** 637-649.

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

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics 123**:** 887-899.

Karasov, T., P. W. Messer and D. A. Petrov, 2010 Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. PLOS Genetics 6**:** e1000924.

Karlsson, C., A. M. Korayem, C. Scherfer, O. Loseva, M. S. Dushay *et al.*, 2004 Proteomic analysis of the *Drosophila* larval hemolymph clot. Journal of Biological Chemistry 279**:** 52033-52041.

Katzenberger, R. J., B. Ganetzky and D. A. Wassarman, 2016 Age and diet affect genetically separable secondary injuries that cause acute mortality following traumatic brain injury in *Drosophila*. G3: Genes|Genomes|Genetics 6**:** 4151-4166.

Keebaugh, E. S., and T. A. Schlenke, 2012 Adaptive evolution of a novel *Drosophila* lectin induced by parasitic wasp attack. Molecular Biology and Evolution 29**:** 565-577.

Kenmoku, H., A. Hori, T. Kuraishi and S. Kurata, 2017 A novel mode of induction of the humoral innate immune response in *Drosophila* larvae. Disease Models & Mechanisms 10**:** 271.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160**:** 765.

Konopka, R. J., and S. Benzer, 1971 Clock mutants of *Drosophila melanogaster*. Proceedings of the National Academy of Sciences of the United States of America 68**:** 2112-2116.

Krautz, R., B. Arefin and U. Theopold, 2014 Damage signals in the insect immune response. Frontiers in Plant Science 5**:** 342-342.

Kryazhimskiy, S., D. P. Rice, E. R. Jerison and M. M. Desai, 2014 Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science 344**:** 1519.

Kwak, H., N. J. Fuda, L. J. Core and J. T. Lis, 2013 Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339**:** 950-953.

Larson, G., E. K. Karlsson, A. Perri, M. T. Webster, S. Y. W. Ho *et al.*, 2012 Rethinking dog domestication by integrating genetics, archeology, and biogeography. Proceedings of the National Academy of Sciences 109**:** 8878.

Law, C. W., Y. Chen, W. Shi and G. K. J. G. B. Smyth, 2014 voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15**:** R29.

Lawrence, M., W. Huber, H. Pagès, P. Aboyoun, M. Carlson *et al.*, 2013 Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology 9**:** e1003118.

Lazzaro, B. P., and Madeline R. Galac, 2006 Disease pathology: wasting energy fighting infection. Current Biology 16**:** R964-R965.

Leamy, L. J., S. A. Kelly, J. Nietfeldt, R. M. Legge, F. Ma *et al.*, 2014 Host genetics and diet, but not immunoglobulin A expression, converge to shape compositional features of the gut microbiome in an advanced intercross population of mice. Genome Biology 15**:** 552-552.

Lee, E.-M., T. T. B. Trinh, H. J. Shim, S.-Y. Park, T. T. T. Nguyen *et al.*, 2012 *Drosophila* Claspin is required for the G2 arrest that is induced by DNA replication stress but not by DNA double-strand breaks. DNA Repair 11**:** 741-752.

Lemaitre, B., and J. Hoffmann, 2007 The host defense of *Drosophila melanogaster*. Annual Review of Immunology 25**:** 697-743.

Leroy, G., T. Mary-Huard, E. Verrier, S. Danvy, E. Charvolin *et al.*, 2013 Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. Genetics Selection Evolution 45**:** 1.

Leulier, F., C. Parquet, S. Pili-Floury, J.-H. Ryu, M. Caroff *et al.*, 2003 The *Drosophila* immune system detects bacteria through specific peptidoglycan recognition. Nature Immunology 4**:** 478.

Levashina, E. A., S. Ohresser, P. Bulet, J.-M. Reichhart, C. Hetru *et al.*, 1995 Metchnikowin, a novel immune-inducible proline-rich peptide from *Drosophila* with antibacterial and antifungal properties. European Journal of Biochemistry 233**:** 694-700.

Levine, A. J., J. Momand and C. A. Finlay, 1991 The p53 tumour suppressor gene. Nature 351**:** 453-456.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74**:** 175.

Ley, R. E., F. Bäckhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight *et al.*, 2005 Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America 102**:** 11070-11075.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25**:** 2078-2079.

Li, M., B. Wang, M. Zhang, M. Rantalainen, S. Wang *et al.*, 2008 Symbiotic gut microbes modulate human metabolic phenotypes. Proceedings of the National Academy of Sciences 105**:** 2117.

Lim, M. Y., H. J. You, H. S. Yoon, B. Kwon, J. Y. Lee *et al.*, 2017 The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. Gut 66**:** 1031.

Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe *et al.*, 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438**:** 803-819.

Liu, D., Z. Shaukat, R. B. Saint and S. L. Gregory, 2015 Chromosomal instability triggers cell death via local signalling through the innate immune receptor Toll. Oncotarget 6**:** 38552-38565.

Long, Y., X. Wang, D. T. Youmans and T. R. Cech, 2017 How do lncRNAs regulate transcription? Science Advances 3.

Lotterhos, K. E., and M. C. Whitlock, 2015 The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Molecular Ecology 24**:** 1031-1046.

Love, M. I., W. Huber and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15**:** 550.

Lü, J., C. Yang, Y. Zhang and H. Pan, 2018 Selection of reference genes for the normalization of RT-qPCR data in gene expression studies in insects: A systematic review. Frontiers in Physiology 9**:** 1560-1560.

Mak, T. W., L. Hauck, D. Grothe and F. Billia, 2017 p53 regulates the cardiac transcriptome. Proceedings of the National Academy of Sciences of the United States of America 114**:** 2331-2336.

Malaspinas, A.-S., O. Malaspinas, S. N. Evans and M. Slatkin, 2012 Estimating allele age and selection coefficient from time-serial data. Genetics 192**:** 599.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17**:** 10-12.

Maynard, J., and J. Haigh, 2007 The hitch-hiking effect of a favourable gene. Genetics Research 89**:** 391-403.

McKean, K. A., C. P. Yourth, B. P. Lazzaro and A. G. Clark, 2008 The evolutionary costs of immunological maintenance and deployment. BMC Evolutionary Biology 8**:** 76.

McKnite, A. M., M. E. Perez-Munoz, L. Lu, E. G. Williams, S. Brewer *et al.*, 2012 Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. PLOS One 7**:** e39191.

McMurdie, P. J., and S. Holmes, 2014 Waste not, want not: why rarefying microbiome data is inadmissible. PLOS Computational Biology 10**:** e1003531.

Messer, P. W., and R. A. Neher, 2012 Estimating the strength of selective sweeps from deep population diversity data. Genetics 191**:** 593.

Messer, P. W., and D. A. Petrov, 2013 Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology & Evolution 28**:** 659-669.

Mi, H., A. Muruganujan, D. Ebert, X. Huang and P. D. Thomas, 2018 PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research 47**:** D419-D426.

Milaneschi, Y., W. K. Simmons, E. F. C. van Rossum and B. W. J. H. Penninx, 2019 Depression and obesity: evidence of shared biological mechanisms. Molecular Psychiatry 24**:** 18-33.

Miles, C. M., and M. Wayne, 2008 Quantitative trait locus (QTL) analysis. Nature Education 1**:** 208.

Minelli, E. B., A. M. Beghini, S. Vesentini, L. Marchiori, G. Nardo *et al.*, 1990 Intestinal microflora as an alternative metabolic source of estrogens in women with uterine leiomyoma and breast cancer. Annals of the New York Academy of Sciences 595**:** 473-479.

Miyaki, M., and T. Kuroki, 2003 Role of Smad4 (DPC4) inactivation in human cancer. Biochemical and Biophysical Research Communications 306**:** 799-804.

Miyamoto, T., and H. Amrein, 2017 Gluconeogenesis: An ancient biochemical pathway with a new twist. Fly 11**:** 218-223.

Montero, P., and J. A. Vilar, 2014 TSclust: an R package for time series clustering. Journal of Statistical Software 62**:** 1-43.

Mukhopadhyay, N. D., and S. Chatterjee, 2006 Causality and pathway search in microarray time series experiment. Bioinformatics 23**:** 442-449.

Mullighan, C. G., X. Su, J. Zhang, I. Radtke, L. A. A. Phillips *et al.*, 2009 Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. The New England Journal of Medicine 360**:** 470-480.

Mumtaz, P. T., S. A. Bhat, S. M. Ahmad, M. A. Dar, R. Ahmed *et al.*, 2017 LncRNAs and immunity: watchdogs for host pathogen interactions. Biological Procedures Online 19**:** 3.

Mundorf, J., C. D. Donohoe, C. D. McClure, T. D. Southall and M. Uhlirova, 2019 Ets21c governs tissue renewal, stress tolerance, and aging in the *Drosophila* intestine. Cell Reports 27**:** 3019-3033.e3015.

Myers, M. P., K. Wager-Smith, A. Rothenfluh-Hilfiker and M. W. Young, 1996 Light-Induced Degradation of TIMELESS and Entrainment of the *Drosophila* Circadian Clock. Science 271**:** 1736.

Myllymäki, H., S. Valanne and M. Rämet, 2014 The *Drosophila* Imd signaling pathway. The Journal of Immunology 192**:** 3455.

Nakad, R., and B. Schumacher, 2016 DNA damage response and immune defense: links and mechanisms. Frontiers in Genetics 7**:** 147-147.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Research 15**:** 1566-1575.

Novembre, J., and M. Stephens, 2008 Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40**:** 646-649.

O'Connor, A., P. M. Quizon, J. E. Albright, F. T. Lin and B. J. Bennett, 2014 Responsiveness of cardiometabolic-related microbiota to diet is influenced by host genetics. Mammalian Genome 25**:** 583-599.

Org, E., B. W. Parks, J. W. J. Joo, B. Emert, W. Schwartzman *et al.*, 2015 Genetic and environmental control of host-gut microbiota interactions. Genome Research 25**:** 1558-1569.

Orr, H. A., and A. J. Betancourt, 2001 Haldane's sieve and adaptation from the standing genetic variation. Genetics 157**:** 875.

Parker, H. G., L. V. Kim, N. B. Sutter, S. Carlson, T. D. Lorentzen *et al.*, 2004 Genetic structure of the purebred domestic dog. Science 304**:** 1160.

Paschos, G. K., and G. A. FitzGerald, 2017 Circadian clocks and metabolism: implications for microbiome and aging. Trends in Genetics 33**:** 760-769.

Pavlidis, P., D. Živković, A. Stamatakis and N. Alachiotis, 2013 SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. Molecular Biology and Evolution 30**:** 2224-2234.

Pearce, Laura R., N. Atanassova, Matthew C. Banton, B. Bottomley, Agatha A. van der Klaauw *et al.*, 2013 KSR2 Mutations Are Associated with Obesity, Insulin Resistance, and Impaired Cellular Fuel Oxidation. Cell 155**:** 765-777.

Pegoraro, M., and E. Tauber, 2018 Photoperiod-dependent expression of MicroRNA in *Drosophila*. bioRxiv**:** 464180.

Pennings, P. S., and J. Hermisson, 2006a Soft Sweeps II: molecular population genetics of adaptation from recurrent mutation or migration. Molecular Biology and Evolution 23**:** 1076-1084.

Pennings, P. S., and J. Hermisson, 2006b Soft Sweeps III: the signature of positive selection from recurrent mutation. PLOS Genetics 2**:** e186.

Petitjean, A., M. I. W. Achatz, A. L. Borresen-Dale, P. Hainaut and M. Olivier, 2007 TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. Oncogene 26**:** 2157-2165.

Pietruska, J. R., and A. B. Kane, 2007 SV40 oncoproteins enhance asbestos-induced DNA double-strand breaks and abrogate senescence in murine mesothelial cells. Cancer Research 67**:** 3637.

Ploetz, R. C., 1994 Panama disease: Return of the first banana menace. International Journal of Pest Management 40**:** 326-336.

Plottel, Claudia S., and Martin J. Blaser, 2011 Microbiome and malignancy. Cell Host & Microbe 10**:** 324-335.

Podani, J., and I. Miklós, 2002 Resemblance coefficients and The Horseshoe Effect in Principal Coordinates Analysis. Ecology 83**:** 3331-3343.

Poh, Y.-P., V. S. Domingues, H. E. Hoekstra and J. D. Jensen, 2014 On the prospect of identifying adaptive loci in recently bottlenecked populations. PLOS One 9**:** e110579.

Pritchard, J. K., J. K. Pickrell and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology 20**:** R208-R215.

Przeworski, M., G. Coop and J. D. Wall, 2005 The signature of positive selection on standing genetic variation. Evolution 59**:** 2312-2323.

Reinsch, C. H., 1967 Smoothing by spline functions. Numerische Mathematik 10**:** 177-183.

Ridaura, V. K., J. J. Faith, F. E. Rey, J. Cheng, A. E. Duncan *et al.*, 2013 Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science 341**:** 1241214-1241214.

Rideout, J. R., Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell *et al.*, 2014 Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ 2**:** e545-e545.

Rimbault, M., H. C. Beale, J. J. Schoenebeck, B. C. Hoopes, J. J. Allen *et al.*, 2013 Derived variants at six genes explain nearly half of size reduction in dog breeds. Genome Research 23**:** 1985-1995.

Robinson, M. D., D. J. McCarthy and G. K. Smyth, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26**:** 139-140.

Robinson, M. D., and A. Oshlack, 2010 A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11**:** R25.

Rolff, J., and M. T. Siva-Jothy, 2002 Copulation corrupts immunity: a mechanism for a cost of mating in insects. Proceedings of the National Academy of Sciences of the United States of America 99**:** 9916-9918.

Rothschild, D., O. Weissbrod, E. Barkan, A. Kurilshikov, T. Korem *et al.*, 2018 Environment dominates over host genetics in shaping human gut microbiota. Nature 555**:** 210.

Round, J. L., and S. K. Mazmanian, 2009 The gut microbiota shapes intestinal immune responses during health and disease. Nature Reviews Immunology 9**:** 313-323.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419**:** 832-837.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449**:** 913.

Sackton, T. B., B. P. Lazzaro and A. G. Clark, 2010 Genotype and gene expression associations with immune function in *Drosophila*. PLOS Genetics 6**:** e1000797.

Sadeghi, R., M. Moradi-Shahrbabak, S. R. Miraei Ashtiani, F. Schlamp, E. J. Cosgrove *et al.*, 2018 Genetic diversity of Persian Arabian horses and their relationship to other native Iranian horse breeds. Journal of Heredity 110**:** 173-182.

Sanduzzi Zamparelli, M., A. Rocco, D. Compare and G. Nardone, 2017 The gut microbiota: a new potential driving force in liver cirrhosis and hepatocellular carcinoma. United European Gastroenterology Journal 5**:** 944-953.

Scheipl, F., S. Greven and H. Küchenhoff, 2008 Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Computational Statistics & Data Analysis 52**:** 3283-3299.

Schlamp, F., J. van der Made, R. Stambler, L. Chesebrough, A. R. Boyko *et al.*, 2016 Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. Molecular Ecology 25**:** 342-356.

Schmutz, S. M., and Y. Melekhovets, 2012 Coat color DNA testing in dogs: theory meets practice. Molecular and Cellular Probes 26**:** 238-242.

Schoenebeck, J. J., and E. A. Ostrander, 2013 The genetics of canine skull shape variation. Genetics 193**:** 317.

Schrider, D. R., F. K. Mendes, M. W. Hahn and A. D. Kern, 2015 Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics 200**:** 267.

Schwenke, R. A., and B. P. Lazzaro, 2017 Juvenile hormone suppresses resistance to infection in mated female *Drosophila melanogaster*. Current Biology 27**:** 596-601.

Schwenke, R. A., B. P. Lazzaro and M. F. Wolfner, 2016 Reproduction–immunity trade-offs in insects. Annual Review of Entomology 61**:** 239-256.

Sefer, E., M. Kleyman and Z.-B. Joseph, 2016 Tradeoffs between dense and replicate sampling strategies for high throughput time series experiments. Cell Systems 3**:** 35-42.

Seth, A. K., A. B. Barrett and L. Barnett, 2015 Granger causality analysis in neuroscience and neuroimaging. Journal of Neuroscience 35**:** 3293-3297.

Shannon, L. M., R. H. Boyko, M. Castelhano, E. Corey, J. J. Hayward *et al.*, 2015 Genetic structure in village dogs reveals a Central Asian domestication origin. Proceedings of the National Academy of Sciences 112**:** 13639.

Shapira, M., 2016 Gut microbiotas and host evolution: scaling up symbiosis. Trends in Ecology & Evolution 31**:** 539-549.

Shin, J., S. Lee, M.-J. Go, S. Y. Lee, S. C. Kim *et al.*, 2016 Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. Scientific Reports 6**:** 29681.

Shinoda, T., and K. Itoyama, 2003 Juvenile hormone acid methyltransferase: a key regulatory enzyme for insect metamorphosis. Proceedings of the National Academy of Sciences of the United States of America 100**:** 11986-11991.

Short, S. M., and B. P. Lazzaro, 2010 Female and male genetic contributions to post-mating immune defence in female *Drosophila melanogaster*. Proceedings of the Royal Society B: Biological Sciences 277**:** 3649-3657.

Short, S. M., M. F. Wolfner and B. P. Lazzaro, 2012 Female *Drosophila melanogaster* suffer reduced defense against infection due to seminal fluid components. Journal of Insect Physiology 58**:** 1192-1201.

Snijders, A. M., S. A. Langley, Y.-M. Kim, C. J. Brislawn, C. Noecker *et al.*, 2016 Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome. Nature Microbiology 2**:** 16221.

So, W. V., L. Sarov-Blat, C. K. Kotarski, M. J. McDonald, R. Allada *et al.*, 2000 *takeout*, a novel *Drosophila* gene under circadian clock transcriptional regulation. Molecular and Cellular Biology 20**:** 6935-6944.

Soukup, S. F., J. Culi and D. Gubb, 2009 Uptake of the necrotic serpin in *Drosophila melanogaster* via the Lipophorin Receptor-1. PLOS Genetics 5**:** e1000532.

Spies, D., and C. Ciaudo, 2015 Dynamics in transcriptomics: advancements in RNA-seq time course and downstream analysis. Computational and Structural Biotechnology Journal 13**:** 469-477.

Spies, D., P. F. Renz, T. A. Beyer and C. Ciaudo, 2017 Comparative analysis of differential gene expression tools for RNA sequencing time course data. Briefings in Bioinformatics 20**:** 288-298.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert *et al.*, 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102**:** 15545.

Sutter, N. B., M. A. Eberle, H. G. Parker, B. J. Pullar, E. F. Kirkness *et al.*, 2004 Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. Genome Research 14**:** 2388-2396.

Svenson, K. L., D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng *et al.*, 2012 High-resolution genetic mapping using the Mouse Diversity Outbred population. Genetics 190**:** 437-447.

Szpiech, Z. A., and R. D. Hernandez, 2014 selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Molecular Biology and Evolution 31**:** 2824-2827.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123**:** 585.

Teshima, K. M., G. Coop and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Research 16**:** 702-712.

Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58**:** 267-288.

Troha, K., J. H. Im, J. Revah, B. P. Lazzaro and N. Buchon, 2018 Comparative transcriptomics reveals CrebA as a novel regulator of infection tolerance in *D. melanogaster*. PLOS Pathogens 14**:** e1006847.

Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan *et al.*, 2009 A core gut microbiome in obese and lean twins. Nature 457**:** 480-484.

Turpin, W., O. Espin-Garcia, W. Xu, M. S. Silverberg, D. Kevans *et al.*, 2016 Association of host genome with intestinal microbial composition in a large healthy cohort. Nature Genetics 48**:** 1413-1417.

Ubhi, T., and G. W. Brown, 2019 Exploiting DNA replication stress for cancer treatment. Cancer Research 79**:** 1730-1739.

Ueda, H. R., A. Matsumoto, M. Kawamura, M. Iino, T. Tanimura *et al.*, 2002 Genome-wide transcriptional orchestration of circadian rhythms in *Drosophila*. Journal of Biological Chemistry 277**:** 14048-14052.

Vaysse, A., A. Ratnakumar, T. Derrien, E. Axelsson, G. Rosengren Pielberg *et al.*, 2011 Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLOS Genetics 7**:** e1002316-e1002316.

Veiga, P., C. A. Gallini, C. Beal, M. Michaud, M. L. Delaney *et al.*, 2010 *Bifidobacterium animalis* subsp. *lactis* fermented milk product reduces inflammation by altering a niche for colitogenic microbes. Proceedings of the National Academy of Sciences of the United States of America 107**:** 18132-18137.

Velagapudi, V. R., R. Hezaveh, C. S. Reigstad, P. Gopalacharyulu, L. Yetukuri *et al.*, 2010 The gut microbiota modulates host energy and lipid metabolism in mice. Journal of Lipid Research 51**:** 1101-1112.

Vitti, J. J., S. R. Grossman and P. C. Sabeti, 2013 Detecting natural selection in genomic data. Annual Review of Genetics 47**:** 97-120.

Vogelstein, B., S. Sur and C. Prives, 2010 p53: the most frequently altered gene in human cancers. Nature Education 3**:** 6.

Voight, B. F., S. Kudaravalli, X. Wen and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLOS Biology 4**:** e72.

Wang, J., S. Kalyan, N. Steck, L. M. Turner, B. Harr *et al.*, 2015 Analysis of intestinal microbiota in hybrid house mice reveals evolutionary divergence in a vertebrate hologenome. Nature Communications 6**:** 6440.

Wang, J., L. B. Thingholm, J. Skiecevičienė, P. Rausch, M. Kummen *et al.*, 2016 Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nature Genetics 48**:** 1396-1406.

Wang, L., L. Yang, M. Debidda, D. Witte and Y. Zheng, 2007 Cdc42 GTPase-activating protein deficiency promotes genomic instability and premature aging-like phenotypes. Proceedings of the National Academy of Sciences of the United States of America 104**:** 1248-1253.

Weigand, H., and F. Leese, 2018 Detecting signatures of positive selection in non-model species using genomic data. Zoological Journal of the Linnean Society 184**:** 528-583.

Wen, L., R. E. Ley, P. Y. Volchkov, P. B. Stranges, L. Avanesyan *et al.*, 2008 Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nature 455**:** 1109-1113.

White, R. J., J. E. Collins, I. M. Sealy, N. Wali, C. M. Dooley *et al.*, 2017 A high-resolution mRNA expression time course of embryonic development in zebrafish. eLife 6**:** e30860.

Wilson, B. A., D. A. Petrov and P. W. Messer, 2014 Soft selective sweeps in complex demographic scenarios. Genetics 198**:** 669.

Wolowczuk, I., C. Verwaerde, O. Viltart, A. Delanoye, M. Delacre *et al.*, 2008 Feeding our immune system: Impact on metabolism. Clinical and Developmental Immunology 2008**:** 19.

Yang, B., Y. Wang and P.-Y. Qian, 2016 Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17**:** 135.

Yang, J., Q. Tan, Q. Fu, Y. Zhou, Y. Hu *et al.*, 2017 Gastrointestinal microbiome and breast cancer: correlations, mechanisms and potential clinical implications. Breast Cancer 24**:** 220-228.

Yoshimoto, S., T. M. Loo, K. Atarashi, H. Kanda, S. Sato *et al.*, 2013 Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. Nature 499**:** 97.

Zackular, J. P., M. A. M. Rogers, M. T. Ruffin and P. D. Schloss, 2014 The human gut microbiome as a screening tool for colorectal cancer. Cancer Prevention Research 7**:** 1112.

Zbinden, M., C. R. Haag and D. Ebert, 2008 Experimental evolution of field populations of *Daphnia magna* in response to parasite treatment. Journal of Evolutionary Biology 21**:** 1068-1078.

Zerofsky, M., E. Harel, N. Silverman and M. Tatar, 2005 Aging of the innate immune response in *Drosophila melanogaster*. Aging Cell 4**:** 103-108.

Zhang, Y., R. Papazyan, M. Damle, B. Fang, J. Jager *et al.*, 2017 The hepatic circadian clock fine-tunes the lipogenic response to feeding through RORα/γ. Genes & Development 31**:** 1202-1211.

Zhao, Z., and A. J. Zera, 2004 A morph-specific daily cycle in the rate of JH biosynthesis underlies a morph-specific daily cycle in the hemolymph JH titer in a wing-polymorphic cricket. Journal of Insect Physiology 50**:** 965-973.

Zhu, J., M. Liao, Z. Yao, W. Liang, Q. Li *et al.*, 2018 Breast cancer in postmenopausal women is associated with an altered gut metagenome. Microbiome 6**:** 136.

Ziyatdinov, A., M. Vázquez-Santiago, H. Brunel, A. Martinez-Perez, H. Aschard *et al.*, 2018 lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. BMC Bioinformatics 19**:** 68-68.

Zsámboki, J., G. Csordás, V. Honti, L. Pintér, I. Bajusz *et al.*, 2013 *Drosophila* Nimrod proteins bind bacteria. Central European Journal of Biology 8**:** 633-645.