On the prevalence and role of epistasis in shaping fitness within and between populations

Onuralp Söylemez

Tesi Doctoral UPF
Barcelona, 2015

Thesis Supervisor

Fyodor A. Kondrashov, PhD

Evolutionary Genomics Laboratory
Bioinformatics and Genomics Programme,
Centre for Genomic Regulation (CRG)
Universitat Pompeu Fabra (UPF)

To my parents and sister

Acknowledgements

I would like to thank my thesis supervisor, Fyodor Kondrashov, for giving me the opportunity to pursue my passion at his research group along with many bright colleagues, and for his generous mentorship throughout my doctoral studies.

I am grateful to the thesis committee members, Roderic Guigó, Xavier Estivill and Santiago Elena for their guidance and encouragement.

I would like to acknowledge the financial support by the Spanish Ministry of Science and Education through a pre-doctoral Formación del Profesorado Universitario (FPU) grant AP2008-01888.

I am grateful to Nikki for her unwavering support. I dedicate this thesis to my parents and sister, without their unconditional love, patience and support this work would not have been possible.

Table of contents

Abstract

The role of epistasis – inter-dependent contributions of alleles to fitness – in shaping genetic variation within and between populations is an important question in evolutionary biology with significant implications for our understanding of the factors contributing to phenotypic variation. While epistasis has been shown to play an important role in evolutionary processes such as speciation and adaptive evolution, many aspects of this role remains poorly understood. In particular, there is much debate on whether observing prevalent epistasis in evolution can be taken as evidence for functional epistasis that is relevant to selectable variation. Here, we studied the nature of epistasis in protein evolution, and found a high prevalence of epistatic interactions between amino acid sites in the human genome. We showed that these interactions can help improve accuracy of predicting the impact of genetic variation on the protein structure and function. We also showed that hypothesis-driven search for epistasis in natural populations can detect genomic signatures of epistasis in humans.

Resumen

El papel de la epistasia - contribuciones interdependientes de alelos a la adecuación biológica - en la conformación de la variación genética dentro y entre poblaciones es una cuestión importante en la biología evolutiva con importantes implicaciones para nuestra comprensión de los factores que contribuyen a la variación fenotípica. Mientras la epistasia se ha demostrado que desempeña un papel importante en los procesos evolutivos como la especiación y evolución adaptativa, muchos aspectos de esta función siguen siendo poco conocidos. En particular, hay mucho debate sobre si la observación de epistasia frecuente en la evolución puede ser tomada como evidencia de epistasia funcional que es relevante a la variación heredable. Aquí, se estudió la naturaleza de la epistasia en la evolución de proteínas, y encontramos una alta prevalencia de interacciones epistaticas entre sitios de aminoácidos en el genoma humano. Hemos demostrado que estas interacciones pueden ayudar a mejorar la precisión de predecir el impacto de la variación genética en la estructura y función de las proteínas. También se puso de manifiesto que la búsqueda de investigación basada en hipótesis por epistasia en poblaciones naturales puede detectar firmas genómicas de epistasia en los humanos.

"[C]ontext and interaction are of the essence."
Richard Lewontin (1974)
"This relentless and futile search for intraspecific epistasis needs to be abandoned!"
Brian Charlesworth (1985)

Preface

One of the fundamental goals of the contemporary evolutionary genetics is to bridge the gap between microevolution and macroevolution, that is, to understand the connection between the genetic variation within a species (*genetic basis of evolutionary change*) and the genetic variation between species (*evolutionary basis of genetic change*). As the same adaptive and non-adaptive evolutionary forces underlie the diversity both in the short-term and the long-term evolution, it is conceivable to identify the confounding factors that are not trivial to take into consideration or the conditions under which the disconnection may be informative about the relative importance of those factors. Presumably, the most relevant context to study the nature of genetic variation is to understand how genetic variation relates to fitness, the ultimate appraisal for heredity.

Mapping genetic variation to phenotypic variation is often complicated by *epistasis*, a term used to describe the instances when the phenotypic or fitness effect of an allele depends on the genetic background. To what extent

epistasis, inter-dependent contribution of alleles to fitness, shapes the standing genetic variation or contributes to genetic divergence between species remains an open and controversial issue. To address this issue, we have decided to quantify the prevalence of epistasis on the macro-evolutionary and micro-evolutionary scales, and to discuss the implications for our understanding of the open questions in molecular evolution and evolutionary biology.

This thesis work includes three manuscripts that form the main text and are presented in respective chapters, and two supplementary manuscripts arising from collaborations during the doctoral studies. Chapter 1 provides an introduction to epistasis, presents examples of epistasis from computational and experimental studies, and briefly discusses the open questions about the role of epistasis in molecular evolution and genomic medicine. Chapter 2 refers to the first manuscript in which we estimated the rate of irreversibility of protein evolution by identifying epistatically interacting sites along the mammalian phylogeny. Chapter 3 refers to the second manuscript in which we assessed the concordance between the prevalence of epistasis in *von Willebrand Factor* (vWF) gene estimated based on the expected prevalence of epistatic interactions and the empirical fitness landscape of the same gene revealed empirically by substrate phage display. Chapter 4 refers to the third manuscript in which we

provide a novel method to identify instances of epistatically interacting alleles among natural polymorphisms in the human population, which can in principle be extended to other model organisms with reasonably high level of genetic variation. Chapter 5 includes the concluding remarks arising from the three main manuscripts and discusses briefly the future perspective on the epistasis studies. Supplementary manuscripts are provided in the appendix.

Chapter 1

Introduction

1.1. Epistasis and the concept of a sequence space

One of the fundamental goals of evolutionary biology is to understand the relationship between *genotype* (an organism's hereditary material) and *phenotype* (the observable characteristics of the organism produced by the hereditary material). A good understanding of how genotype relates to phenotype has significant implications for our ability to predict the consequences of genetic changes on quantitative traits and medically-relevant phenotypes.

While this relationship – often conceptualized as *mapping* genotype to phenotype – implies causation, studying the precise nature of this relationship remains to be a challenge due to a variety of intervening factors such as the influence of environment (Burga and Lehner 2013), the intrinsic stochasticity associated with the conception of genotypes (Burga et al. 2011) and the developmental processes that define the fine details of phenotypes. However, it is conceivable to study the influence of these intervening factors by identifying systematic deviations from the null expectation of a direct causal link between genotype and phenotype without any intervening factor.

Clearly, the most intriguing genotype-phenotype map refers to the relationship between the molecular sequence of an organism and the organism's fitness, that is, how a particular genotype (*allele*) at the sequence level contributes to organismal fitness.

To illustrate the structure and nature of the molecular sequence space, the theoretical evolutionary biologist Sewall Wright introduced the concept of fitness landscape describing the map of allele combinations and their contribution to fitness. (Wright, 1932) (**Figure 1**) Wright's visualization of the sequence space illustrates the fundamental features of molecular evolution: (i) evolution proceeds gradually by accumulating many small changes (one allele change or mutation at a time), and (ii) evolution can only take mutational paths involving functional intermediates corresponding to allele combinations with adaptive values.

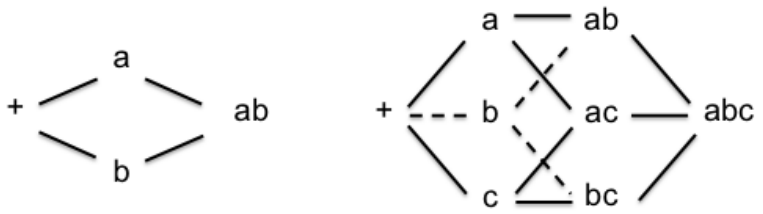

Figure 1. Sewall Wright's illustration of genotype space as described by a network of allele combinations. The cross represents the wild-type, and the contours correspond to different adaptive values for mutants. Adapted from (Wright, 1932)

In 1970, the evolutionary biologist John Maynard Smith introduced a powerful analogy of using a popular word game to demonstrate the same fundamental features of molecular evolution that Wright conveyed visually in his network of allele combinations. (Maynard Smith, 1970) The objective of the word game is to convert one word to another word by following two rules: you can change one letter at a time, and you have to change a letter such that the intermediate words along the way must be meaningful in the English language. (**Figure 2**)

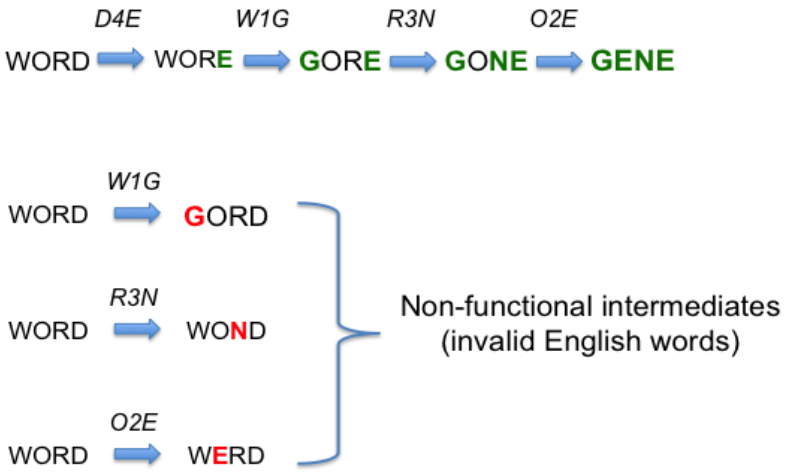

Figure 2. The word game analogy for fundamental features of molecular evolution. Adapted from (Smith, 1970).

Following the original example shown in Figure 2, three of the four mutations – G at position 1, N at position 3, and E at position 2 – required to convert WORD to GENE would not result in valid words in English language if introduced at the first step. That is, these mutations become acceptable only at later stages when there are other mutations. This context dependency or interdependency between individual mutations is called epistasis, which implies interactions between the effects of different mutations on fitness. The concept of a sequence space and mutational paths involving interdependent mutations provides a general framework to study how epistasis influences evolution by reconstructing possible mutational

paths that a putative ancestral sequence could have diverged to give rise to the contemporary sequences.

1.2. Patterns of epistasis in nature

Epistasis refers to instances where the phenotypic or fitness effect of an allele at one locus depends on alleles at other loci. The epistatic or combined effect of interacting alleles can be qualitatively and quantitatively different for a given trait depending on the genetic architecture of the phenotype or fitness for that trait. In general, we can distinguish different types of epistasis based on how the combined effect deviates from the individual effect of the alleles on the phenotype. Given a simple case of allelic combinations of two mutations, there are four different types of epistasis that can be observed when moving from the initial allele combination of *ab* to allele combination of *AB* with a higher fitness value (**Figure 3**).

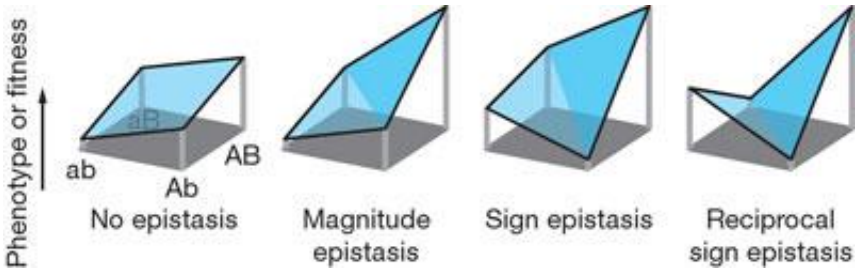

Figure 3. Different types of epistasis in the simple case of only two mutations (Poelwijk et al. 2007)

If the fitness value of a mutation at one locus (*a* to *A*) does not depend on whether there is a *b* or *B* at the other locus, there is no epistasis. Alternatively, if the fitness value on one background is different in magnitude from that on another background, then this type of epistasis is called magnitude epistasis. Clearly, the fitness value of the combined effect may change in sign, that is, the mutation at a given locus is beneficial or deleterious depending on the state of the other locus – this type of epistasis is called sign epistasis. Maynard Smith's word game described in previous section illustrates a good example of sign epistasis where certain letters can only be introduced (beneficial or acceptable) on the background of letters at other positions in the word, otherwise their introduction would result in invalid words (deleterious or unacceptable).

Understanding to what extent the actual sequence space harbors epistasis, and sign epistasis in particular, remains to be an active research area, which has fundamental implications for molecular evolution and clinical comparative medicine. The precise structure or nature of the genotype-phenotype map ("topology of the fitness landscape") determines the ways in which adaptive and non-adaptive evolutionary forces can shape fitness in populations. Clearly, the type of epistasis we observe in nature reflects the relevant context and the scale at which we study epistasis. Therefore, it is conceivable that studying

epistasis at the population level might reveal one type of epistasis while the epistatic interactions at the gene network level in the same population might indicate a different type of epistasis, or no epistasis at all. Thus, it is essential to appreciate the context when searching for epistasis and making inferences about the underlying topology of the fitness landscape.

Accordingly, the literature on epistasis offers a rich diversity of experimental and theoretical studies in their choice of complexity of the organism (low or high), of the molecular level (genes or networks of genes), of the types of epistatic interactions (synergistic or antagonistic), to name the notable preferences. These preferences are usually informed by the study design questions pertinent to the characteristics of the phenotype of interest. In particular, it is desirable that the phenotype can be measured precisely and monitored – for example, through sequencing – accurately across many rounds or generations, and that the phenotype approximates well the fitness. These considerations have prompted two broad approaches to studying epistasis – hypothesis-driven and hypothesis-free methods.

Hypothesis-driven approaches to studying the nature of epistatic interactions typically involve a single gene with a well-characterized structure and function, and the

distribution of epistatic effects are studied among all possible allele combinations of a modest number of sites that are known or suspected to contribute to the phenotype of interest. In principle, these studies can detect epistatic interactions of considerable effect size and provide the best examples of epistasis that can be clearly linked to the gene function. For example, Natarajan and colleagues studied structural and functional variation among segregating amino acid variants that contribute to adaptive functional variation (oxygen affinity) in deer mice hemoglobin (Natarajan et al. 2013). Among all combinatorial permutations of allelic variants that affect oxygen affinity, they found very few combinations that were prevalent in highland and lowland deer mice while the remaining combinations were deleterious. Moreover, the effects of individual mutations in these combinations on adapting to high altitudes were shown to depend on the allelic state of other residue positions, which is indicative of sign epistasis. This study underscores the value of information on structure and the availability of a phenotype that relates to fitness reasonably well.

Another approach to studying epistatic interactions in a hypothesis-driven manner relies on accurately reconstructing the ancestral sequence of contemporary proteins to investigate how the ancestral sequence could have diverged in sequence space to give rise to the protein

with the existing function. The most notable examples highlighting the value of this approach include two studies on the intra-molecular interactions involved in glucocorticoid receptor evolution (Ortlund et al. 2007; Bridgham, Ortlund and Thornton 2009) and a recent study on the inter-molecular interaction between an ancient transcription factor and its DNA regulatory elements (Anderson, McKeown and Thornton 2015). These studies show the importance of historical contingency when evolving a novel function, which involves permissive and restricted mutations that allow and block, respectively, certain evolutionary trajectories. In particular, the studies on glucocorticoid receptor evolution found that the evolutionary trajectory leading to the novel specificity involves an amino acid substitution that is neutral with respect to the specificity, however, is required to accommodate the function-switching substitutions further along the way.

Historical contingency – the importance of temporal order of mutations – can also occur on much shorter time scales such as during the antibiotic resistance. Indeed, studies on resistance to beta-lactam antibiotics (e.g., penicillin) (Weinreich et al. 2006) and to pyrimethamine in the malaria parasite (Lozovsky et al. 2009) show that majority of the possible mutational trajectories are selectively inaccessible due to pervasive intragenic sign epistasis between sites, and therefore, resistance develops

in a step-wise fashion following a limited number of temporal configurations involving successive substitutions. (**Figure 4**)

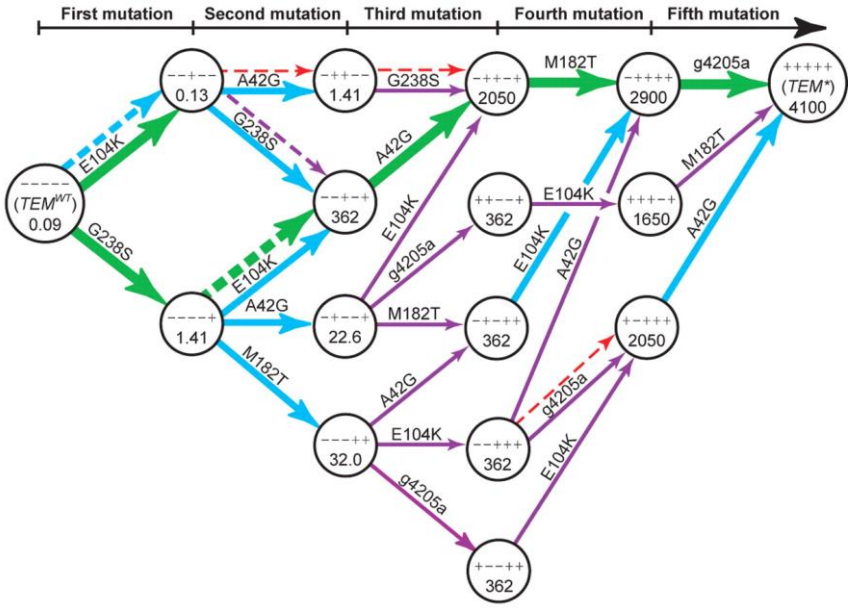

Figure 4. The 10 most probable mutational trajectories among a total of 120 possible trajectories. Width of the edges correspond to probabilities of each beneficial mutation (Weinreich et al. 2006)

Alternatively, hypothesis-free approaches refer to experimental studies on the topology and dynamics of fitness landscapes, and seek an understanding of the prevailing pattern of epistasis and the characteristics of the distribution of epistatic effects. Viruses (e.g., Bonhoeffer et al. 2004; Gong et al. 2013) and bacteria are the ideal model organisms for this type of studies as the high mutation rate in these organisms provide a rapid turnover of genotypes in

sequence space to test many combinations, they are relatively easy to work with experimentally, and they often have significant implications for human health.

Reference	Pattern of epistasis	Organism
Mukai (1969)	Synergistic epistasis	Invertebrate
deVisser et al. (1996, 1997a)	Synergistic epistasis	Algae
deVisser et al. (1997b)	Unclear / both	Fungus
Whitlock and Bourguet (2000)	Synergistic epistasis	Invertebrate
Elena (1999)	No epistasis	Virus
Wloch et al. (2001)	No epistasis	Eukaryote
Bonhoeffer et al (2004)	Antagonistic epistasis	Virus
Burch and Chao (2004)	Antagonistic epistasis	Virus
Jasnos and Korona (2007)	Antagonistic epistasis	Eukaryote
Elena and Lenski (1997)	Unclear / both	Bacteria
Gong et al. (2013)	Antagonistic epistasis	Virus
Sanjuan et al. (2004)	Antagonistic epistasis	Virus
Maisnier-Patin et al. (2005)	Unclear / both	Bacteria

Table 1. Average direction or general pattern of epistasis reported by various experimental studies on the distribution of epistatic effects.

Table 1 summarizes the average direction of epistasis reported by a number of selected experimental studies on organisms ranging from RNA viruses and prokaryotes such as E.coli to more complex eukaryotes including insects and yeast, and it clearly shows that there is no universal pattern of epistasis that prevails across

species. These studies describe a global fitness landscape summarizing the evolutionary properties of the underlying genotype-phenotype map. Fortunately, recent advances in high-throughput sequencing technologies have enabled generating all possible single mutants and double mutants in a specific region of a gene (Bank et al. 2014) or even the entire gene (Sarkisyan et al. 2015, see Appendix I), allowing studying the distribution of epistatic effects by reconstructing local fitness landscapes. A complete understanding of the dynamics of fitness landscapes requires the synthesis of local fitness landscapes (i.e., inferring the biophysical properties of the landscape) and global fitness landscapes (i.e., evolutionary properties of the landscape).

1.3. Searching for signatures of epistasis in humans

While the role of epistasis on macroevolutionary scale has been widely documented (Breen et al. 2012), whether epistasis contributes to standing genetic variation remains controversial (Hemani et al. 2013; Mackay 2014; Hill and Maki-Tanila 2014) despite its importance for our understanding of the genetic architecture of quantitative traits and disease. Notably, epistasis has significant implications for medically relevant issues such as missing heritability (Manolio et al. 2009; Zuk et al. 2012) and

accurate prediction of the functional impact of genetic variants, as well as for fundamental questions in evolutionary biology including the evolution of sex (Kondrashov 1988; de Visser and Elena 2007) and the mode and tempo of molecular evolution (Povolotskaya and Kondrashov 2010; Breen et al. 2012; McCandlish et al. 2013).

The controversy over the importance of epistasis within populations usually acknowledges the presence of epistasis of varying degree in segregating populations, however, is primarily concerned with the relevance of observed epistasis to heritable genetic variation. Most of the empirical evidence for the prevalent epistasis found in model organisms including yeast and fruit fly typically refers to epistatic interactions involving mutations of large phenotypic effects (Mackay 2014), whereas epistatic interactions, if exist, in natural populations are arguably expected to have small effects that have proven to be non-trivial to detect. Indeed, recent studies on the estimation of epistasis within populations – notably, in Drosophila (Huang et al 2012) and humans (Brown et al. 2014; Hemani et al. 2014) – report a low amount of epistasis, and therefore, it is argued that the contribution of epistatic interactions to genetic variation is inconsequential. Although ignoring epistasis has proven to be convenient in quantitative genetics (Nelson, Pettersson and Carlborg 2013) and in

important evolutionary practices such as animal breeding (Crow 2010), and moreover, is in complete accordance with the theoretical predictions (Hill et al. 2008), however, it remains unclear to what extent the observation of little epistasis reflects the true nature of epistasis or the conceptual and technical limitations on our current ability to detect epistasis.

It is important to note that the arguments for the insignificance of epistasis in natural populations traditionally presumes that epistatically interacting alleles can only contribute to non-additive components of genetic variation. As the studies on genetic variation in quantitative traits consistently show that most variation can be attributed to additive component (Hill et al. 2008), epistasis is often accordingly ruled out as a viable factor affecting the variation. However, it is not clear whether epistatic interactions can manifest their contribution as part of the additive component (Greene et al. 2009). Thus, this type of epistasis based on variance components is termed *statistical epistasis* and describes the average effect of an allele at the population-level. Alternatively, it is also possible to define epistasis at the individual level describing the interdependent contribution of alleles to phenotype or fitness. This implies that the effect of an individual allele on a trait depends on the presence of other alleles elsewhere in the genome (genetic background). This type of epistasis

based on context-dependency is termed *functional or biological epistasis*. It is important to emphasize that statistical and functional epistasis refer to different phenomenon, and the presence (absence) of one does not necessarily implies the absence (presence) of other. Some of the controversy in the field can be attributed to prematurely taking observation of functional epistasis as evidence of statistical epistasis.

In humans, uncovering the prevalence and role of epistasis in shaping genetic variation has been a challenging task due to statistical and technical limitations, and the issues pertinent to the challenge have been best described by the issue of missing heritability whereby genetic variations underlying most complex traits that are identified by the genome-wide association studies (GWAS) are in fact shown to explain only a small fraction of heritability observed for these traits. A poster case for the apparent missing heritability is height in humans, which is highly heritable and influenced by many loci. A GWAS in 183,727 individuals revealed 180 loci that influence adult height, however, these loci were shown to have a predictive power of only 10% of the observed phenotypic variation (Lango et al. 2010). Invoking an alternative prediction model whereby many common variants of very small effects are considered, the proportion of variation in height explained by such common variants has improved the predictive

power to about 40% (Yang et al. 2011), however, these estimates are far from the estimates of about 80% based on relatedness of relatives (Fisher 1918; Visscher et al. 2008). Epistasis has recently been argued to be important for the missing heritability (Manolio et al. 2009; Frazer et al. 2009; Eichler et al. 2010), in particular by causing overestimation of added variability (Zuk et al. 2012).

The idea of personalized medicine relies on our ability to accurately predict the consequence of genetic variants on quantitative and disease traits, and factors underlying the missing heritability may significantly hinder the predictive power of existing models. However, current literature on studies searching for epistatic interactions in humans points to an even more problematic issue: majority of the reported epistatic interactions do not replicate when tested in an independent data set, or are false positives due to poor study designs (Combarros et al. 2008; **Table 2**). While it is reasonable to expect the genuine epistatic interactions to replicate independent of the data set, it is also conceivable that epistatic interactions, particularly those involving rare variants, may not replicate across population due to issues related to sample size or phenotypic variability (Greene et al. 2009).

Study	Trait	Sample size†	Search space¥	Remarks
Carrasquillo et al. 2002	Hirschsprung's disease	43	1,494	Interaction between RET and EDNRB
Combarros et al. 2008	Alzheimer's disease	Meta-analysis of >100 studies	Meta-analysis of >100 studies	Majority of reported interactions are false positives
Barrett et al. 2009	Type-I diabetes	7,514	841,622	HLA
Evans et al. 2011	Ankylosing spondylitis	3,023	2,223,620	HLA-B27 and ERAP1
Wan et al. 2010	WTCCC††	14,925	356,441	Variants in MHC
Prabhu and Pe'er 2012	WTCCC†† (only for bipolar disorder)	1,868	374,481	Unable to replicate in independent cohort
Hemani et al. 2014	BSGS†††	846	528,509 (7,339 genes)	Few significant epistatic pairs
Wood et al. 2014	InCHIANTI study	450	30 (Hemani et al. 2014)	Detected pairs better explained by multi-locus LD

Brown et al. 2015	TwinsUK cohort	765	13,660 genes	Possibly contaminated by haplotype effects
Lippert et al. 2013	WTCCC††	14,925	356,441	Variants in MHC

Table 2. Some of the large-scale studies reporting empirical evidence for epistasis in humans. Notably, most examples in the list include an interaction involving a large effect contributed by the *HLA* or *MHC*, which may be due to haplotype effects.

¥ Search space refers to the total number of single nucleotide polymorphisms considered for the data analysis. The initial search space is typically reduced substantially in the later stages in line with the study design. Some studies included other types of genetic variants, including microsatellites or structural variants, in addition to single nucleotide polymorphisms.

† The number of individuals refers to cases, and does not include controls and replicate cohorts, and may vary among studies using the same dataset due to different filtering criteria.

†† Wellcome Trust Case Control Consortium (WTCCC) includes seven common diseases: bipolar disorder, coronary artery disease, hypertension, Crohn's disease, rheumatoid arthritis, type-I diabetes, and type-II diabetes.

††† Brisbane Systems Genetics Study includes data from 846 individuals on gene expression levels measured in whole blood.

Recent studies on the search for epistasis in humans have made use of available data on gene expression levels, which typically have large effect sizes and therefore are in

principle less affected by statistical limitations than epistatic interactions involving small effects (Hemani et al. 2014; Brown et al. 2014). These studies employ a two-stage search whereby an exhaustive search of all pairwise effects is followed by an attempt to replicate the significant hits from the previous search in an independent data set. Using this approach Hemani *et al.* searched for segregating polymorphisms in humans that influence expression levels in an epistatic manner, and found 501 significant hits among an initial search of all pairwise effects of about half a million SNPs. When the authors made an effort to replicate these hits in another data set, they were able to replicate only 30 interactions. In addition to very few instances of epistatic pairs, they estimated that the most of the phenotypic variation are due to large additive effects rather than large epistatic effects.

Suspecting that the apparent epistasis between two variants can instead be due to presence of a third variant associated with the two variants (multi-locus LD), Wood *et al.* replicated the 30 interactions reported by Hemani *et al.* in another independent data set explicitly controlling for the possibility of a third variant that could better explain what appears to be epistasis between the other two variants. Indeed, the authors were able to show for 27 out of 28 cases analyzed that the presence of a single causal variant moderately associated with the reported variants can better

explain the variation in expression levels. Notably, Hemani *et al.* controlled for haplotype effects in their study design by filtering based on pairwise LD, however, the haplotype effects due to multi-locus LD was missed. In a similar study design, Brown *et al*. tested for significant epistatic interactions influencing gene expression levels using RNA-sequence data from lymphoblastoid cell lines, and found an initial set of 508 associations of which 57 epistatic interactions were shown to replicate in an independent data set. Unlike Hemani *et al.*, Brown *et al.* concluded that about half of those interactions, epistatic variance explained more than additive variance. Nevertheless, a careful re-examination of the reported interactions showed that haplotype effects confounded all of these interactions even after Brown *et al*. made an effort to explicitly control for possible haplotype effects by replicating the significant interactions using whole genome sequence data (Wood et al. 2014).

In summary, detecting and replicating epistasis in humans have been challenging due to a variety of confounding factors including haplotype effects that can result in false positives even in carefully designed studies, cohort or population specific signals that are unlikely to replicate in another cohort or population.

References

- Allen HL et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature.*
- Anderson DW, McKeown AN, Thornton JW. (2015) Intramolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife.*
- Bank C et al. (2014) A systematic survey of an intragenic epistatic landscape. *Molecular Biology and Evolution.*
- Barrett JC et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics.*
- Bonhoeffer S et al. (2004) Evidence for positive epistasis in HIV-1. *Science.*
- Breen M et al. (2012) Epistasis as the primary factor In molecular evolution. *Nature.*
- Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constraints the direction of glucocorticoid receptor evolution. *Nature.*
- Brown AA et al. (2014) Genetic interactions affecting human gene expression identified by variance association mapping. *eLife.*

- Burch CL and Chao L (2004) Epistasis and its relationship to canalization in the RNA virus phi6. *Genetics.*
- Burga A, Casanueva MO, Lehner B (2011) Predicting mutation outcome from early stochastic variation in genetic interaction partners. *Nature.*
- Burga A and Lehner B (2012) Beyond phenotype to genotype: why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience. *FEBS Journal.*
- Carrasquillo MM et al. (2002) Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nature genetics.*
- Combarros O et al. (2008) Epistasis in sporadic Alzheimer's disease. *Neurobiology of aging.*
- Crow JF (2010) On epistasis: why is it unimportant in polygenic directional selection. *Philosophical Transactions B.*
- deVisser J, Hoekstra RF, den Ende HV (1996) The effect of sex and deleterious mutations on fitness in Chlamydomonas. *Proceedings of the Royal Society B.*
- deVisser J, Hoekstra RF, den Ende HV (1997a) An experimental test for synergistic epistasis and its application in Chlamydomonas. *Genetics.*

- deVisser J, Hoekstra RF, den Ende HV (1997b) Test of interaction between genetic markers that affect fitness in Aspergillus niger. *Evolution.*
- Eichler EE et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Review Genetics.*
- Elena SF and Lenski R (1997) Test of synergistic interactions between deleterious mutations in bacteria. *Nature.*
- Elena SF (1999) Little evidence for synergism among deleterious mutations in a nonsegmented RNA virus. *Journal of Molecular Evolution.*
- Evans DM et al (2011) Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nature Genetics.*
- Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh.*
- Frazer KA et al. (2009) Human genetic variation and its contribution to complex traits. *Nature Review Genetics.*
- Gong LI, Suchard MA, Bloom JD (2013) Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife.*

- Greene CS et al. (2009) Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS ONE.*
- Hemani et al (2013) Detection and replication of epistasis influencing transcription in humans. *Nature.*
- Hill GW, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics.*
- Hill GW and Maki-Tanila A (2014) Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. *Journal of Animal Breeding and Genetics.*
- Huang W et al. (2012) Epistasis dominates the genetic architecture of Drosophila quantitative traits. *PNAS.*
- Jasnos L and Korona R (2007) Epistatic buffering of fitness loss in yeast double deletion strains. *Nature genetics.*
- Kondrashov AS (1988) Deleterious mutations and the evolution of sexual reproduction. *Nature.*
- Lippert C et al. (2013) The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific reports.*
- Lozovsky ER et al. (2009) Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *PNAS.*

- Mackay TF (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Review Genetics.*
- Maisnier-Patin S et al. (2005) Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nature genetics.*
- Manolio TA et al. (2009) Finding the missing heritability of complex diseases. *Nature.*
- Maynard Smith J (1970) Natural selection and the concept of a protein space. *Nature.*
- McCandlish DM et al. (2013) The role of epistasis in protein evolution. *Nature.*
- Mukai T (1969) The genetic structure of natural populations of Drosophila melanogaster. *Genetics.*
- Natarajan C et al. (2013) Epistasis among adaptive mutations in deer mouse hemoglobin. *Science.*
- Nelson RM, Pettersson ME, Carlborg O (2013) A century after Fisher: time for a new paradigm in quantitative genetics. *Trends in genetics.*
- Ortlund EA et al. (2007) Crystal structure of an ancient protein: evolution by conformational epistasis. *Science.*
- Poelwijk FJ et al. (2007) Empirical fitness landscapes reveal accessible evolutionary paths. *Nature.*
- Povolotskaya I and Kondrashov FA (2010) Sequence space and the ongoing expansion of the protein universe. *Nature.*

- Prabhu S and Pe'er I (2012) Ultrafast genome-wide scan for SNP-SNP interactions in common complex diseases. *Genome Research.*
- Sanjuan R, Moya A, Elena SF (2004) The contribution of epistasis to the architecture of fitness in an RNA virus. *PNAS.*
- Sarkisyan KS et al. (2015) Local fitness landscape of the green fluorescent protein. *(See Appendix I)*
- Wan X et al. (2010) Detecting two-locus associations allowing for interactions in genome-wide association studies. *Bioinformatics.*
- Wei W, Hemani G, Haley CS (2014) Detecting epistasis in human complex traits. *Nature Review Genetics.*
- Weinreich DM et al. (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science.*
- Whitlock MC and Bourguet D (2000) Factors affecting the genetic load in Drosophila: synergistic epistasis and correlations among fitness components. *Evolution.*
- Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era – concepts and misconceptions. *Nature Review Genetics.*
- Wloch DM et al. (2001) Direct estimate of the mutation rate and the distribution of fitness effects in the yeast Saccharomyces cerevisiae. *Genetics.*

- Wood AR et al. (2014) Another explanation for apparent epistasis. *Nature.*
- Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the sixth international congress of genetics.*
- Yang J et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature.*
- Zuk O et al. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS.*

Chapter 2

Estimating the rate of irreversibility in protein evolution

Soylemez, O. & Kondrashov, F.A.
Genome Biology and Evolution 480, 250-253 (2012)

Preface

In this work we obtain the first estimate of the propensity of irreversible evolution on the molecular level of protein sequence.

The issue of whether or not evolution is irreversible has been discussed by Darwin and elaborated in its present form by Dollo in 1893. Muller and Dobzhansky in 1930's have also touched on this question in their study of genetic incompatibilities in the course of speciation. However, despite the importance and attention to this problem over the course of the last century, mostly on the level of phenotype, we believe that we are the first to look at this question from a quantitative perspective on the level of amino acid sequence of specific proteins.

In this work we relate the available data on human disease mutations to the reconstructed amino acid sequence in the common ancestor of human and other placental mammals. We find that between 10 and 50

percent of all ancestral amino acid states in the placental phylogeny match a described disease mutation to humans. Such cases represent instances of inherent evolutionary irreversibility as the modern human lineage cannot revert to these ancestral states because they cause disease and, therefore, cannot achieve fixation at present. The high fraction of irreversibility of all molecular evolution is surprising and we hypothesize that it is caused by the nature of protein structure and function.

Soylemez O, Kondrashov FA. Estimating the rate of irreversibility in protein evolution. Genome Biol Evol. 2012;4(12):1213-22. doi: 10.1093/gbe/evs096

Chapter 3

Massively parallel enzyme kinetics reveals the substrate recognition landscape of ADAMTS13

Kretz C., Dai M., **Soylemez O.**, Yee A., Desch K., Siemieniak D., Tomberg K., Kondrashov F.A., Meng F., Ginsburg D. *(2015) Proceedings of the National Academy of Sciences doi:10.1073/pnas.1511328112*

Preface

Accurately predicting the impact of genetic variation on the protein structure and function is an important challenge in evolutionary biology and genomic medicine. Currently available predictive computer algorithms such as Polyphen and SIFT can only explore a fraction of all possible mutations in the human exome space, and moreover, these predictions do not take into account the biological context (e.g., protease/substrate pairs), which may lead to an incomplete and poor understanding of the functional annotation of the variants.

In this manuscript, we report a method based on high-throughput sequencing of a phage library to rapidly screen all possible substrate residues in the coagulation protein von Willebrand factor (VWF) to assess important interaction sites with its cognate protease ADAMTS13. The

method provides a comprehensive picture of the substrate recognition landscape, successfully identifying the key sites as well as amino acid changes that are important for the interaction between ADAMTS13 and VWF.

Reconstructing empirical substrate recognition landscapes can provide insights into the impact of genetic variation on the fitness, and the method described here can be broadly applicable to many other protease/substrate pairs.

Kretz CA, Dai M, Soylemez O, Yee A, Desch KC, Siemieniak D, Tomberg K, Kondrashov FA, Meng F, Ginsburg D. Massively parallel enzyme kinetics reveals the substrate recognition landscape of the metalloprotease ADAMTS13. Proc Natl Acad Sci U S A. 2015 Jul 28;112(30):9328-33. doi: 10.1073/pnas.1511328112.

Chapter 4

Epistasis among natural polymorphisms in humans detected by structural analysis

Soylemez, O., Ivankov D.N., & Kondrashov, F.A.

Abstract

Whether or not epistasis, the interdependent contribution of alleles to fitness, shapes standing genetic variation is crucial for our understanding of the issue of missing heritability (Zuk et al. 2012), personalized medicine (Mackay and Moore 2014), the evolution of sex (Kondrashov 1988) and molecular evolution (Breen et al. 2012), yet remains the subject of intense controversy (Hill et al. 2008; Mackay 2014). Indirect evidence, mostly from model organisms, suggests that epistatically interacting alleles might contribute to standing genetic variation (Mackay 2014). However, the status quo in human genetics is that epistasis is hard to detect or is very infrequent as to matter for heritable variation, and as a result epistatic interactions continue to be considered inconsequential to the study of human genetic traits. Here, we combine data on single nucleotide polymorphisms (SNPs) in the human genome and information on spatial proximity of amino acid residues in the protein tertiary structure to detect the signal of epistasis shaping standing human genetic variation.

From high-resolution protein tertiary structures of 6685 human protein-coding genes, we obtain data on structurally-interacting sites in the three-dimensional protein structure. We find that pairs of nonsynonymous SNPs that are found in the same individual are more likely to occur in sites that belong to a structurally-linked residue pair compared to synonymous SNP pairs, likely reflecting the pressure of epistasis maintaining these polymorphisms in the same haplotype. We estimate that several dozen pairs of nonsynonymous SNPs within the available 2504 human genomes were influenced by intragenic epistasis with a structural basis. Our approach can be used to detect epistasis in other species, and may aid in the resolution of the persisting controversies on standing variation.

Introduction

Whether or not epistasis, the non-additive contribution of alleles to fitness, plays a role in shaping standing genetic variation has been controversial issue in genetics throughout the last several decades (Fisher 1918; Visscher, Hill and Wray 2008; Hill, Goddard and Visscher 2008; Crow 2010; Hill and Maki-Tanila 2014). In theory, epistasis can shape standing variation through selection maintaining favorable allele combinations in linkage disequilibrium (LD), whereby such combinations would be found in the same haplotypes more frequently than by

chance alone (Lewontin 1974). On the other hand, non-selective evolutionary forces, including genetic drift and population structure, play a large role in shaping LD across genomes (Ardlie, Kruglyak and Seielstad 2002; Slatkin 2008). At present, epistasis is typically not thought to play a large role in shaping standing genetic variation with the few studies reporting a signature of epistasis in standing variation (Hemani et al. 2014; Brown et al. 2014) have been met with some criticism (Wood et al. 2014). Nevertheless, the debate is likely to continue as it may be crucial for our understanding of complex phenotypes, including those relevant for medical genetics (Mackay and Moore 2014). By contrast, evidence indicating the importance of epistatic interactions between amino acid substitutions across long evolutionary timescales is accumulating (Breen et al. 2012, McCandlish et al. 2013). A substantial fraction of such interactions is thought to have been driven by intra-protein interactions for the maintenance of structural stability of proteins (Ferrer-Costa, Orozco and de la Cruz 2007; Baresic et al. 2010; Ivankov, Finkelstein, and Kondrashov 2014; Sikosek and Chan 2014). Therefore, in the present study we focus on the role of epistasis among variants found at structurally-interacting sites in shaping standing variation in the human population.

Results

We obtained high-resolution protein tertiary structures of 6685 proteins coded in the human genome from PDB (Berman et al. 2008) and identified structurally-interacting sites, those located in close proximity in the three-dimensional protein structure (see **Methods**). We then obtained data on human SNPs from 2504 individuals from the 1000 Genomes Project (McVean et al. 2012) determining two common measures of LD – r^2 and D' –, between all intragenic nonsynonymous and synonymous SNPs in the 6685 genes with available data on protein structure and residue contacts (**Table 1**). We related the data on human SNPs to the tertiary protein structures, recording the fraction of SNPs found in codons that code for structurally interacting amino acid residues. The distribution of pairs of synonymous SNPs is expected to be independent relative to the protein tertiary structure, whereas pairs of nonsynonymous SNPs may be epistatic for the maintenance of protein structural stability. Thus, an excess of nonsynonymous SNP pairs under high LD relative to such synonymous SNP pairs would indicate the action of epistasis.

	Mean	Standard deviation	Number of genes
# of pairs of structurally-interacting sites	1450	1365	6588
# of nonsynonymous SNPs	11.5	17	6057
# of synonymous SNPs	10	12	6283

Table 1. Summary of data on structurally-interacting sites and number of codons with nonsynonymous and synonymous SNPs, respectively, for each gene with available data on structure.

We find that pairs of nonsynonymous SNPs were significantly more likely to be found at structurally interacting sites compared to synonymous SNP pairs when the pair was under high LD for r^2 (**Figure 1)**. Amino acid residues located in close vicinity in structure are also more likely to be coded by codons located on the DNA closer than random. This is reflected in the observation that synonymous SNP pairs under high LD are more often found in codons coding for residues close together in structure. The signature of epistasis for nonsynonymous SNP pairs, therefore, is presented in the higher fraction of nonsynonymous SNP pairs under high LD compared with synonymous SNP pairs that occur in codons coding for residues in direct structural interaction. Overall, we identify 95 pairs of nonsynonymous SNPs in 59 genes that are

likely to interact in an epistatic manner due to structural constraints.

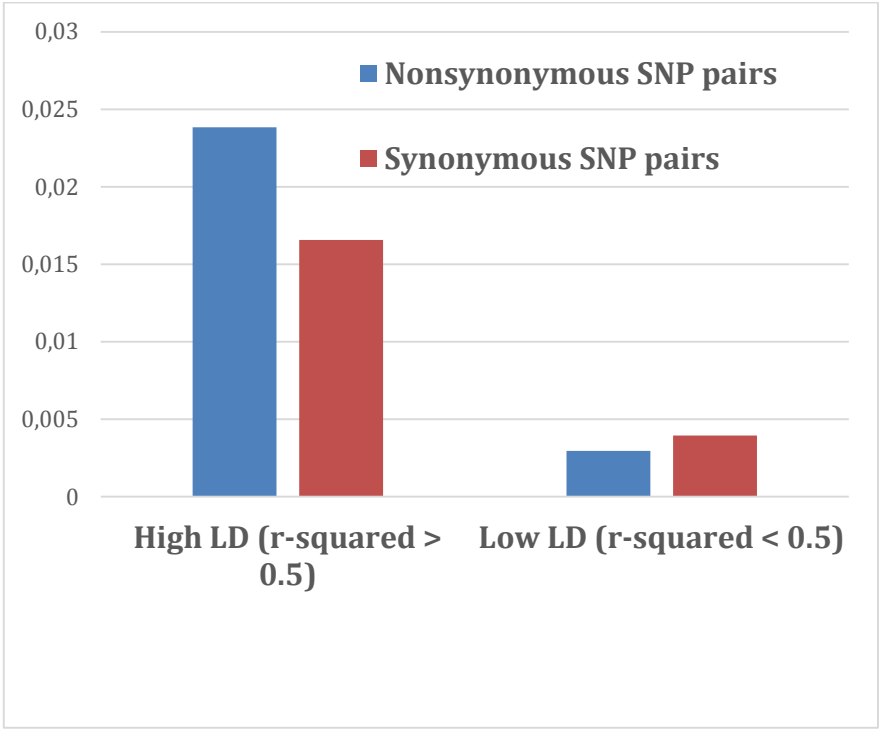

Figure 1. Fraction of SNP pairs under high and low LD when measured by r^2 that correspond to structurally-interacting sites in the 3D protein structure. All comparisons are statistically significant. (Fisher's exact test at p-value=0.05)

In contrast, the pairs under high LD when measured by *D'* do not show the pattern observed for the pairs under high LD for r^2, possibly reflecting that *D'* can artificially inflate LD estimate when one of the alleles is rare and is more likely to include nearly or completely neutral alleles unlike r^2 which is likely to register non-neutral alleles. **(Figure 2)**

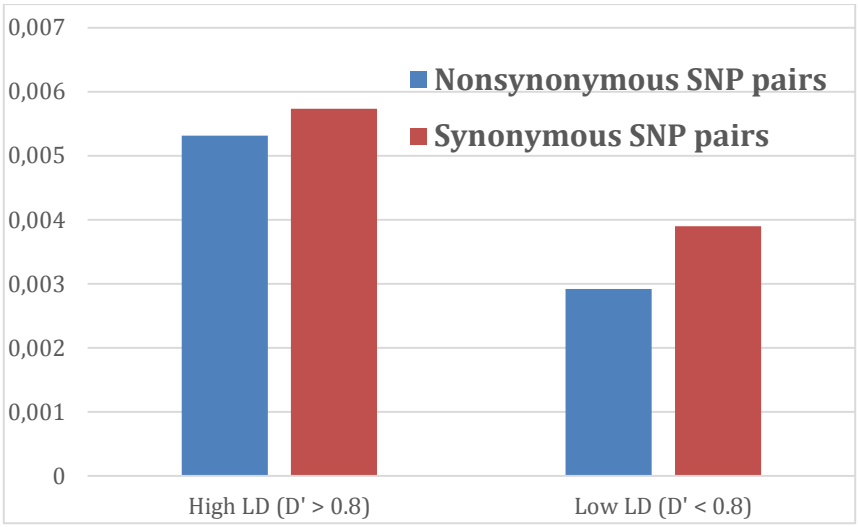

Figure 2. Fraction of SNP pairs under high and low LD when measured by *D'* that correspond to structurally interacting sites in the 3D protein structure. All comparisons are statistically significant. (Fisher's exact test at p-value=0.05)

The r^2 measure of LD normalizes the co-occurrence of alleles by the allele frequency of the minor derived allele of the two SNPs. For a SNP pair A1A2 and B1B2, where A and B are sites and 1 and 2 ancestral and derived alleles, respectively, the r^2 measure assigns a high LD value to instances when A1B1 and A2B2 genotypes are frequent while both A1B2 and A2B1 genotypes are rare. Alternatively, the *D'* measure does not normalize by the derived allele frequency, such that instances when A1B1, A2B1, and A2B2 genotypes are common while A1B2 genotypes are rare may also be considered under high LD as long as the frequency of the A1 allele is sufficiently high.

To investigate whether the observed difference in the overall pattern of fitness effects of SNP pairs presented for r^2 and D' can be explained by the difference in the type of epistasis that these two measures are expected to catch – reciprocal and non-reciprocal epistasis, respectively – we compared the prevalence of each of all four possible allele combinations in a SNP pair across long-term evolution. For each SNP pair under high LD ($r^2 > 0.5$ or $D' > 0.8$) that are found in structurally-interacting sites, we calculated the proportion of each of four possible allele combinations found in the multiple sequence alignments of the vertebrate sequences orthologous to the genes analyzed in this study. The expected difference between the two distributions of allele combinations across two different measures of LD would in principle demonstrate whether the intermediate allele combinations involving only a single mutant is found more often, that is presumably less deleterious, in the long-term evolution than the combination with the double mutant involving both minor alleles.

We found that the double mutant (*ab*) is found more often than the less prevalent of the intermediate combination (*Ab*) when LD is measured by r^2, that is, co-occurrence of both minor alleles (*ab*) has a higher fitness (more prevalent) than the less prevalent intermediate combination (*Ab*); however, in contrast to expectation, more

prevalent of the intermediate combinations (*aB*) is observed more often in the evolution than the double mutant, providing no evidence for sign epistasis. (**Figure 3**)

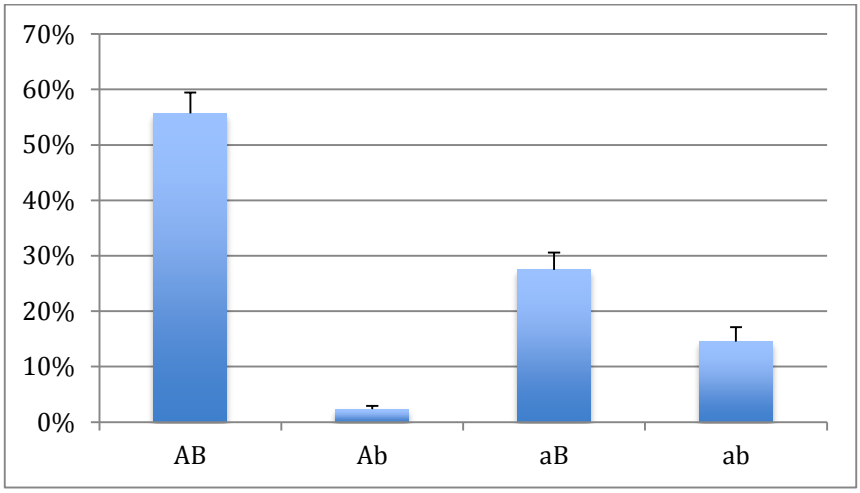

Figure 3. Frequency of allele combinations of SNP pairs under high LD (r^2 > 0.5) that correspond to a structurally-interacting sites across vertebrate evolution. *AB* represents the most prevalent allele combination in the human population. *Ab* is chosen to represent the less prevalent of the two intermediate allele combinations. Standard error bars are indicated.

Similarly, when pairs under high LD as measured by *D'* are considered for their prevalence across long-term evolution, we found that the less prevalent of the intermediate allele combinations (*Ab*) is found less prevalent than the double mutant while the more prevalent of the intermediate allele combinations (*aB*) is practically indistinguishable from the wild-type (*AB*) reflecting that *D'* assigns high values to pairs where one allele is relatively common and therefore likely to be neutral. **(Figure 4)**

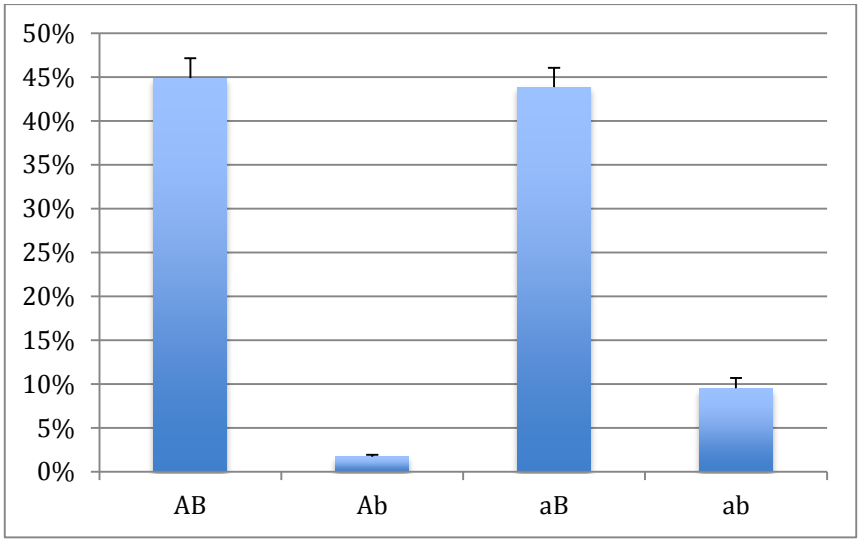

Figure 4. Frequency of allele combinations of SNP pairs under high LD (*D'* > 0.8) that correspond to a structurally-interacting sites across vertebrate evolution. *AB* represents the most prevalent allele combination in the human population. *Ab* is chosen to represent the less prevalent of the two intermediate allele combinations. Standard error bars are indicated.

In the case of LD as measured by r^2, we observe that each intermediate allele combination (*Ab* and *aB*) involves deleterious (fitness-lowering) changes with respect the wild-type combination (*AB*). Therefore, we can evaluate the type of epistasis observed for these combinations by comparing the prevalence of the double mutant with the fitness loss due to single mutants revealing positive epistasis as the fitness loss due to double mutant is smaller than the fitness loss combined for the single mutants alone. Similarly, pairs calculated using *D'* show positive epistasis provided that the

more prevalent of the two intermediate combinations (*aB*) is considered as slightly deleterious.

Discussion

Different lines of evidence indicate that co-evolving sites are interacting in an epistatic manner. First, the overall abundance of epistasis in shaping molecular evolution (Breen et al. 2012; McCandlish et al. 2012) that often has a structural basis for epistatic interaction (Ferrer-Costa, Orozco and de la Cruz 2007; Baresic et al. 2010; Ivankov, Finkelstein, and Kondrashov 2014; Sikosek and Chan 2014). Second, analysis of co-evolving sites is used in protein structure prediction whereby co-evolving sites are demonstrated to be physically interacting in the protein structure. Given the structural basis of many instances of epistatically interaction sites we reasoned that co-evolving sites can be used as a proxy for epistatically interacting sites. Using this reasoning, we provide the first quantification of the degree to which epistasis shapes human genetic variation. Our simple, sequence-based approach estimates dozens of epistatic pairs of SNPs, that some of them can be revealed to have a structural basis for their interaction. Our method can be applied to non-coding genes and other organisms provided enough data on SNPs and orthologues, which is currently unavailable but is a matter of time, especially for model organisms.

Our estimate is likely to be underestimate of the true degree of epistasis because some sites that are apparently correlated in the course of evolution but not occur in close proximity in the protein structure may harbor long-range functional interactions between these sites (Socolich et al. 2005; Knaggs et al. 2007; Noivirt-Brik, Unger and Horovitz 2009; Kowarsch et al. 2010). Second, we estimate the degree of epistasis involving only pairwise interactions while higher order epistasis involving more than two sites may be also common and important. Interactions involving multiple sites are particularly pertinent to the estimates of true extent of epistasis as they may give rise to haplotype blocks that can result in false positives when searching only for pairwise interactions. Indeed, two recent studies reporting for the first time evidence for epistasis influencing transcription levels in humans have been shown to be biased by the haplotype effects such that the effect attributed to the supposedly epistatic interaction involving variants at site 1 and site 2 can be better explained by a single causal variant that is under LD with site 1 and site 2. (Hemani et al. 2014; Brown et al. 2014; Wood et al. 2014) While our study design should in principle be immune to this type of haplotype effects as we consider explicitly those SNP pairs that are under LD in the human population, the haplotype effect can still affect some of the pairs we detect by structural analysis as the haplotype blocks can maintain

allele combinations due to non-selective forces whereas proposed structural basis for allele combinations invoke adaptive explanations.

Our analysis of epistatic interactions using data on genetic variation and structurally-interacting sites can provide insights into the discussion of the role and prevalence of epistasis in natural population, the relative importance of structural and evolutionary information when studying the distribution of genetic variants in a population, and accurate / context-dependent evaluation of the functional impact of genetic variants that remains to be the biggest challenge facing personalized genomics.

Methods

SNP pairs under linkage disequilibrium

We obtained data on SNPs from the version 5 of the latest release of the Phase 3 variant set from the 1000 Genomes Project. We converted the nucleotide-based variation files provided by the 1000 Genomes Project into codon-based variation files as multiple SNPs may correspond to the same codon and result in an amino acid residue different from the residue that might be obtained via individual variants. Excluding singletons and private doubletons, pairwise linkage disequilibrium (LD) – as measured by r^2 and D' – between all intragenic nonsynonymous and

synonymous SNP pairs was calculated for representative transcripts of the protein-coding genes (GENCODE v19, and SNP pairs with $r^2 > 0.5$ (or $D' > 0.8$) were considered to be under high LD.

Identification of structurally-interacting sites

For 17512 human protein-coding genes from ENSEMBL, ran protein-protein BLAST ("blastp", BLAST version 2.2.26+) against local copy of database "pdbaa" (downloaded on January, 31st 2015) using tabular output option ("-outfmt 6"). We retained hits having a sequence identity of at least 50% and e-value of at most 1E-03, resulting in a total of 9349 PDB structures, of which 41 were not found in the local mirror of PDB. For the remaining 9308 PDB structures we identified residue pairs having at least one pair of non-hydrogen atoms located at the distance of 6Å or less by in-house PERL script. We mapped found contacts to the original human sequences by aligning fragment of human sequences to the fragment of sequence from PDB file found by BLAST. For this mapping, we used an in-house PERL script that implements Needleman-Wunsch algorithm with BLOSUM62 matrix, gap initiation penalty equal to -1 and gap elongation penalty equal to -1. If we found residue contact in different PDB structures, we retained only one contact corresponding to lower e-value, or the one with higher sequence identity if e-values were identical. In case that both the sequence identity and the e-

value are identical, the first was chosen. Finally, the contacts between neighboring residues were excluded because they are always within 6Å from each other.

References

- Ardlie KG, Kruglyak L and Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nature Review Genetics.*
- Baresic A et al. (2010) Compensated pathogenic deviations: analysis of structural effects. *Journal of Molecular Biology.*
- Berman HM et al. (2000) The Protein Data Bank. *Nucleic Acids Research.*
- Breen M et al. (2012) Epistasis as the primary factor in molecular evolution. *Nature.*
- Brown AA et al. (2014) Genetic interactions affecting human gene expression identified by variance association mapping. *eLife.*
- Crow JF (2010) On epistasis: why is it unimportant in polygenic directional selection. *Philosophical Transactions B.*
- Ferrer-Costa C, Orozco M and de la Cruz X (2007) Characterization of compensated mutations in terms of structural and physic-chemical properties. *Journal of Molecular Biology.*
- Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance.

Philosophical Transactions of the Royal Society of Edinburgh.

- Hemani et al (2013) Detection and replication of epistasis influencing transcription in humans. *Nature.*
- Hill GW, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics.*
- Hill GW and Maki-Tanila A (2014) Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. *Journal of Animal Breeding and Genetics.*
- Ivankov DI, Finkelstein AV and Kondrashov FA (2014) A structural perspective of compensatory evolution. *Current Opinion in Structural Biology.*
- Knaggs MH et al. (2007) Insights into correlated motions and long-range interactions in CheY derived from molecular dynamics simulations. *Biophysics Journal.*
- Kowarsch A et al. (2010) Correlated mutations: A hallmark of phenotypic amino acid substitutions. *PLoS Computational Biology.*
- Lewontin RC (1974) The genetic basis of evolutionary change. *Columbia University Press.*
- Mackay TF (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Review Genetics.*

- Mackay TF and Moore JH (2014) Why epistasis is important for tackling complex human disease genetics. *Genome Medicine.*
- McCandlish DM et al. (2013) The role of epistasis in protein evolution. *Nature.*
- McVean et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature.*
- Noivirt-Brik O, Unger R and Horovitz A (2009) Analysing the origin of long-range interactions in proteins using lattice models. *BMC Structural Biology.*
- Sikosek T and Chan HS (2014) Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society Interface.*
- Slatkin M (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Review Genetics.*
- Soccolich M et al. (2005) Evolutionary information for specifying a protein fold. *Nature.*
- Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era – concepts and misconceptions. *Nature Review Genetics.*
- Wood AR et al. (2014) Another explanation for apparent epistasis. *Nature.*
- Zuk O et al. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS.*

Chapter 5

Concluding remarks

The aim of this thesis work was to investigate the role of epistasis in short-term and long-term protein evolution. In particular, the studies presented here were set out to bridge the gap between micro- and macro-evolution by addressing the following questions:

1. What can we learn from studying genomic incompatibilities across genetic backgrounds at the species level about the nature of epistasis in protein evolution?
2. Can the evolutionary analysis of epistatic interactions in the long-term protein evolution help improve accuracy of predicting the impact of genetic variation on the protein function?
3. To what extent evolutionary and structural information gathered from long-term evolution can be informative about the genetic variation in natural populations?

Compensatory interactions in protein evolution

In line of these broad questions, we first studied the prevalence of epistatic interactions in protein evolution by surveying genomic incompatibilities across species. These

incompatibilities correspond to instances where a certain allele is acceptable in one genetic background but the same allele is deleterious (e.g., disease-associated variant) in another genetic background. Remarkably, there are many amino acid changes associated with disease in humans while the causal variant is present in non-human species, apparently without invoking the drastic phenotypic changes and disease traits observed in humans. It is conceivable to reason that the effect of this variant on the fitness depends on the genetic context including, for example, other genetic changes in the same protein or in the interacting partner that modify the effect.

In our work on compensatory evolution (Soylemez & Kondrashov, 2012), we used data on known pathogenic mutations to answer two important questions regarding such context-dependent changes. First, we were interested in finding out whether such changes – disease-associated mutations in humans that appear to be wild-type states in non-humans – are prevalent in evolution. Second, we used a phylogenetic approach to estimate how far back in time we need to go to find an amino acid state in the ancestral protein sequence that is currently deleterious in humans, or how fast an amino acid state that was acceptable in our ancestors has become deleterious in modern humans.

We used a phylogenetic approach to reconstruct ancestral states at sites in human protein-coding genes at which disease-associated mutations are known, and determined how often the disease-associated mutation was present as the ancestral state. Such states are indicative of epistasis, because ancestral states presumably did not confer the low-fitness phenotypes. We reasoned that ancestral states become unacceptable in modern humans because of epistatic changes at other sites that occurred in the lineage leading to the humans. We found that, correcting for potential sources of bias, about 10% of all amino acid changes are irreversible.

There are two limitations in our analysis. First, it is important to highlight the distinction between the definition of irreversibility used in our study, which refers to reversing a particular mutation on the current human genetic background to the ancestral state, and the definition employed by others in the literature (Bridgham et al. 2009 and Gore), which refers to entire evolutionary trajectories leading back to the ancestral state of the protein. In principle, it is arguable that the latter definition provides a more comprehensive picture of the dynamics of evolutionary landscapes, which would require identification of all evolutionary events involved, whereas our definition can be more informative about the evolutionary

mechanisms underlying the compensatory changes and less about the dynamics of evolutionary landscape.

Second, we cannot conclusively distinguish the permissive and restrictive mutations that might have played a role leading to the observation that some ancestral states are currently not allowed in humans. Suppose an amino acid substitution, A53T, is associated with disease in humans, and the disease-causing state T was inferred to be present in the ancestor. Assuming that the ancestral state was not associated with disease, there are three possible scenarios for explaining why state T is currently not acceptable in humans: 1) the acceptable state A in humans is acceptable only because, for instance, of the epistatic interaction with the state R at position 86, which only allows A (not T) at position 53; therefore, state T at position 53 becomes unacceptable because of a restrictive mutation in the human lineage, 2) state T at position 53 might perhaps be associated with disease in the ancestor but was not in non-human species because of the epistatic interaction with the derived state S at position 65 in non-human species, which allowed state T at position 53; therefore, state T at position becomes unacceptable in humans but is acceptable in other species because of a permissive mutation in the non-human lineage, or 3) the acceptable state T in the ancestor was acceptable only because of the epistatic interaction with the state M at position 95, which

only allows T at position 53; therefore, state T at position 53 was acceptable because of a restrictive mutation on the ancestral genetic background. We cannot currently distinguish between these scenarios without identifying all the epistatic changes involved, and it remains elusive whether restrictive or permissive mutations contribute more often to epistatic interactions in evolution.

Concordance between the evolutionary analysis of epistatic sites and the reconstructed empirical fitness landscape

Studying compensatory interactions in protein evolution by making use of known pathogenic mutation can evidently provide insights into the prevalence of epistatic interactions, however, to what extent putative compensatory or epistatic changes identified using computational prediction methods are in concordance with the empirical observation? The concordance is particularly crucial when assessing the impact of genetic variants that are of clinical relevance and importance.

Recent advances in rapid and low-cost sequencing technologies have enabled exhaustive search for epistatic interactions by assaying fitness for all possible mutants in a region (Kretz et al. 2015, see Chapter 3) or even an entire gene (Sarkisyan et al. 2015, see Appendix I) of interest. As described in Kretz et al. 2015, we used a novel method

based on high-throughput sequencing of a phage library for the rapid screening of individual substrate residues in von Willebrand Factor (VWF) protein to evaluate key interaction sites with its cognate protease ADAMTS13. This method allowed us to study ADAMTS13-dependent enzymatic kinetics in the VWF73 (73 amino acid long substrate region of VWF), and measure reaction rates (k_{cat}/K_m values) for every individual and multiple mutations.

A comparison of the k_{cat}/K_m values obtained for either single variant alone to the values for pairs of amino acid substitutions revealed no significant epistatic interactions. Moreover, an evolutionary analysis showed that sites in VWF73 with higher impact on k_{cat}/K_m values tended to evolve slower than sites with low impact, indicating that the VWF73 evolutionary sequence conservation reflects the average impact of a mutation in the human VWF73 sequence. Despite the apparent lack of epistasis in VWF73 based on mutational and evolutionary analyses, however, it is possible that this particular region may be not epistatic or not as much epistatic as the rest of the gene. Indeed, VWF73 region of the gene lacks substantial tertiary structure (Crawley et al., 2011), and therefore, may be less prone to epistatic interactions, in particular, with a structural basis. To test the hypothesis that the VWF73 and VWF differ in their propensity for epistasis, we employed the method described previously

(Kondrashov, Sunyaev and Kondrashov 2002; Soylemez and Kondrashov 2012) and found that the proportion of disease mutations found in other species within VWF73 was not different from the rest of the VWF protein. We also observed that the proportion of mutations that lower k_{cat}/K_m values found in other species was similar to the proportion of mutations that increase k_{cat}/K_m values. This observation from the evolutionary analysis is particularly illustrative in light of the observation from the mutational analysis regarding the known pathogenic mutations associated with von Willebrand disease (VWD). Analysis of the k_{cat}/K_m values for known VWD-2A mutations within VWF73, which lead to VWD by enhancing the cleavage by ADAMTS13 (i.e., higher k_{cat}/K_m values), showed that more than half of the known VWD-2A mutations instead resisted cleavage. These observations suggest potential value of evolutionary information to predict the functional impact of mutations in VWF73. In particular, evolutionary information on possible epistatic interactions can complement the empirical data that may be underpowered – by design such as quasi-proxy choice of phenotype for fitness or due to technical limitations such as imprecise measurements – to detect interactions with relatively small effects.

Searching for signatures of epistasis in humans

Genome-wide association studies do not take into consideration epistasis as such the phenotypic variation in a quantitative trait is estimated by the joint effect of individual loci contributing while ignoring interaction terms. While this practice has proven to be useful for many quantitative traits for which the variation is mostly additive and hence arguably rendering the contribution of possible epistatic interactions inconsequential, accurate prediction of the impact of genetic variants on quantitative traits and clinically-relevant phenotypes requires a good understanding of the role of epistasis in the genotype-phenotype association.

The quest for identifying instances of epistatic interactions in humans and availability of large amount of data on genotypic and phenotypic variation has recently prompted novel association study designs to search for epistasis while keeping in mind the important statistical limitations such as the challenge of detecting epistasis involving variants of small effect size. Accordingly, there is an interest in searching for epistasis using available data on gene expression levels, which typically have large effects and therefore may be less vulnerable to statistical limitations. Nevertheless, most of the epistatic interactions that are reported to influence gene expression levels in humans are not robust examples of true epistasis but are instead often confounded by haplotype effects even after

making an effort to control for these effects (Hemani et al. 2014; Brown et al. 2014).

In light of these studies, we reasoned that a hypothesis-driven search for epistasis by narrowing the search space and rigorous follow up on validating the identified interactions can help detect epistasis, if any, in humans. Similar to studying epistasis in gene expression level, we decided to focus on a phenotype that arguably has intermediate to large effect sizes and therefore may serve as a candidate set against which putative interactions can be tested. Moreover, we aimed to leverage the known haplotype effects confounding the signal of epistasis by specifically searching for epistatic interactions among variants that are in high linkage disequilibrium. Assuming that structurally-linked contacts within the tertiary protein structure introduce spatial constraints on residue pairs, we quantified the degree to which pairs of intragenic variants in the human genome correspond to residue-residue contacts in protein structure, indicating putative epistatic interactions with a structural basis. To control for random associations, we used the correspondence between structurally-linked contacts and pairs of intragenic synonymous variants, which are presumably neutral in their effect on the protein structure, as the baseline or neutral expectation for the random correspondence. Thus, we showed that intragenic pairs of nonsynonymous variants are 25% more likely to

overlap a structurally-linked contact than their synonymous counterparts, and identified dozens of instances of epistatic interactions between nonsynonymous variants found, unsurprisingly, in genes such as HLA.

In principle, there are two ways to test whether the epistatic pairs identified in our study are indeed genuine examples of epistasis and not just an apparent association due to haplotype effects indicating a structurally-linked third, causal variant that can better explain the correspondence. First, we can make an attempt to validate the functional association by showing that observed pairs of variants result in a structurally more stable conformation than the unobserved pairs of variants. This validation requires a robust prediction of the double mutant on the structural stability, however, current structural prediction methods are not capable of achieving such robustness. Second, we could check whether possible multi-locus linkage disequilibrium may have a structural basis with tightly packed groups of residues in physical space. While the absence of cluster of contacts may hint at contamination by haplotype effects, the presence of such clustering is not sufficient to dismiss the haplotype effects. Thus, while our method is useful for identifying instances of epistatic interactions, it is not sufficiently robust to known confounding factors such as haplotype effects.

Appendix

Appendix I

Local fitness landscape of green fluorescent protein
Sarkisyan K.S., Bolotin D.A., Meer M.V., Usmanova D.R., Mishin A.S., Sharonov G.V., Ivankov D.N., Bozhanova N.G., Baranov M.S., **Soylemez O**, Bogatyreva N.S., Vlasov P.K., Egorov E.S., Logacheva M.D., Kondrashov A.S., Chudakov D.M., Putintseva E.V., Mamedov I.Z., Tawfik D.S., Lukyanov K.A., Kondrashov F.A.

(Submitted for peer-review)

Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, Bogatyreva NS, Vlasov PK, Egorov ES, Logacheva MD, Kondrashov AS, Chudakov DM, Putintseva EV, Mamedov IZ, Tawfik DS, Lukyanov KA, Kondrashov FA. Local fitness landscape of the green fluorescent protein. Nature. 2016 May 11;533(7603):397-401. doi:10.1038/nature17995

Appendix II.

Structure and evolutionary history of a large family of NLR proteins in the zebrafish

Howe K., Schiffer, P.H., Zielinski, J., Wiehe, T., Laird, G.K., Marioni, J., **Soylemez O**, Kondrashov F.A., Leptin, M.

(Submitted for peer-review)

Pre-print available at: http://dx.doi.org/10.1101/022061

Howe K, Schiffer PH, Zielinski J, Wiehe T, Laird GK, Marioni JC, Soylemez O, Kondrashov F, Leptin M. Structure and evolutionary history of a large family of NLR proteins in the zebrafish. Open Biol. 2016 Apr;6(4):160009. doi: 10.1098/rsob.160009