3 MODIFIER EFFECTS BETWEEN REGULATORY AND PROTEIN-CODING VARIANTS

In this chapter I will:

- Outline how interactions between genetic variants have an impact on phenotypes.
- Explain why detecting interactions is challenging.
- Put forth a biological framework that can be used to test for interactions between regulatory (eQTLs) and protein-coding variants (nsSNPs) with an impact on gene expression.
- Demonstrate a modification effect in cis, arising from the eQTL-nsSNP interaction, that also has a trans effect on gene expression.
- Discuss the biological implications of this interaction.

3.1 CONTEXT-DEPENDENT EFFECTS ON PHENOTYPES: INTERACTIONS

To date, most association studies attempt to link single genetic variants to a specific phenotype (Brem, Yvert et al. 2002; Morley, Molony et al. 2004; Stranger, Forrest et al. 2005; Goring, Curran et al. 2007). Most of the systems that underlie cellular, developmental and physiological function however are composed of many elements that interact with one another, often in complex ways (Phillips 2008). As a result the extent to which a phenotype is shaped by genetic factors may not be a simple reflection of their independent effects, but is likely to arise in part from context-dependent effects, such as interactions between genetic factors, as well as interactions between genetic factors and the environment (Gibson 2008; Phillips 2008; Flint and Mackay 2009). The interaction between genetic variants that results in a phenotypic effect conditional on the combined presence of two or more variants is called epistasis (Brem, Storey et al.

2005; Nagel 2005). Epistasis may arise from a variety of underlying mechanisms. Over the years geneticists have used this term to describe subtly different genetic phenomena including the functional relationship between genes, the genetic ordering of regulatory pathways and the quantitative differences of allele-specific effects (Phillips 2008; Cordell 2009). Philips (2008) defines three forms of epistasis: functional epistasis (the molecular interactions that proteins and other genetic elements have with one another), compositional epistasis (the blocking of one allelic effect by an allele at another locus), and statistical epistasis (the average effect of substitution of alleles at combinations of loci, with respect to the average genetic background of the population). Hartl and Clark (2007) define epistasis as any situation in which the genetic effects of different loci that contribute to a phenotypic trait are not additive. In this thesis I refer to epistasis as a property of specific alleles at two loci whose interaction has an impact on gene expression, and will use the term interchangeably with the term interaction.

3.2 Prevalence and biological significance of interactions

The prevalence and biological significance of epistasis has always been an area of interest in the field of genetics, but its contribution to phenotypic variation has remained obscure, largely because genetic interactions have proven difficult to test (Musani, Shriner et al. 2007; Cordell 2009). This difficulty arises primarily because it is unclear which variant combinations should be tested and under which model of epistasis. To date, such an approach has been most feasible for specific genes or biological pathways that have been well-characterised, mostly in model organisms.

One of the best studied examples of epistasis is coat colour in mammals. In mice, an adaptive transition from dark to light coat colour accompanied the movement of dark-coloured forest mice from the forest to the beach (Steiner, Weber et al. 2007; Phillips 2008). The genetic basis for this transition stems from an interaction between

structural changes to the *agouti* locus and regulatory changes to the *Mc1r* locus. Obesity is another phenotype in mice that is affected by epistatic interactions and an extended network of epistatic QTLs has been discovered on chromosomes 4, 17, and 19 that controls regulation of fat pad depots and body weight (Stylianou, Korstanje et al. 2006).

A classic example of an interaction between regulatory and protein-coding variation is the *Adh* locus in *Drosophila* (Laurie, Bridgham et al. 1991; Stam and Laurie 1996). A series of regulatory SNPs in complex LD and with an impact on protein concentration, modify the effects of a protein-coding variant affecting the catalytic efficiency of this enzyme. Catalytic efficiency and protein levels determine overall enzyme activity. This example illustrates that large effects attributed to a single locus may arise as a consequence of multiple associated interacting variants and is a case of a modification effect in cis where the protein-coding effect is magnified or masked through the action of regulatory variants. More recent studies in *Drosophila* reveal epistatic effects between genes affecting traits such as ovariole number (Orgogozo, Broman et al. 2006) and olfactory avoidance (Sambandan, Yamamoto et al. 2006).

In cases where little is known about the genes sculpting a phenotype, addressing the possibility of epistasis becomes more challenging. A recent study interrogating cardiac dysfunction in *Drosophila* (Ocorr, Crawley et al. 2007) identified a major susceptibility locus for this trait, but highlighted the importance of examining the phenotype in different genetic backgrounds to detect variants whose effects are manifest through interactions with the prime susceptibility locus. The extent of epistasis in a more global way has been demonstrated in yeast where experiments on gene expression revealed that interacting locus pairs are involved in the inheritance of over half of all transcripts (Brem, Storey et al. 2005; Boone, Bussey et al. 2007). Furthermore, a large proportion of the eQTLs attributable to interaction effects were not detected by single locus tests. This suggests that analysis of interaction effects in other systems is likely to uncover additional associations.

In humans, most documented cases of epistasis have been detected in instances where there are biological clues as to which genes should be tested. Epistasis between two multiple sclerosis (MS) associated human leukocyte antigen (HLA) alleles was demonstrated by Gregerson et al. (2006) who showed that one allele modifies the T-cell response that is activated by a second allele, through activation-induced apoptosis contributing to a milder form of MS-like disease. Similarly, Oprea et al. (2008) demonstrated that a specific modifier effect is protective against spinal muscular atrophy (SMA). SMA arises from a homozygous deletion of the *SMN1* gene, but some deletion homozygotes escape the disease phenotype due to the modulating effects of expression of PLS2.

Risk for nicotine dependence and lung cancer was shown to be sculpted by interactions between functional variants in genes belonging to the neuronal nicotinic acetyl choline receptor (nAChR) family (Wang, Cruchaga et al. 2009). nAChR genes encode pentameric ligand-gated ion channels that mediate fast signal transmission at synapses and modulate the release of neurotransmitters. Nicotine is an exogenous agonist of these receptors, and variations in nAChR genes are strong candidate risk factors for nicotine dependence and lung cancer. The authors of this study showed that interactions between a coding variant, that changes amino acid sequence in the α5 nicotine receptor subunit gene *CHRNA5* (D398N), and non-coding variants that regulate the gene's expression levels confer risk for nicotine dependence and lung cancer. They conclude by stating that by establishing this cis modification effect they have identified a potential drug target.

With the explosion of successful GWAS over the past three years, the natural next step is genome-wide interaction testing (Cordell 2009). Detecting epistasis is crucial as it is likely to uncover new variants affecting phenotypes. Additionally, epistasis may mask the genetic impact of variants and impede replication of primary associations. Differential fixation of variants that modulate the primary disease variant can therefore

affect the degree of penetrance of disease alleles and the need to address this property of genes in a systematic, genome-wide approach is becoming increasingly pressing. The case of MS clearly illustrates this: as with most complex disorders, MS has a polygenic heritable component characterised by underlying complex genetic architecture (Oksenberg, Baranzini et al. 2008). Association studies to date have met with modest success in identifying MS-causing genes, and a large proportion of phenotypic variation remains unexplained. The expectation is that this residual variation arises at least in part, as a consequence of gene-gene interactions.

In this study I explored the extent to which regulatory variants modify protein-coding effects in cis and tested whether this modification effect has an impact on gene expression of other genes in the genome in a trans effect. This work has been described in (Dimas, Stranger et al. 2008).

3.3 BIOLOGICAL FRAMEWORK TO DETECT INTERACTIONS

Most strategies that address the effects of epistasis in humans involve millions of agnostic pairwise tests falling into one of two broad categories: exhaustive testing of interactions between all pairs of variants across the genome (Marchini, Donnelly et al. 2005), or testing of interactions between all pairs of variants with an independent main effect on the phenotype (Marchini, Donnelly et al. 2005; Evans, Marchini et al. 2006; Dixon, Liang et al. 2007). It is not entirely clear whether improvements in statistical methods will be sufficient to address the problem of epistasis. Therefore the development of realistic biological models of epistatic interactions may reduce the statistical cost of dealing with many comparisons and facilitate the development of such methodologies.

In this study I present a biological framework for global survey of interaction effects in humans, which avoids exhaustive testing of agnostic pairs and involves

prioritisation of variants to be tested. Two types of functional variants are common throughout the human genome and are present at appreciable frequencies in populations: regulatory variants with an impact on the expression patterns and levels of genes (Pastinen and Hudson 2004; Birney, Stamatoyannopoulos et al. 2007; Forton, Udalova et al. 2007; Spielman, Bastone et al. 2007; Stranger, Nica et al. 2007) and protein-coding variants affecting protein sequence (Rodriguez-Trelles, Tarrio et al. 2003; Birney, Stamatoyannopoulos et al. 2007). To date, the effects of these variants have been considered independently of each other. In this study I evaluated the joint effects of regulatory and protein-coding variants on genome-wide expression phenotypes in humans to highlight an underappreciated angle of functional variation.

As outlined in section 2.6.1 the proposed model brings together quantitative and qualitative variation, by testing the cis and trans impact on gene expression observed when a gene with an identified regulatory variant (eQTL) also contains protein-coding variation (nsSNP). Under such a scenario, and assuming that mRNA levels are indicative of mature protein levels, the resulting protein products will differ in quantity (expression level) and quality (amino acid sequence) among individuals (Figure 9). Depending on the historical rate of recombination between eQTLs and nsSNPs, different allelic combinations (haplotypes) can arise on the two homologous chromosomes in a population (Figure 14). As a consequence, phasing (the arrangement of alleles at each variant position with respect to one another) can differ between individuals in the population. Such an interaction results in a modification (magnification or masking) of the functional impact of the protein-coding variant. If the modified gene product has downstream targets, then expression of these target genes may also be affected in a trans manner.

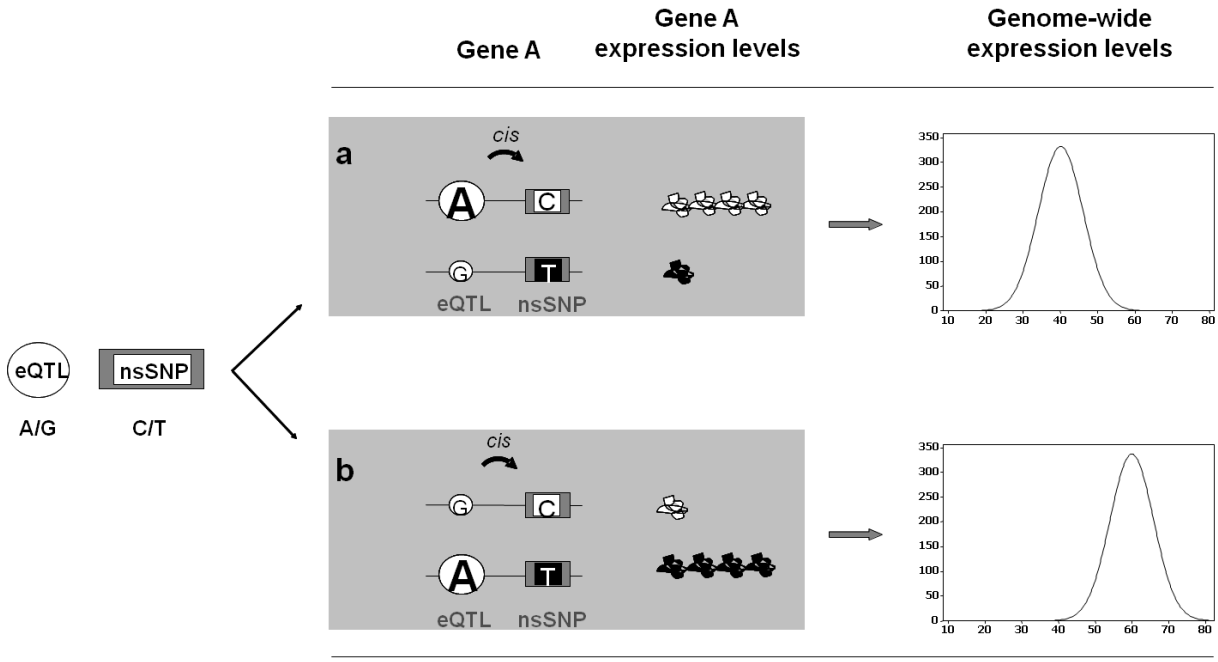

Figure 14. Illustration of a hypothetical epistatic interaction between a regulatory (eQTL) and a protein-coding variant (nsSNP). Two double heterozygote individuals may be genotypically identical, but the phasing of alleles can be different and may result in very distinct phenotypes between individuals. In **a)** the A allele of the eQTL drives high expression levels of the protein arising from the C allele of the nsSNP. In **b)** the G allele of the eQTL drives low expression levels of the protein arising from the C allele of the nsSNP. If the protein-coding variant is functionally important then this interaction in cis can give rise to different means in the distribution of a complex trait phenotype (e.g. genome-wide expression levels) as shown on the right (trans effect).

3.4 MODIFICATION EFFECT IN CIS: DIFFERENTIALLY EXPRESSED NSSNPS

Using this model as a main principle, I explored the degree to which nsSNPs can be modulated by cis eQTLs. eQTLs were identified in a previous study (Stranger, Nica et al. 2007) in LCLs of the unrelated individuals of the Phase 2 HapMap populations (60 CEU, 45 CHB, 45 JPT and 60 YRI) (Table 4). LCLs represent one particular cell type and even though there may be some effect arising from EBV transformation, it has been demonstrated that genetic effects on gene expression, such as the ones I describe below,

are readily identifiable, mappable, and replicate in independent population samples generated decades apart (Dimas, Deutsch et al. 2009).

	0.01 permutation threshold			
Population	LR: significant genes	CEU-CHB-JPT-YRI multi-population	CEU-CHB-JPT multi-population	CHB-JPT multi-population
CEU	606	1,186	1,149	1,071
CHB	634	1,186	1,149	1,071
JPT	679	1,186	1,149	1,071
YRI	742	1,186	1,149	1,071
Non-redundant union	1,746			
4 populations	114			
≥2 populations	533			

Table 4. eQTLs detected in the HapMap Phase 2 populations (0.01 permutation threshold). Adapted from (Stranger, Nica et al. 2007).

Two strategies were applied to detect DE nsSNPs. The first strategy involved scanning genes with known cis eQTLs (Stranger, Nica et al. 2007), for nsSNPs. The aim was to identify nsSNPs that are predicted to be DE as a consequence of a nearby regulatory variant tagged by the eQTL. I identified 606, 634, 679 and 742 genes with at least one eQTL at the 0.01 permutation threshold (estimated FDR of 20%) (Table 4). Of these genes 159, 168, 180 and 202 (union of 484) were found to contain 286, 304, 311 and 393 nsSNPs respectively (union of 909) (Table 5). I infer that these nsSNPs are DE as they reside in genes with experimentally-derived varying expression levels. This means that there are allelic effects on gene expression such that, depending on the genotypes of the eQTL and nsSNP and on the phasing of their alleles, one can make predictions about the relative abundance of the two alleles of a transcript in the cell.

a

	nsSNPs					
Population	Total nsSNPs interrogated	With identified eQTL	Single population association	Single population DE	Multi-population association	Total DE
CEU	5,686	286	242	452		
CHB	5,335	304	276	478		
JPT	5,328	311	267	487		
YRI	6,093	393	255	574		
Non-redundant union	8,233	909	703	1,355	587	1,502

b

	genes					
Population	Total genes interrogated	With identified eQTL	Single population association	Single population DE	Multi-population association	Total DE
CEU	3,579	159	196	307		
CHB	3,412	168	226	322		
JPT	3,410	180	210	325		
YRI	3,692	202	211	364		
Non-redundant union	4,518	484	560	863	461	973

Table 5. a) nsSNPs and b) genes interrogated for differential expression. (DE: differentially expressed)

The second strategy for DE nsSNP discovery involved direct association testing (using LR) between nsSNP genotype and expression levels of the gene in which the nsSNP resides. This strategy aimed to identify DE nsSNPs that are in LD with a regulatory variant that drives expression levels. Depending on the strength of the regulatory effect, such variants may or may not have been detected in the initial scan for eQTLs (Stranger, Nica et al. 2007). Relative distances between eQTLs and nsSNPs can vary, but in the special case where this distance is short in genetic terms, the two variants may be in LD (McVean, Spencer et al. 2005). Under these circumstances it is expected that the nsSNP itself will demonstrate some degree of association with expression levels of the gene in which it resides. I tested for genotype-expression associations in each population separately and in three multiple population sample panels (see section 2.6.3).

For the single-populations analysis, with significance evaluated at the 0.01 permutation threshold, 56 nsSNPs and 34 genes are expected to have at least one significant association by chance. I detected 242, 276, 267 and 255 nsSNPs (union of 703; estimated FDR of 21%) with significant for the CEU, CHB, JPT and YRI populations respectively (Table 6 a). These associated nsSNPs correspond to 196, 226, 210 and 211 genes (union of 560; estimated FDR of 16%) (Table 6 b). For the multiple-population analysis I detected 345, 362 and 417 nsSNPs (estimated FDR of 15%) for the four, three and two population groups respectively (Table 6 a), corresponding to 284, 296 and 320 significant genes (estimated FDR of 11%) (Table 6 b). Overall, the multiple-population analysis yielded a total of 587 nsSNPs with significant associations, corresponding to 461 genes. Taken together, the association analyses indicate that 884 nsSNPs (688 genes) across the four populations are associated with expression levels of the genes they are in, suggesting that they are in LD with regulatory variants driving their expression. In this specific case of association, the nsSNP itself serves as a proxy for the regulatory variant and knowledge of associated nsSNP genotype for an individual provides a prediction of relative abundance of the two transcript alleles.

a

	0.01 permutation threshold						
	1	2	3	4			
Population	significant nsSNPs	CEU-CHB-JPT-YRI multipop	CEU-CHB-JPT multipop	CHB-JPT multipop	Overlap 1&2	Overlap 1&3	Overlap 1&4
CEU	242	345	362	417	111	139	104
CHB	276	345	362	417	126	162	224
JPT	267	345	362	417	136	161	203
YRI	255	345	362	417	102	86	90
Nonredundant	703						
4 populations	34						
≥2 populations	233						

b

	0.01 permutation threshold						
	1	2	3	4			
Population	significant genes	CEU-CHB-JPT-YRI multipop	CEU-CHB-JPT multipop	CHB-JPT multipop	Overlap 1&2	Overlap 1&3	Overlap 1&4
CEU	196	284	296	320	99	117	87
CHB	226	284	296	320	109	129	183
JPT	210	284	296	320	114	125	156
YRI	211	284	296	320	87	77	82
Nonredundant	560						
4 populations	31						
≥2 populations	196						

Table 6. a) nsSNP and b) gene cis associations detected in single and multiple populations.

To summarize, two classes of DE nsSNPs were discovered: a) 909 nsSNPs mapping in genes with a previously identified eQTL (considering nsSNPs of all frequencies) and b) 884 nsSNPs showing a significant association with expression levels of the gene they are in (considering nsSNPs with MAF ≥ 0.05) (Figure 15). From a non-redundant total of 8,233 nsSNPs tested in four populations, 1,502 of these (~18.2%) are predicted to be DE. It is a plausible biological hypothesis that mature protein levels mirror transcript levels on average and as a consequence, this high fraction of DE nsSNPs may have important implications for levels of protein diversity in the cell.

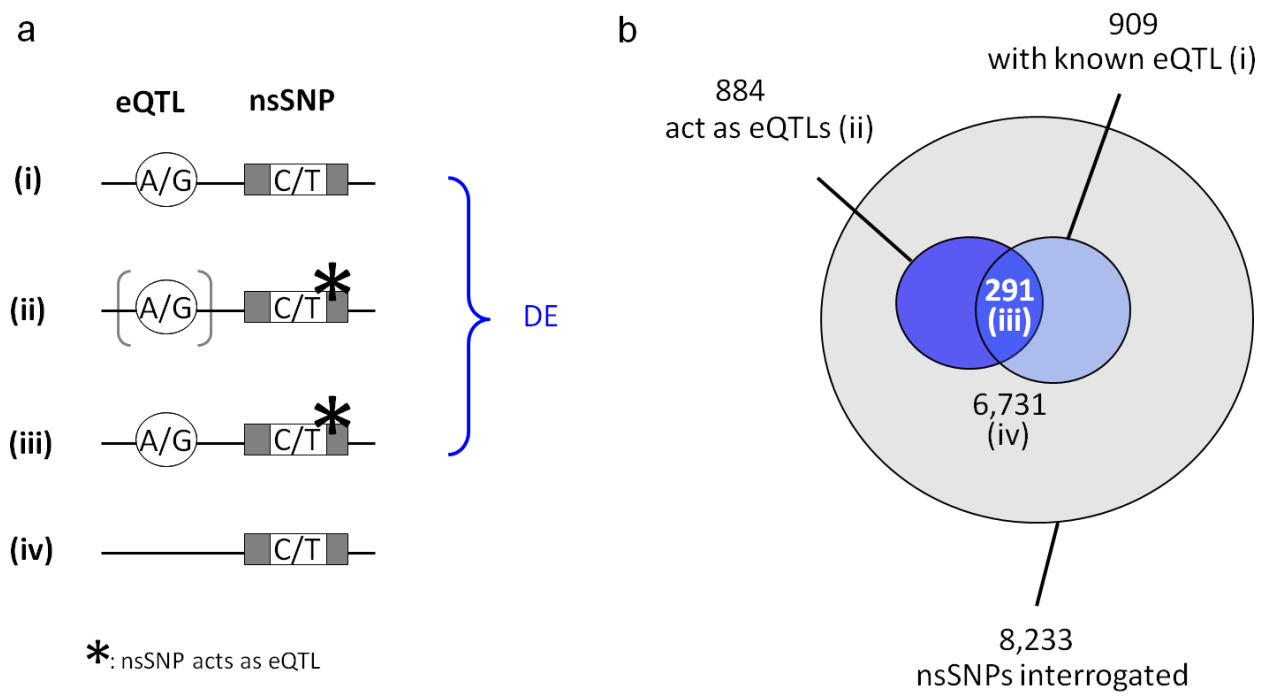

Figure 15. Strategies applied to discover differentially expressed (DE) nsSNPs. a) Two approaches were employed to discover DE nsSNPs: nsSNPs mapping in genes with a known eQTL (i) and nsSNPs that were associated with expression levels of the gene they map in (ii). In (ii) the presence of a cis-acting regulatory variant is implied. For some nsSNPs with a significant association, an identified cis eQTLs also exists (iii). In all other cases the nsSNPs interrogated were not inferred to be to be DE (iv). **b)** Of the 8,233 nsSNPs studied, 909 mapped in a gene with an identified eQTL (i), 884 were found to be associated with levels of expression of the gene they reside in (ii), 291 nsSNPs with an identified eQTL also showed a significant association with expression levels (iii) and 6,731 nsSNPs showed no evidence of differential expression (iv). Taken together over 18% of nsSNPs were found to be DE.

3.4.1 Linkage disequilibrium between eQTLs and nsSNPs

Of the 884 DE nsSNPs detected through association testing, only 291 also possess a previously identified eQTL. This suggests that eQTL detection in our previous study was conservative and that nsSNPs can act as tags of undiscovered regulatory variants. With this in mind, it is expected that LD between eQTL-nsSNP pairs in which the nsSNP had a significant association with gene expression, will be greater than LD between eQTL-nsSNP pairs in which the nsSNP was not associated. To explore this, I

used data from the single population analysis, and compared the distribution of r^2 values between the two eQTL-nsSNP pair types. As expected, much higher LD was found for eQTL-nsSNP pairs where the nsSNP showed a significant association (M-W p-value < 0.0001) (Figure 16). This confirms that in most cases, association of the nsSNP with its gene's expression is due to a regulatory variant tagged by the eQTL.

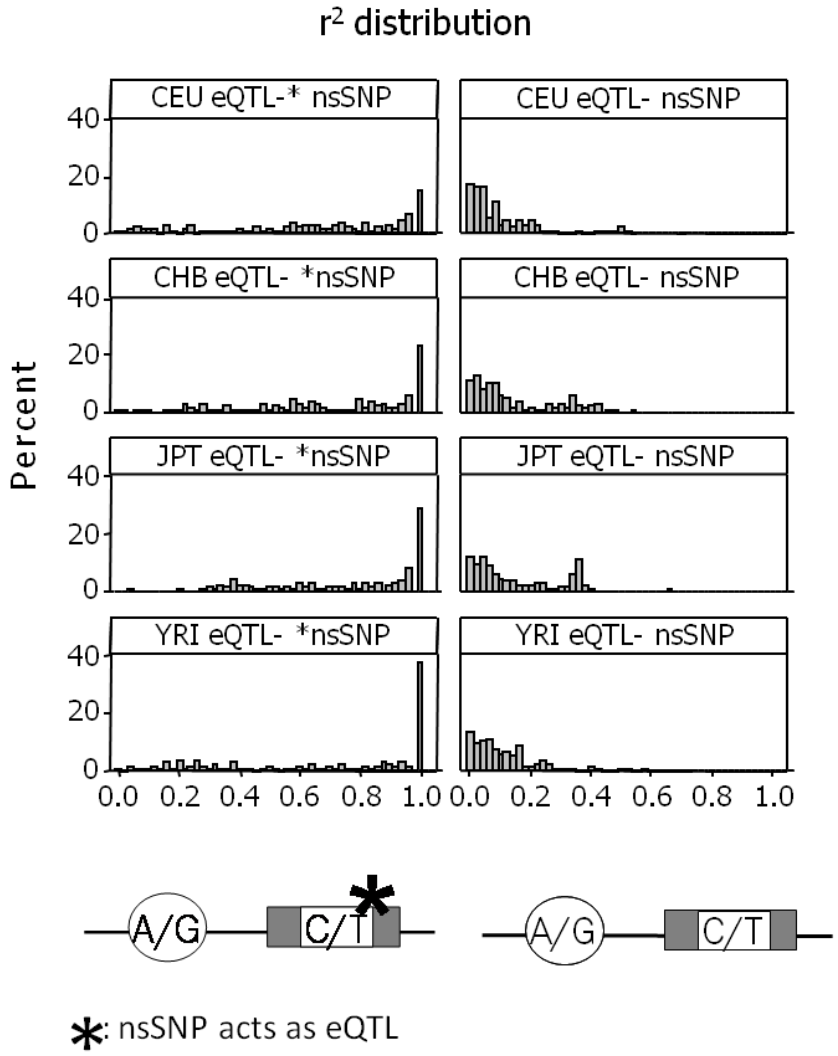

Figure 16. Linkage disequilibrium (LD) properties of eQTL-nsSNP pairs. The distribution of r^2 (a measure of LD) was compared between eQTL-nsSNP pairs in which the nsSNP acts as an eQTL (i.e. showed a significant association with its gene's expression levels) and SNP pairs in which the nsSNP was not associated. As expected, r^2 values are much higher in the first case, where the nsSNP is thought to act as a tag of the functional regulatory variant nearby.

3.4.2 Experimental verification of differentially expressed nsSNPs

Thus far I have described relative abundance estimates for transcripts of genes containing nsSNPs using genotypic associations. To verify the statistical predictions of nsSNP association tests, it was necessary to perform direct allele-specific quantification. A subset of nsSNPs were tested for ASE (Pastinen, Ge et al. 2006; Forton, Udalova et al. 2007) in heterozygote CEU and YRI individuals. The initial experiment included a total of 141 nsSNPs predicted to be DE, but the assay performed was new and proved noisy. As a result it was possible to confirm and analyse signals for 28 nsSNPs, after filtering for association $r^2 > 0.27$ and ASE mean RNA intensity > 12. For heterozygous individuals at each nsSNP, I assigned relative expression of the two alleles and subsequently compared the experimentally derived relative abundance (ASE results) with the predictions of relative abundance from the genotypic association test. Predicted and experimentally-quantified relative expression of nsSNP alleles were in agreement for 89% (16 out of 18) and 90% (9 out of 10) of nsSNPs tested in the CEU (Figure 17 a) and the YRI populations (Figure 17 b) respectively. This is in agreement with the estimated FDR and suggests strongly that the relative abundance of alternative coding transcripts can be inferred reliably by genotypic associations.

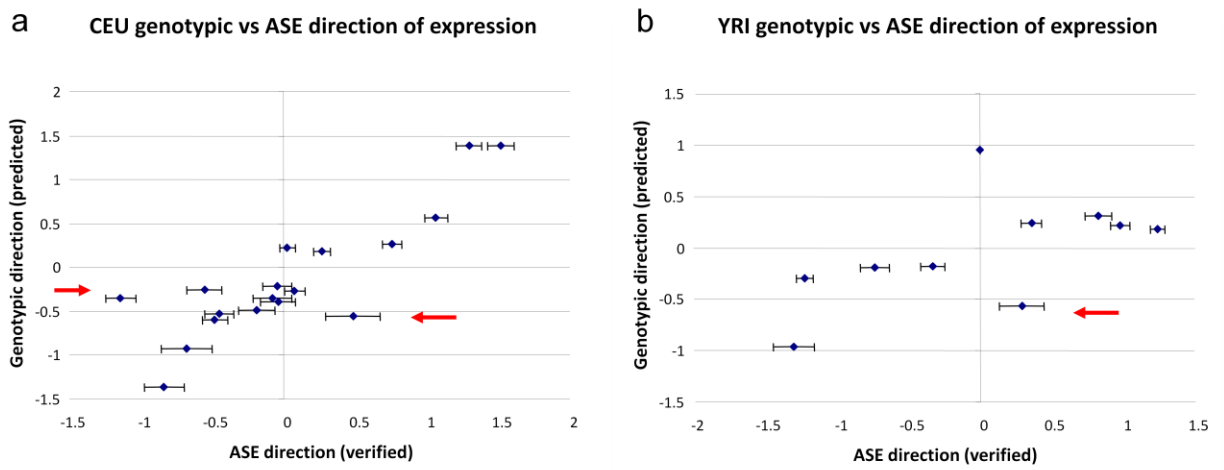

Figure 17. Comparison of statistically predicted and experimentally verified direction of nsSNP allelic effects. The predictions of the nsSNP association test were in agreement with the experimentally verified direction of expression in **a)** 89% and **b)** 90% of the cases studied in the CEU and YRI populations respectively. Red arrows point to the cases where association predictions did not agree with allele-specific expression (ASE) results.

3.4.3 Properties of differentially expressed nsSNPs

To assess the potential biological impact of DE nsSNPs I compared three functional attributes of amino acid substitutions arising from DE nsSNPs and non-DE nsSNPs (testing nsSNPs with MAF ≥ 0.05, to assess common nsSNP consequences). I investigated: 1) the relative position of substitution on the peptide, as different effects may arise depending on whether the nsSNP is at the beginning or the end of the peptide (Figure 18 a), 2) the resulting change in peptide hydrophobicity which may alter the interactions of a protein (Kyte and Doolittle 1982) (Figure 18 b) and 3) the resulting change in Pfam score (a measure of amino acid profile in each position of a protein domain) (Finn, Tate et al. 2008), which assesses the integrity of protein domains that are evolutionary conserved and likely to harbour important functions (Figure 18 c). In all cases the properties of DE nsSNPs were not different from those of non-DE nsSNPs (M-W p-value ≥ 0.05). Though indirect and not comprehensive, this finding

suggests that DE nsSNPs may be a random subset of nsSNPs. If these variants have a functional impact, this will be modified (magnified or masked) by the regulatory variant tagged by the eQTL.

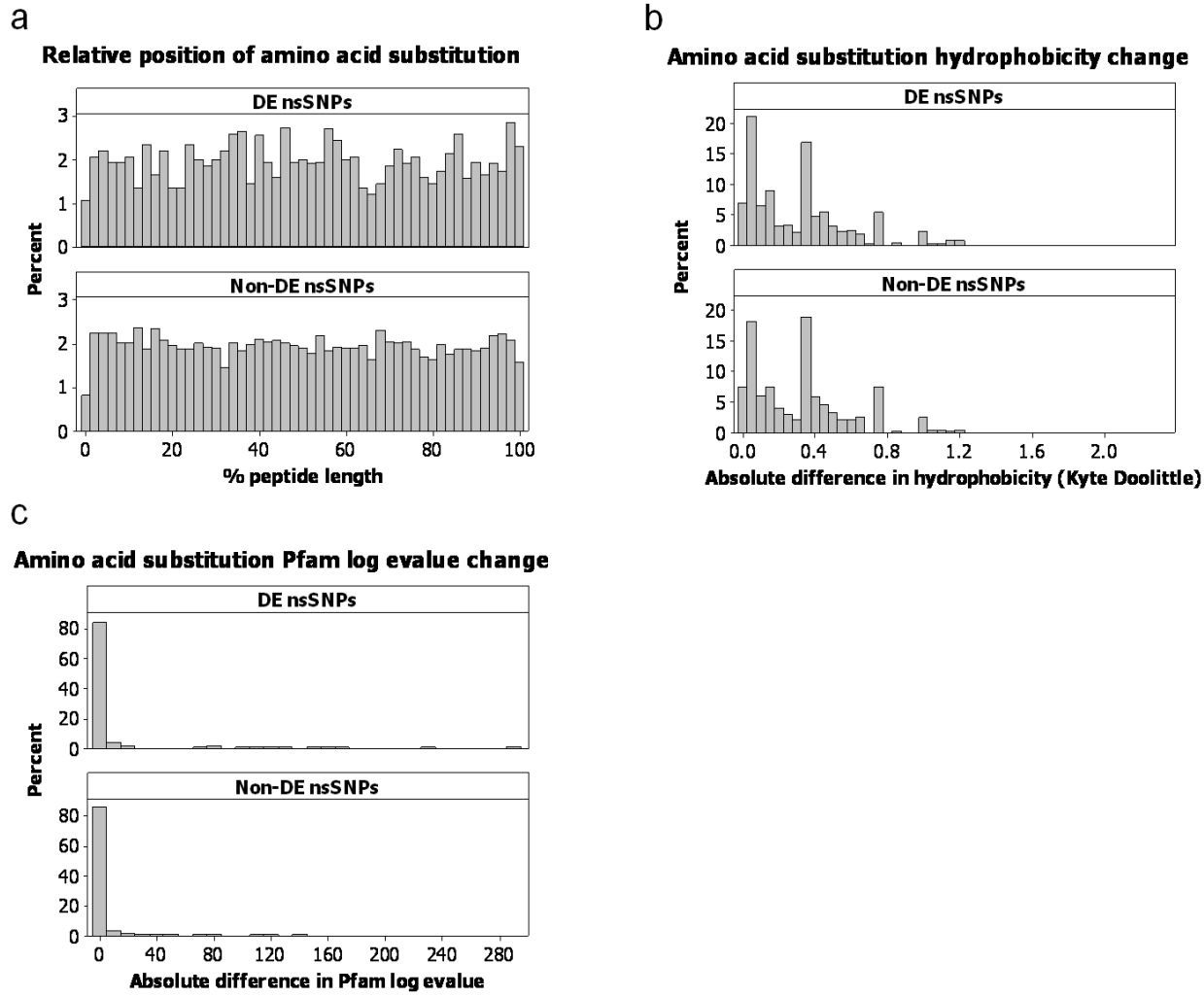

Figure 18. Comparison of biological properties of differentially expressed (DE) vs. non-DE nsSNPs. Three functional attributes of the amino acid substitutions resulting from DE nsSNPs vs. non-DE nsSNPs were compared: **a)** relative position of substitution on the peptide, **b)** resulting change in peptide hydrophobicity and **c)** resulting change in Pfam score when searched against the Pfam profile Hidden Markov Model library. In all cases Mann-Whitney (M-W) tests did not reveal a significant difference between DE and non-DE nsSNPs (M-W p-value ≥ 0.05) and DE nsSNPs appear to be a random subset of nsSNPs. Therefore, if a random nsSNP has a phenotypic effect, this is likely to be magnified or masked through differential expression driven by cis-acting regulatory variants.

To assess how many DE nsSNPs have a known function, I explored the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim/) and found that 71 (out of 1,502) DE nsSNPs had an OMIM entry (OMIM nsSNPs, the genes they map in and the predicted health impact are shown in the Appendix). DE nsSNPs were found to map in genes with a role in cancer susceptibility (*BRAC1* (+113705), *BARD1* (+113705)), asthma and obesity (*ADRB2* (+109690)), Crohn disease (CD) (*DLG5* (*604090)), myokymia (*KCNA1* (*176260)), diabetes (*OAS1* (*164350)), chronic lymphatic leukaemia (*P2RX7* (*602566)) emphysema and liver disease (*PI* (+107400)), severe keratoderma (*DSP* (+125647)), and familial hypercholesterolemia (*ABCA1* (+600046)). In some cases the functional role of the nsSNP is unclear and the noise in reported functional effects in OMIM is well-known and difficult to assess in a study such as the present. However there are examples where specific effects have been attributed to nsSNPs. For example, rs28931610 in *DSP* is predicted to change disulphide bonding patterns and alter the peptide tertiary structure, rs28933383 in *KCNA1* causes a substitution in a highly conserved position of the potassium channel and is predicted to impair neuronal repolarization, rs28937574 in *P2RX7* is a loss of function mutation associated with chronic lymphatic leukaemia, rs28931572 in *PI* entails a replacement of a polar for a non-polar amino acid and is predicted to disrupt tertiary structure of the protein, and rs2230806 in *ABCA1* is associated with protection against coronary heart disease in familial hypercholesterolemia. The modulation of such strong effects by cis regulatory variation may increase the complexity and severity of the biological impact.

3.5 eQTL-nsSNP epistatic effect in trans

Thus far I have presented evidence for a modification effect in cis. In cases where the gene containing the DE nsSNP has downstream targets, then it is likely that the expression of target genes is also affected. The aim of this analysis was to test for the

genome-wide effects of this interaction directly, in a statistical framework. To do this I carried out ANOVA to test the main effects of eQTLs and nsSNPs as well as their interaction term (eQTL x nsSNP) on genome-wide gene expression. The rationale behind this approach is that if an eQTL-nsSNP interaction is biologically relevant, its effect may influence gene expression in trans. The power to detect an interaction is maximized when all combinations of genotypes are present, each at appreciable frequencies in the population. To increase power of interaction detection, rare homozygotes were pooled with heterozygotes into a single genotypic category, creating a 2x2 table of genotypes (section 2.6.7). This does not introduce bias in the test statistic as shown by permutations below. Analyses were performed for the CEU population as CHB and JPT population samples were small (45 individuals) and YRI have shown low levels of trans effects in previous studies (Stranger, Nica et al. 2007). I tested 22 eQTL-nsSNP pairs with low LD ($D' \leq 0.5$) and a MAF ≥ 0.1 for both SNPs, against genome-wide expression. At the 0.001 nominal p-value threshold, roughly 331 significant associations are expected (assuming a uniform distribution of p-values) for the interaction term. I detected 412, which corresponds to an estimated FDR of 80%. This is an overall weak signal, but the signals at the tail of the distribution appear to be real given the limited power of this analysis (Figure 19 a).

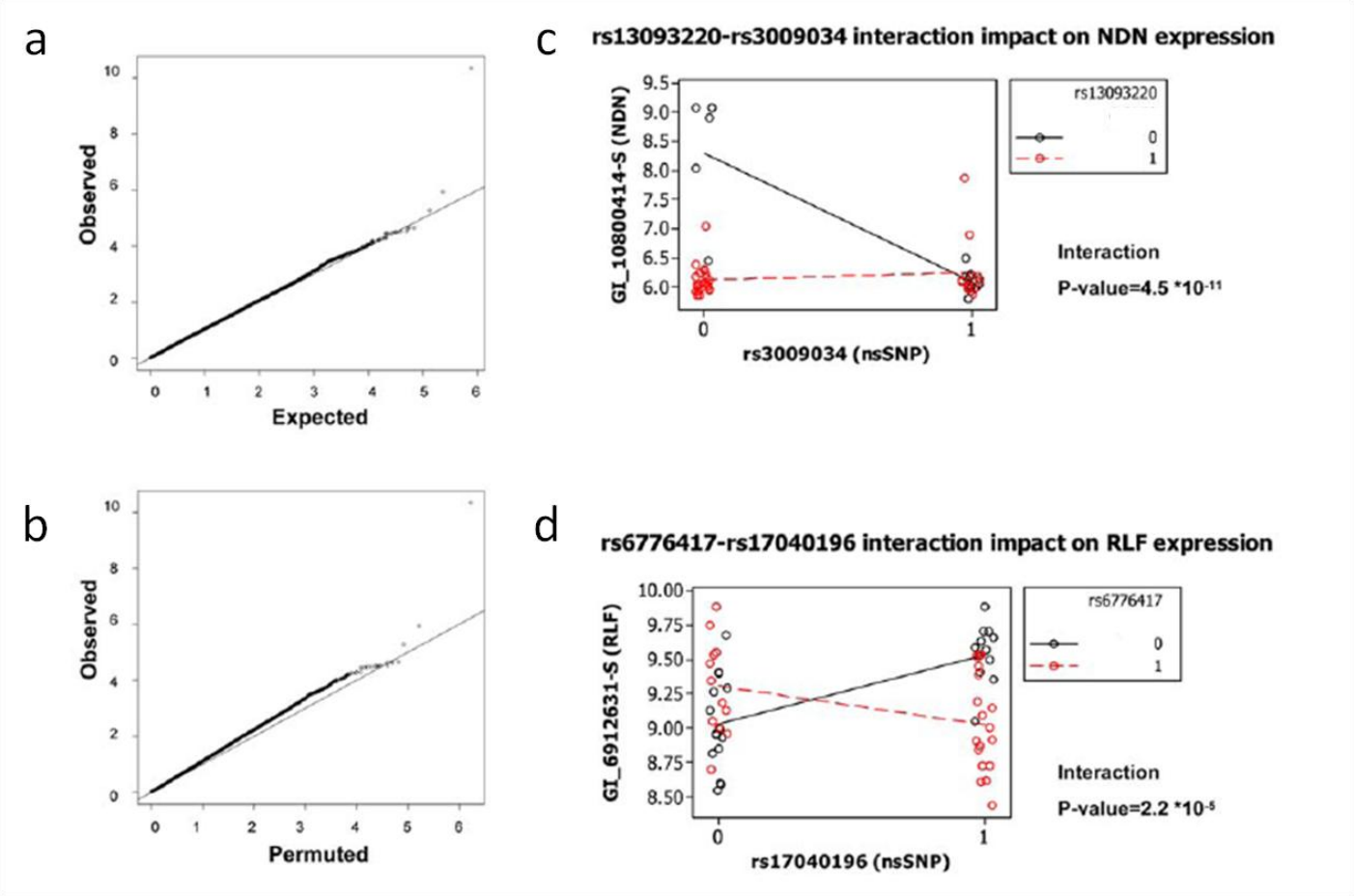

Figure 19. Impact of eQTL-nsSNP genetic interaction on trans gene expression. a) QQ plot of observed vs. expected $-\log_{10}$ p-values of the interaction term from analysis of variance (ANOVA) under the assumption of a uniform distribution of expected p-values. **b)** QQ plot of observed vs. permuted $-\log_{10}$ p-values of the interaction term from ANOVA. **c)** The interaction between rs13093220 (eQTL) and rs3009034 (nsSNP) on chromosome 3 is associated with changes in expression of gene *NDN* (probe ID GI_10800414-S) on chromosome 15 (interaction p-value = $4.5*10^{-11}$). **d)** The interaction between rs6776417 (eQTL) rs17040196 (nsSNP) on chromosome 3 is associated with changes in expression of gene *RLF* (probe ID GI_6912631-S) on chromosome 1 (interaction p-value = $2.2*10^{-5}$).

To test for potential biases in the statistic used, I carried out the same tests using permuted gene expression values (a single permutation was performed by maintaining the correlated structure of gene expression data, see section 2.6.7) relative to the eQTL-nsSNP genotypes. I explored the p-value distribution of the eQTL-nsSNP interaction for observed and permuted data (Figure 19 b) and found an abundance of low p-values in

the observed data. There appears to be some degree of p-value inflation in the observed data relative to the permuted data which is most likely due to correlations in gene expression values. However this does not affect the enrichment of p-values seen at the tails of the observed distribution relative to distributions from expected and permuted values. The observed results therefore show enrichment relative to a uniform distribution of p-values (permutation was not performed to assess significance thresholds, but to assess enrichment of tests with low p-values in the observed data). To further evaluate the robustness of the interactions, I repeated the analysis for the top ten eQTL-nsSNP significant pairs against their corresponding trans-associated gene expression phenotype, after permuting eQTL genotypes relative to nsSNP genotypes and gene expression values. As expected, the significance of the interaction term vanishes in the permuted data. The conditional effects of alleles at the eQTL and nsSNP loci can therefore have a very different impact on the expression of other genes in the cell. This conditional effect on gene expression is illustrated in Figure 19 c and Figure 19 d which show two examples of eQTL-nsSNP interactions (interaction term p-values = 4.5×10^{-11} and 2.2×10^{-5} respectively). In Figure 19 c rs3009034 has an effect on gene expression of gene *NDN* only if the genotype of rs13093220 is homozygous for the common allele. The phenotypic effect of such interactions is even more prominent in Figure 19 d where opposite directions of the effect of rs1704196 are observed. Table 7 shows summary statistics and specific information of SNPs and genes for the ten most significant interactions with a trans effect.

eQTL	nsSNP	source chr	eQTL loc	eQTL MAF	nsSNP loc	nsSNP MAF	trans gene Hugo	trans chr	eQTL p-value	nsSNP p-value	Interaction p-value
rs13093220	rs3009034	3	75853445	0.453	75864268	0.242	NDN	15	1.04E-04	3.41E-05	4.56E-11
rs2280902	rs2272720	8	2064479	0.433	2008828	0.467	TFF2	21	1.88E-02	3.35E-03	1.20E-06
rs2280902	rs2272720	8	2064479	0.433	2008828	0.467	Hs.525661	15	1.03E-01	1.16E-02	5.41E-06
rs6776417	rs17040196	3	14674018	0.267	14720861	0.35	RLF	1	1.46E-01	2.89E-01	2.22E-05
rs6776417	rs6790129	3	14674018	0.267	14730621	0.342	RLF	1	1.46E-01	2.89E-01	2.22E-05
rs6776417	rs17040196	3	14674018	0.267	14720861	0.35	C6orf57	6	3.40E-01	8.91E-02	2.36E-05
rs6776417	rs6790129	3	14674018	0.267	14730621	0.342	C6orf57	6	3.40E-01	8.91E-02	2.36E-05
rs6776417	rs17040196	3	14674018	0.267	14720861	0.35	Hs.121413	2	8.84E-01	8.96E-01	3.10E-05
rs6776417	rs6790129	3	14674018	0.267	14730621	0.342	Hs.121413	2	8.84E-01	8.96E-01	3.10E-05
rs2280902	rs2272720	8	2064479	0.433	2008828	0.467	hmm11130	2	3.51E-01	6.33E-02	3.17E-05

Table 7. eQTL-nsSNP pairs with the most significant interaction effects in trans. Summary statistics and information about mapping location as well as source and trans-affected genes are shown. (chr: chromosome, loc: location, MAF: minor allele frequency)

3.6 Conclusions

I have presented a biological framework to interrogate functional genetic variation by focusing on a specific case of epistasis between regulatory and protein-coding variants. I demonstrated that regulatory variants may have an impact on the protein diversity of cells by differentially modulating the expression of protein-coding variants. In cis, regulatory variants can amplify or mask the functional effects of protein-coding variants. If the coding variant has a role in disease, such an interaction is likely to result in a milder or more severe phenotype to the one expected if only the protein-coding variant were present. Cis interactions were also shown to affect the expression of other genes in the cell in a trans effect, revealed only if an interaction between variants is specifically tested for.

The conditional and context-dependent effects of alleles of variants are likely to have important consequences for complex and quantitative phenotypic traits (Flint and Mackay 2009). In this study I put forth a biological framework for considering and conditioning existing disease associations on known regulatory and protein-coding

variants, in an approach that also provides a potential explanation for the differential penetrance of known disease variants. The abundance of cis regulatory and protein-coding variants in human populations and the generic nature of this type of epistatic interaction (no assumptions made about specific biological pathways) makes it likely that such interactions are common genetic factors underlying complex traits and their consideration is likely to reveal important associations that have not been detected to date. Furthermore, this consideration is particularly important for studies that fail to replicate primary disease associations in newly tested populations, since some of the failures may be due to differential frequency of modifier alleles between the first and second population. Consideration of such interactions may assist in better interpretation of non-replicated signals.