Chapter 5

Conclusion

5.1 Overview and limitations

The main motivation for the work in this thesis was to find evidence for how or if non-linear encoding of genetic information contributes to phenotypic variance. After applying quality control measures and establishing the additive association baselines in Chapter 2, I searched for statistical epistasis using standard statistical methods and NN approaches in Chapter 3 and Chapter 4, respectively. Like the two parallel efforts that were comparable in scope to my work (Bellot et al., 2018; Xu et al., 2020), neither the standard nor the NN approaches produced evidence for contributions of statistical epistasis to phenotypic variance.

I believe that the greatest practical limitation of my work was that I restricted myself to only perform recombination block level tests, which precluded the possibility of detecting interactions within blocks. I thought that this was necessary due to the potential for a perfect overlap between genuine statistical epistasis and haplotype effects to exist (Wood et al., 2014). Such haplotype effects are a physical property of the DNA molecule, whereas I was interested in non-linear effects that describe information encoding. However, taking this highly conservative approach meant that I no longer had the ability to identify (the biologically potentially more plausible) local interactions, and this may have contributed to the overall negative results of my analyses. In the future, once WGS data becomes standard, large-scale fine mapping databases (such as the CausalDB (Wang et al., 2019)) and methods that could handle multiple causal signals (Wang et al., 2020) become more widely used, interaction tests that involve fine mapped causal loci may be performed without the danger to be mistaken for pure haplotype effects.

5.2 Reflections on non-linear genetic effects

Current estimates of the fraction of the human genome that is truly functional range from 8.2% to 80% (Dunham et al., 2012; Rands et al., 2014). All functional areas of the genome would be expected to work in concert to produce a genome-wide phenotype, which would then arise as an emergent property from the activity of all active parts of the base sequence. From this perspective, non-linearity appears to be an inevitable property that arises from the compression of information to produce complex biological systems.

When I began my work, I started with the very sensible, although now I what believe to be erroneous, intuition that to achieve the aforementioned non-linear encoding of genetic information, the process that generates or maintains trait variance should also be non-linear. There were numerous supporters of this view who made plausible arguments based on either simulations (Carter et al., 2005) and theoretical grounds (Mackay and Moore, 2014) or by citing examples from model organisms (Hansen, 2013). However, after reflecting on my findings, and revisiting some of the same literature that shaped my initial views, I have now come to believe that my initial views were misguided.

After working with and developing methods for real biological data for several years, I find simulations and theoretical arguments less convincing than before, as biology is a science of what is, rather than what could be. Arguments based on how the apparent additive profile of traits could also be explained by alternatively parameterised models that involve epistasis, such as those made by Huang and Mackay (2016) that I covered in the Introduction under section 1.1.5.1, now leave me unimpressed, as these theories cannot be proved or disproved by current statistical frameworks or datasets. State-of-the-art evidence from a very recent study by Hivert et al. (2020) that performed large-scale variance component analysis in the UKBB across 70 complex traits, found no significant contribution to phenotypic variance from epistatic effects.

I now believe that those who think that evidence from model organisms or artificial populations imply that statistical epistasis may be relevant to humans may have made the same conceptual mistake I did. This thinking ultimately stems from conflating functional with statistical epistasis or alternatively phrased, the trait with the variance in a trait. To clarify, I will illustrate this with a quote from Mackay and Moore (2014), where the authors stated that "*quantitative variation in phenotypes and disease risk must result in part from the perturbation of highly dynamic, interconnected and non-linear networks [...] by multiple genetic variants; thus, gene-gene interactions are likely.*" Here, they reasoned that because the trait itself is a product of complex non-linear systems, trait variation would also require non-linear effects, whereas these two do not necessarily imply each other. For another illustrative example of this thinking, see the study by Kuzmin et al. (2018). Here, after finding that artificially introduced variation generated abundant higher-order interactions in yeast, the authors suggested that epistasis may explain missing heritability in humans. However, I now posit that this non-naturally segregating variation that they introduced via mutagenesis merely exposed latent functional epistasis by generating artificial heterozygosity at normally non-polymorphic loci. This in turn caused the disruption of the normal functionality of the genome (such artificial variation is almost always deleterious), rather than provided insights on normal trait variation. While such deleterious mutations may also occur in natural populations, as they are deleterious, they would not persist or be present in great enough numbers to affect population variance. Chance (rare) mutations, which selection have not had time to eliminate yet, in the constrained part of the genome could also conceivably manifest as statistical epistasis in humans; however, these would be either transient, or make up a very small fraction of the total trait variance. On the other hand, (missing) heritability is a property of normal trait variance due to naturally segregating variation in the base sequence. Considering this alternative explanation, I now do not believe that the evidence presented by the authors supports the conclusion that statistical interactions may be relevant to humans.

Finally, as for the few demonstrated cases of epistasis in humans (for example in rheumatoid arthritis (Dang et al., 2016; Génin et al., 2013; The Australo-Anglo-American Spondyloarthritis Consortium (TASC) et al., 2011)), these only support my current view that such effects are scarce, and in the larger landscape of phenotypic variation they account for little relative to additive associations. Indeed, the GWAS Catalog recorded an exponentially increasing number of additive associations in the last 10 years (Buniello et al., 2019), whereas there has been no comparable progress in epistasis detection.

The apparent lack of a direct contribution of non-linear genetic effects that would impact trait variance may appear puzzling at first, especially given the almost unimaginably complex encoding of information in the genome. However, this paradox may be resolved by Fisher's original explanation for this phenomena that I first described in the Introduction in section 1.1.5.2. Fisher proposed that non-linear information encoding may be achieved by one change at a time through purely additive processes. Under this model the probability of a new allele's frequency rising or falling is conditioned on the (potentially) fixed parts of the genome. Fisher remarked on this topic in his book The Genetical Theory of Natural Selection, where he wrote:

'[...] the effects by which any gene-substitution is recognized depend on the results of interactions with, possibly, all other ingredients of the germ plasm [...]' (p52, 2nd ed.).

In conclusion, perhaps the strongest argument against the importance of epistasis to trait variance is that it is simply not necessary. Given that nature tends to prefer parsimonious solutions where possible, if non-linear information encoding can occur from purely additive processes, then there is no need to introduce or even to maintain non-linear population genetic variance.

5.3 Outlook and future work

Even if non-linearity does not (substantially) directly contribute to phenotypic variance in a population, the way information is stored is still a crucial attribute of the genome, and decoding it will be essential to deepen our understanding of how genetic variation impacts complex traits and disease risk. Thus, with the benefit of hindsight, I feel that I spent my time looking for non-linearity in the wrong place, between polymorphic loci, rather than where it resides in abundance, in the rest of the genome. Therefore, my research interests now turn towards considering the non-linearity within fixed areas of the genome as a prior, and finding ways to connect that back to phenotypic variance.

Relating fixed parts of the genome to phenotypic variance may seem like an impossibility at first, as loci which do not vary in the population, by definition, cannot contribute to trait variance; thus, their function may appear inscrutable. However, sequence analysis is about relating the different parts of the genome against each other, loci which do not vary in the population still vary with respect to other regions of the genome; thus, invariant sequence context may be used to infer the effects of polymorphic loci. Therefore, examining the (local) sequence context of causal loci may reveal information about what makes, say, a height SNP a height SNP. If this information can be learned, then this may be used to predict a prior for polymorphic loci elsewhere; thus, accomplishing the goal to relate non-linear genetic effects in the fixed parts of the genome to phenotypic variance. Considering the wider field of how the NN framework is applied in genomics, I see a trend converging towards this goal.

The prevailing trend in most successful NN projects so far was to link narrow molecular phenotypes, such as TF binding, to local sequence context of ~1000bps. The scope of more recent sequence analyses have been gradually expanded to encompass larger and larger areas of the genome, which grew from 1000bp to ~131Kb, to consider more distal regulatory features (Kelley et al., 2018), even up to ~1Mb to study genome folding (Fudenberg et al., 2019). The complexity of the target phenotypes have also been increasing. Early efforts aimed to predict basic molecular phenomena, such as the presence of regulatory features; however, more recent studies have realised more ambitious goals, such as relating sequence features to gene expression via the integration of many smaller models over 40Kbs (Zhou et al., 2018). However, none have yet managed to explicitly tie non-linear genetic effects directly to genome-wide phenotypes, such as complex traits and diseases. Thus, I expect that connecting non-linearity in sequence data to phenotypic variance may be the next major challenge to be overcome in the coming years.

At this junction, it is also necessary to re-examine the ceiling of the maximum level of functional inference possible from NN based sequence analysis. As I covered in the Introduction in section 1.7, NNs perform best under a large data regime, where the outcome depends on non-linear combinations of the input features. To explore this argument further, I need to introduce a new concept which I will refer to as the 'self-containedness' of the problem being modeled. To clarify, this concept describes the observation that the class label of an image only depends on the pixels in the image, or that for games like GO (Silver et al., 2016) all the relevant information is included on the game board. For these types of prediction tasks a NN based model may achieve near perfect accuracy in prediction, as all the elements that contribute to the outcome are present in the training data. However, some may argue that biology is different, as biological systems potentially depend on input from external sources. Inference in this context would be equivalent to training from and then predicting trait SNP coefficients, such as those obtained from a GWAS. The model from which the SNP coefficients are obtained from include (covariates and) a noise term; thus in expectation, a SNP coefficient is the pure genetic effect driven by the base sequence alone, and is therefore predictable from the base sequence. Of course, the more complex the trait, the wider the context that would need to be considered; however, as the ultimate source of causality is still the base sequence, predicting SNP coefficients should also remain possible in theory. The overall trait inference possible from the sequence alone is quantified by heritability, which also represents a direct measurement of this aforementioned 'self-containedness' of the system. Recent heritability analyses revealed that for a great many traits the nucleotide base sequence is the ultimate origin of causality for the majority of trait variance (Polderman et al., 2015); thus in theory, the limits of trait inference from pure sequence data are also correspondingly high. From this perspective, the phenotype may be viewed as a non-linear transformation of the base sequence, up to level of broad-sense heritability. This view also implies that all intermediate stages such as cell, tissue and organ differentiation, expression levels, micro-biome (or at least the parts of these systems that are relevant to the traits), are also in turn determined by the base sequence. Thus, at least in theory, there would have to exist a direct non-linear map from the base sequence to the genetic component of the phenotype which does not depend on any further information from biological samples or environmental covariates. Given certain parallel developments in genomic studies described below, this suggests that modelling non-linearity may become increasingly important in the future.

The size of GWAS cohorts have been steadily increasing. Back in the 2000s studies typically numbered in the low thousands of individuals, whereas today meta-analyses have reached the ~1 million watermark (Lee et al., 2018). This trend is going to continue in the future, with biobank scale efforts in the UK alone set to reach ~5 million individuals with the *5 million genomes project* in 2023 (GEL, 2020). With other countries following suit, it seems highly likely that within a decade meta-analyses will reach cohort sizes on the order of tens of millions. This brings me to one of the rarely appreciated advantages of the GWAS design, which is the way its resource costs scale relative to studies that rely on more intrusive biological samples (such as specific cells or tissues). The biological data required for a GWAS is minimal, a saliva sample is sufficient, which may be collected during routine visits to one's GP. As electronic health records are becoming common (which may serve the the target phenotypes), GWAS may be considered as a mostly information based study that lends itself to large-scale automation, which could encompass entire populations in the near future.

Let us now consider studies that require more involved biological samples, such as biopsies of tissue samples, single-cell sequencing or microbiome data etc. These types of studies scale linearly with the number of sample donors, as they rely on manual and often labour intensive sample collection procedures. Also, the ceiling for cohort sizes would be limited to the fraction of the population willing to undergo such invasive procedures. Thus, it may be reasonably expected that while the costs of GWAS-like studies scale less than linearly with sample size, so these will likely to reach tens or even hundreds of millions of individuals, studies that require biological samples will grow in size at a far lower rate. I believe that this difference in scaling up may also mean that the relative importance of GWAS-like studies will grow disproportionately in the long term. This likely increase in the importance and size of GWAS type studies may also create more opportunities for methods that could provide mechanistic insights into the function of the genome based primarily on sequence information. Much of statistical genetics today is about recovering a faint signal from a very noisy source, whereas NNs excel in the task of modelling a highly complex non-linear signal when sample size is no longer a limiting factor. As the volume and the resolution of available genomic data increases, the field of genetics may start to resemble more closely the domains where NNs traditionally excel; hence, I expect the areas where NNs are applied in genomics to increase in the near future.

The current paradigm to infer mechanism relies on empirical evidence from the experimental manipulation of the genome that may yield insights on functionality, which could then be used to identify drug targets for example. While traditional GWAS is limited to identify one-to-one associations between individual genetic variants and phenotypic outcomes, the previously described trends, which will result in an orders of magnitude increase in genetic information, may open up opportunities for alternative approaches that could offer more mechanistic insights via advanced computational approaches. Methods such as NNs and their algorithmic descendants, which employ non-linear modelling of genetic effects, are uniquely suited to extract functional insights purely from genetic information by the virtue of the learned non-linearity. To illustrate why this is the case with a general example, consider a (fine mapped) GWAS SNP. Because of the additive nature of the GWAS association, it cannot reveal anything about its genomic context by itself. However, if associations would be made via implicating SNPs together with their relevant sequence context (that may include potentially non-polymorphic regulatory targets), then each association could also become biologically informative. Therefore, despite my own negative results in this work, I am cautiously optimistic about the future applicability of non-linear methods to genetic data, and I see a potential future where large biobank-scale GWAS and NNs are applied in complementary roles, as the former would generate the data and additive associations, and the latter could provide inference on mechanism of effect.