Utrecht University

Graduate School of Life Sciences
Master Molecular and Cellular Life Science
Writing Assignment

Omnigenic existential crisis: looking for validation

Author:
Jolien Rietkerk
6485677

Daily Supervisor:
Dr. P. Deelen
UMC Examiner:
Prof. Dr. J. de Ridder

May 24, 2020

Laymen's summary

Research has focused on the relationship between genotype (DNA) and phenotype (e.g. traits). The main question: what effect does a mutation in the DNA (a single nucleotide polymorphism, or SNP) have on a specific trait? Classical genetic theory, as described by Mendel, states that a single SNP or even a whole gene, can affect the complete trait. This is the case in rare diseases. Because we know that this effect is direct, we can easily predict if someone has a disease or if offspring will likely get the disease. It is not that simple with complex diseases. These diseases have a genetic background influences a trait in such a way that cannot be explained by the Mendelian rules. This is a problem, because there are many complex diseases for which treatment is needed, or from which the molecular mechanism is very unclear.

In recent years, researchers have hypothesized on a model that explains the heritability and genetic background of complex diseases. This new model, called the omnigenic model, contains two other concepts: the infinitesimal model and the concept of polygenicity.

In the infinitesimal model, all genes in an organism together explain the heritability of a trait in equal amounts. Polygenic traits are also highly complex, but do not state that all genes explain heritability, but merely a lot of genes explain heritability.

The omnigenic model also states that the genetic background of complex traits exists of many genes that together explain a lot of the heritability, but goes into more detail in specifying that all genes are part of a network. In this network there is a different types of genes: core genes and peripheral genes. The current work discusses the specific definition of these genes to be as follows: Core genes are genes that have a direct effect on the phenotype, the mechanism of which it does this is not specifically explained by the omnigenic model. The peripheral genes are non-core genes that can regulate and interact with core genes, but do not have a direct effect on the phenotype. For the omnigenic model to be statistically correct, there are more peripheral genes than core genes. So much more peripheral genes even, that the amount of heritability explained by core genes would not be detectable in genome-wide studies.

Last, there is a specific type of peripheral gene, the peripheral master regulator, that can regulate many peripheral and core genes. Because of the central position of these peripheral master regulators, they can have a large effect on the network and are therefore often easily detected.

The current review explores literature that addresses the omnigenic model and compares the interpretations of the omnigenic model in literature to the definition of core genes in order to get better insight on the validity of the omnigenic model.

In general, literature is in agreement with the omnigenic model, however interpretations of the core gene definition differ, which make it difficult to compare and interpret results. The mechanism of the direct effect that core genes would have on a phenotype remains unclear and it is advised to focus research on network analysis and integration of multiple data types in order to get a better definition of core genes and explain more details of the genotype - to - phenotype relationship.

Abstract

A main endeavour in genetics is to unravel the mechanisms that govern the relationship between genotype and phenotype. The authors of the omnigenic model [1] have created a starting point for discussion on the molecular mechanism that governs this relationship. The omnigenic model states that most genes in disease relevant cells contribute to the heritability of a trait. These relevant genes are in a network where core genes are at the center, surrounded by trans-acting peripheral genes and peripheral master regulators. Literature appears undecided on the official definition of core genes and so I defined the core gene definition using the most recent reiteration of the omnigenic model by Liu et al. [2]. This definition exists of three parts: first, core genes have a direct effect on a phenotype. Second, peripheral genes are non-core genes that affect the phenotype indirectly, and third, peripheral master regulators can regulate multiple genes and so be more often detected in Genome-Wide Association Studies (GWAS). This prioritized papers on their mentioning of the core gene definition, omnigenic model or interesting new views on the overall genotype - phenotype relationship. I discuss how literature is mostly in line with the omnigenic model, but that more research into genotype-to-phenotype mechanisms is needed in order to truly validate the core gene definition, suggesting to focus on network analysis and the combination of different molecular data types.

Contents

Chapter 1

Introduction

A main endeavour in genetics is to unravel the mechanisms that govern the relationship between genotype and phenotype. Classical genetics by Mendel states that this relationship can be straightforward, with a phenotype having a basis in a single gene that can be predictably inherited by offspring. These types of genes have a large additive effect on the phenotype, since altering the function of these genes or removing these genes can change a phenotype drastically, often leading to severe diseases. In case of these severe diseases, evolutionary selection keeps the mutations at a minimum and so these severe diseases are very rare. [3]

The mutations (or Single Nucleotide Polymorphisms: SNPs) that occur often and naturally are not selected against, since they do not have such a high effect on a phenotype and so do not have any severe disease phenotypes. However, researchers found that common variants are likely involved in the heritability of complex diseases. Complex diseases, by definition, are based on multiple variants and do not follow classical Mendelian inheritance. Many examples exist for complex diseases, like Schizophrenia, Alzheimer's Disease [4] and Amyotrophic Lateral Sclerosis (ALS). [5][6]

Researchers aim to understand the molecular mechanisms that govern disease etiology in order to, among others, identify risk factors and possible targets to be involved in treatment of the diseases. This research is conducted through a number of methods, currently centered around high-throughput methods such as SNP arrays, Whole Genome Sequencing (WGS) and Next Generation Sequencing (NGS).

One of the analyses done on SNP arrays is Genome-Wide Association Studies (GWASs). In GWAS one associates genetic variants on the SNP array with the phenotype in question through statistical testing, comparing 'healthy' controls to patients with the trait in question. [7] The SNPs that associate to the phenotype are called variants. An important quality of these variants is that many of them are often inherited together, due to the low frequency of recombination that occurs between variants. This phenomenon is called Linkage Disequilibrium (LD). This, together with the fact that SNP arrays do not cover the complete genome makes that the variants associated with the phenotype (or variants that explain most of the variation in the phenotype) are rarely directly causal to the phenotype. And so the main challenge in GWAS analysis is to identify causal variants and possibly causal genes or pathways to explain disease etiology. [3][6] [8]

To this end, multiple models have been proposed to explain the relationship between genotype and phenotype based on experimental evidence from genetics studies such as GWAS, WGS, NGS and even combining them with other layers of biology like measures of transcription (mRNA expression levels) and protein expression levels. [6] [9]

The first model that used genetics to explain the genotype-phenotype relationship in complex traits was the infinitesimal model. [10] This model describes that eventually all heritability of a phenotype (or trait) can be explained by adding together the explained variation of all genes in the genome. This implies that each individual gene has an infinitesimal effect size on the phenotype. Recently, discussion on GWAS result interpretations have revisited this older model and found it to be a reasonable explanation of the problem of 'missing heritability' [11] that was proposed when researchers first found how small the individually explained variance per variant is. [6]

A concept that is, to some extent, similar to the infinitesimal model is polygenicity. [8] The idea of polygenicity adds directionality to genetic interactions and does not necessarily state that all genes explain a trait, but that many genes can explain a trait. With the conceptual background of polygenicity, the overall

effect sizes of variants are averaged, making it hard to identify individual unique combinations of up or down regulation effects of certain variants and so the effect size of the variants that are significant are found to be small. [6] [12]

Another iteration of the infinitesimal model that combines it with the concept of polygenicity as well, is the omnigenic model. This model was first proposed in 2017 by Boyle and colleagues and is based upon the idea that genes are a part of a network in trait-relevant cells [13]. More specifically, it makes the distinction between two types of genes within this network. Core genes and peripheral genes. [1] [2] This distinction can be compared to that between cis and trans interactions. *Cis* interactions are those of genes that are of short genetic distance of each other, whereas *trans* interactions are interactions of genes that are far apart in the genome, sometimes even interactions between different chromosomes.

In the omnigenic model, core genes are genes in the network that have a direct effect on the phenotype, but are highly outnumbered by peripheral genes. Core genes could carry some more rare variants, due to the higher biological impact of core genes on the phenotype. Peripheral genes, by exclusion, are non-core genes that can act through trans interactions upon the core genes and are sometimes directly involved in regulation of core genes. It is the large number of peripheral genes that explain the most of the heritability of a trait. The fact that they outnumber the core genes makes it so that core genes, regardless of their biological importance, explain little heritability. The connectivity in the network that is seen in this model is similar to the polygenic model. [14]

The initial publication of the omnigenic model [1] has been interpreted differently in literature. Most importantly, the definition of core genes is often not fully explained or the distinction between core genes and peripheral genes is not made clear. The definition of core genes that was first proposed by Boyle et al. [1] was quite broad and open for interpretation. However the more recent publication of core gene definition [2] was less broad, but still has room for uncertainty.

This review aims to clearly describe the definition of core genes using its fundamentals in Gene Regulatory Networks (GRNs) and explain the shortcomings of this definition. Secondly, since literature seems undecided, this review tests the hypothesis that core genes exist. This is done by reviewing iterations of the omnigenic model, interpretations of the core gene definition, and other views on the genotype-phenotype relationship that is presented in current literature.

Chapter 2

Defining 'core genes'

At the center of the omnigenic model are core genes. This chapter provides an overview of how the most recent definition of core genes came to be, and highlight remaining uncertainties. In the next chapter I discuss the level to which literature is in accordance with this definition.

2.1   Rate-limiting genes are the basis of the 'core gene' definition

The definition of core genes in the omnigenic model published by Boyle et al. [1] is as follows: "'a modest number of genes or gene pathways with specific roles in disease etiology, as well as their regulators" and "such genes will tend to have biologically interpretable roles in disease...". This definition takes the assumption that genes are in a network and cites a paper by Chakravarti and Turner, who describe the concepts of Gene Regulating Networks (GRNs) that contain rate-limiting genes. [13] Interestingly, rate-limiting genes are somewhat in line with the concept of core genes.

Chakravarti and Turner describe that all genes are part of a GRN. However, not all genes have the same functionality in this network, not all effects are of the same size or have the same directionality. Also, not all SNPs affecting a GRN are necessarily shared between all individuals. The authors argue that each GRN contains a rate-limiting gene that has the largest additive effect size on a trait. To explain, the authors give the example of the RET gene. The RET gene is one of 15 genes related to Hirschsprung disease and is shared between all disease cases. Disruption of the RET gene function is not enough to create the phenotype, but it is necessary for the phenotype. It is this quality that makes the RET gene the rate-limiting gene in this context. For visualisation, figure 2.1a shows a hypothetical GRN with genes X and Y indicated. To determine which of the genes X and Y in this network is the rate-limiting gene (or the gene with the largest additive effect size) I indicated the effects of deleting either X or Y from the network in figure 2.1b and c respectively. Clearly, the effects on the GRN by deleting X are larger than deleting Y from the network. Specifically, X has a more general effect than Y. And so, X would be the rate-limiting gene here.
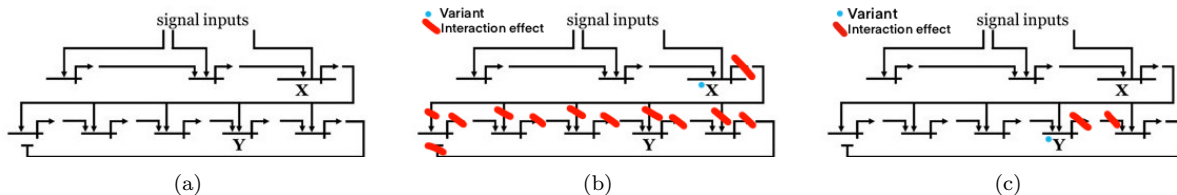

Figure 2.1: Adapted figures from Chakravarti and Tuner [13] figure 2. Genetic interactions that are affected by a certain deletion or functional impairment are indicated with red hashes. (a) a hypothetical gene regulatory network (GRN) comprising of different genes with directional effects. X and Y are genes that are taken for comparison. Lines and arrows indicate directional interactions of genes. (b) The effect on the genetic interactions within the hypothetical GRN of (a) when a deleterious mutation of gene X occurs.(c) The effect on the genetic interactions within the hypothetical GRN of (a) when a deleterious mutation on gene Y occurs. Note that Chakravarti and Turner argue that gene X is the rate-limiting gene here as it has a larger effect on the GRU than Y, where the effect of Y is purely a local effect and more regulation exists for the network to compensate for the loss of function in Y.

2.2 A specification of the 'core gene' definition is needed

The concept of rate-limiting genes can be seen in the definition of core genes as first proposed by Boyle et al.: "a modest number of genes or gene pathways with specific roles in disease etiology, as well as their direct regulators". However, the omnigenic model and core genes are clearly more generally defined, as it also includes the possibility of having multiple rate-limiting genes/core genes. As rightly argued by Wray et al. [14] this could be problematic since it makes the difference between core genes and peripheral genes quite unclear and implies that a model with a large number of core genes might exist, which could become similar to having no core genes defined in a network.

It is this rapid deviation from a single rate-limiting gene to the idea of multiple core genes and pathways that make it necessary to keep the definition of core genes specific and possibly more binary than first intended by Boyle et al..

2.3 The latest 'core gene' definition is open for interpretation

Indeed, a specification has been given by Liu et al. [2] in an iteration of the omnigenic model. The definition of core genes in this paper is mentioned several times, one more specific then the other:

1. "Core genes can affect disease directly, while peripheral genes can only affect risk indirectly through trans-regulatory effects on core genes"

2. "We define a gene as a core gene if and only if the gene product (protein, or RNA for a non-coding gene) has a direct effect - not mediated through gene regulatory networks - on cellular and organismal processes leading to a change in the expected value of a particular phenotype"

3. "... by exclusion, all other genes expressed in relevant cell types are considered peripheral genes, and can only affect the phenotype indirectly through regulatory effects on core genes"

4. "This definition of core genes implies that the phenotype of an individual is conditionally independent of the peripheral genes, given the expression levels and coding sequences of the core genes."

Besides defining core genes and peripheral genes, Liu et al. describe another term: "peripheral master regulators"(PMRs). These regulators are said to be the key regulators of multiple (core) genes and have a potential to have "a relatively large effect size". The authors argue that their importance might not be represented in the GWAS signals of such genes since the selective constraints on these regulators could be strong and so variants here would be rare as well.
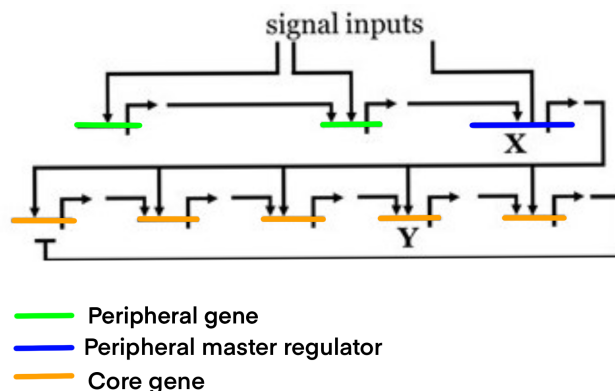

Figure 2.2: Adapted figures from Chakravarti and Tuner [13] figure 2 to visualise the definitions of core gene (orange), peripheral gene (green) and peripheral master regulator (blue) as proposed by Liu et al. [2] and as used in this thesis. Note that X in this case is a peripheral master regulator, not a core gene or 'rate-limiting gene' as was described in figure 2.1. Note that the identities of genes with these terms differ per network and can be ambiguous, for example master regulator can be the last to act on most downstream effector molecules, so X in this case, or it could be a regulator that regulates core genes as well as peripheral genes.

In contrast to the paper by Boyle et al. [1], where the definition of core genes appeared to be in line with the rate-limiting genes, the definition of core genes in the paper by Liu et al. is no longer in line with the idea of rate-limiting genes. In this paper the definition of PMRs fits the rate-limiting concept better, since the rate-limiting gene was part of a GRN and would have the largest effect size. [2][13]

The definition of core genes versus peripheral genes by Liu et al. has taken a turn in a different direction in stating that core genes have a direct effect on a trait, which narrows down the possible candidates for a core gene. To have a definition for comparison to interpretations in literature I will define the core genes using this last definition, in the definition that core genes are: **genes that have a direct effect on cellular or organismal functioning via their direct protein or RNA products and which do not regulate other proteins. Additionally, core genes cannot be peripheral master regulators. All other genes in relevant cell types are considered general peripheral genes.**(figure 2.2)

It is important to note that even though in this definition a clear cutoff appears to exist between core and peripheral genes, the definition remains ambiguous. This is due to the concept that core genes have a direct effect on a trait, is an assumption about the relationship between genotype and phenotype, where it would be clear that genes that directly affect phenotype exist. The latter is the topic of the next chapter, where I review papers that provide evidence for the genotype to phenotype relationship, core gene, peripheral gene and master regulator definitions.

Chapter 3

Validating 'core genes'

The previous chapter specified the term 'core genes': core genes are genes that have a direct effect on cellular or organismal functioning via their direct protein or RNA products and which do not regulate other proteins or genes. Additionally, core genes cannot be peripheral master regulators. All other genes in relevant cell types are considered peripheral genes. (for clarity referred to as 'the core gene definition' throughout this chapter)

This definition can be split into three parts: 1. core genes and the effect they have on a phenotype; 2. peripheral master regulators and 3. general peripheral genes.

This chapter presents literature that addresses the omnigenic model and specify their interpretations of it, in order to validate each of these 3 parts of the core gene definition.

3.1 Core genes and their effect on phenotypes

Core genes themselves are seldom discussed in depth in literature. Most papers discuss the 'omnigenic side' of the model, being that their results fit the omnigenic model, or it is in line with the concept of trans-acting variants that are present in large numbers in the data. (see supplements) Various papers seem to have taken the core gene definition with some liberty. One of them is a paper by Valdés et al. [15] which already shows an interpretation of core genes that is not in line with the omnigenic model reference they use (Boyle et al.[1]). Specifically, the authors state: "... the recently proposed omnigenic model hypothesis, which states that complex traits can be influenced mostly by genes outside not only by 'core genes' mainly found by the genome-wide significant SNPs, but also the rest of genes outside of the 'core pathways' with apparent unrelated biological functionality."

In contrast to the definition of Boyle et al. where core genes are often not strong hits in GWAS studies due to the conceitedness of the GRN and outnumbering peripheral genes, Valdés et al clearly state that they considered core genes to be main hits found by GWAS. Interestingly, it appears their discussion of the omnigenic model is flawed, as they present some functionality in their prioritized SNPs to be related to protein binding. The high presence of protein binding hits in GWAS data was an argument used by Boyle et al. in favour of the existence of peripheral genes. Thus, it appears that even though the authors state that their results are in line with the omnigenic model and core genes, they merely identified peripheral genes and are so only partially in line with the core gene definition.

A paper with a core gene definition more close to the core gene definition above, is by Wang et al. [9] who state: "the core gene, which has a direct effect on a change in the expected value of a phenotype...".

In this paper, the authors go into the genotype-phenotype relation by defining six logical patterns of relationships between SNPs, mRNA levels and protein levels. The authors search for each pattern in a data set from mice and humans. All patterns were found to be present in different amounts, but the pattern in line with the omnigenic model was one of two major patterns. This pattern describes how SNPs and mRNA expression levels independently influence protein levels, which is in line with the above mentioned core gene definition. This confirms that such 'direct effects' would exist and that it is present in relatively large numbers. The authors hypothesize on the probability of peripheral genes and a more complex GRN that might contain their patterns, but it is promising that they find SNP-protein and mRNA-protein patterns.

It is important to note that this implies a definition of phenotype that takes proteins themselves as a phenotype, which not necessarily for GWAS variants found using systemic phenotypes.

A review by Hughes et al. [16] also argues that in research to cardiovascular disease (CAD) the omnigenic model is well accepted because of the many variants that add to heritability. However, they also focus their review on three variants previously proposed by Harst and Verweij [17] which "may represent core genes but they may be signals that are context or cell specific to CAD". This seems to imply their agreement with the core gene definition in that they influence the phenotype, CAD. Later, the authors refer to these variants as core variants that are related to pathways driving CAD and so, these variants could be core genes and this paper is in line with the core gene definition.

Another paper that argues in favour of the omnigenic model and core genes, is by Broekema et al. [8] It defines core genes as: "the genes on which many peripheral genes converge (core genes) generally have stronger effects on the phenotype. As such, the variants that affect core genes are more likely to be Mendelian disease variants." This definition is more in line with the first version of the omnigenic model, where single Mendelian - like variants were hinted to be core genes, whereas the second version and definition above suggest multiple genes to be core genes. The figure accommodating this definition however, shows multiple genes and so their statement on Mendelian variants is somewhat misleading. An addition to their interpretation of the core gene definition is given in their discussion, where the authors explain that if expression levels of genes are strongly correlated, it is likely that these genes are core genes in a GRN associated to disease. This implies a joint effect of core genes on a phenotype, which does not contradict the above described definition of core genes. And so this paper seems in line with my definition.

In a review on the psychiatry of Schizophrenia, the omnigenic model appears promising: "The omnigenic model even proposed that almost every gene could contribute to disease risk and that 'peripheral genes' cumulatively contribute more heritability through trans-effects than 'core genes' do through cis-effects." and "although only a few specific genes are the primary focus of many hypothesis, they account for a tiny fraction of disease genetic associations." This is in accordance with the omnigenic model and the core gene definition. The limitation of core genes to cis-effects even further strengthens their case to be in line with the core gene definition, since often regulatory effects by proteins and genes are trans-effects, however cis-effects on other core genes might occur. [18]

A final paper stating to be in line with the core gene definition is published by O'Connor et al.[19]. They are very clearly in agreement with the omnigenic model and the definition described above: "A limited number of core genes have direct effects on a trait, but due to densely connected cellular networks, thousands of other peripheral genes- perhaps including every gene expressed in a relevant cell type- also contributes to heritability. Our results support the distinction between core and peripheral genes, suggesting that only a limited number of genes have direct phenotypic effects if mutated, and that these genes explain little heritability... one might expect that core genes, even if they explain a minority of heritability due to the larger number of peripheral genes, would usually harbor the strongest common-variant associations. However, our results suggest that they may not, due to being strongly contained by (negative) selection."

This clearly shows their agreement with the omnigenic model. The results this section discusses are referring to the authors research into the effects of negative selection on the distribution of SNP effect sizes vs explained heritability. They introduce a concept called flattening, which they test by computational modelling. The full extent of which is out of the scope of this review, however it is an interesting argument in favour of the omnigenic model, in the sense that the heritability of genes is spread across the genome.

In conclusion, the papers discussed here are in line with core genes and their direct effect on a phenotype, to different extents. Interestingly, some papers appear to refer to the existence of peripheral genes to hint towards a valid omnigenic model as well.

3.2   General peripheral genes

The existence of these general peripheral genes is implied by the validity of the above mentioned core gene definitions, since the distinction between core genes and peripheral genes go hand in hand. However, the selection of the papers discussed in this review (see supplements) highlights many papers that cite the omnigenic model but merely state an accordance with peripheral genes, not core genes. Since the distinction is key to the omnigenic model (as well as both concepts) I discuss here some papers that merely state the existence of peripheral genes.

First, as explained above, the paper by Valdés et al. [15] is not necessarily in line with the core gene definition. The same applies to another paper in the selection: Jakobson et al. [20]. "The genome wide distribution of QTLs we observe thus reconciles the statistical and theoretical architecture of the omnigenic model with our observation that not all segregating variants impact every quantitative trait."

Earlier in their text, the authors do mention that the QTLs found are spread across the genome. This is all in line with the omnigenic model and high number of peripheral genes and their distribution on the genome. The QTLs the authors found were found using *Saccharomyces cerevisiae* yeast as a model organism for heritability. An interesting approach for followup research, as discussed in the next chapter.

These two papers are a mere few of the papers stating their accordance with peripheral genes (see supplements). This validates the existence of peripheral genes that is required for the definition of core genes to be true as well.

3.3 Peripheral master regulators

The final part of the core gene definition is that core genes are not peripheral master regulators, implying that peripheral master regulators exist. Out of the papers prioritized (see supplements), two argue to be in line with the omnigenic model and core genes, but with the current definition, are better suited to be PMRs.

Pullabhatla et al. [21] conclude their studies by stating "our results support extreme phenotype sampling and *de novo* mutation discovery to aid a hypothesis - driven search for rare variant associations with complex diseases in the hunt to determine core disease genes". Their focus on rare variants with a high phenotypic effect size is somewhat in accordance with the core gene definition, however the authors mention that genes are part of many phenotypes and a high focus on identifying rare variants is more a search for Mendelian-type variants likely to be PMRs when the effect sizes are delectably high. This is confirmed when the authors directly cite Chakravarti and Turner [13] who described the rate-limiting gene concept. And so, this paper aligns with the omnigenic model, but describes PMRs, not core genes.

Second, a similar concept is seen in the paper by Rammos et al. who state: "Genes identified in GWASs or rare variant studies may be core genes that serve as the basis of developing networks used to identify peripheral genes". Note that the focus here is also on rare variants for core genes, like Pullabhatla et al.. Therefore, this paper is arguably also more in line with PMRs than with core genes.

Both these papers are in line with the initial definition of core genes, which can be expected, since these papers date back to before the second iteration of the omnigenic model was published. Nevertheless, it is important to note this distinction and it is important that this part of the core gene definition is verified as well. Additionally, PMRs can still be interesting targets for treatment as was stated by Liu et al.. [2]

Chapter 4

Discussion and future prospects

Since it was first proposed, the omnigenic model has been reiterated, interpreted and evolved. The omnigenic model makes a distinction between core genes and peripheral genes. The specific definition of core genes and peripheral genes has been interpreted differently as well as the validity of the model itself. In order to validate the omnigenic model and its implications using literature this review discussed the question: do core genes exist?

Initially, the omnigenic model seemed to be built around the key concept of the distinction between core genes and peripheral genes within a GRN. Core genes are implied to be a small number of genes that drive disease etiology but are outnumbered by trans-acting peripheral genes. As a result, core genes often go undetected and add little to the variance explained in a GWAS experiment compared to the vast number of peripheral genes in the network. I conclude that, with the second iteration of the model that implements a much needed specification of the definition, the authors intend core genes to be those directly affecting a phenotype, peripheral genes being all other genes by exclusion, and peripheral master regulators that could be high hits in GWAS as a result of their large effect size, considering the hypothetical large number of proteins such a regulator would regulate.

Even though this definition eliminates the broad interpretation of the original definition, the authors specifically mention that their model implies larger numbers of core genes, hinting at core pathways that would govern disease etiology.

When validating the existence of core genes, the review addressed three parts that go together with the core gene definition: core genes and their effect on the phenotype, peripheral genes and peripheral master regulators.

The open definition of core genes has created some irregular interpretations of the core gene definition. Nevertheless, the papers reviewed here are mostly in line with the core gene definition. With regard on the direct effect on a phenotype that these genes might have, it is hard to come to a conclusion whether such an influence exists. This mostly because of a gap between GWAS and mapping methods, from mapping variants to protein expression to doing GWAS on system-wide phenotypes. One solution to bridge this gap is proposed by Liu et al. [18] who propose the term intermediate phenotypes or cellular phenotypes which could help associate variants to diseases on a new level, creating a more clear view on cellular networks and the mechanisms that could govern the direct effect of core genes on phenotypes.

One effect that has been discussed however, is that of SNPs and mRNA levels on protein levels (among other patterns) by Wang et al. [9] which elucidates a mechanism that is often present in molecular networks.

It is studies like this that dive deeper into the cellular networks that could elucidate the effect that core genes could have on a phenotype on a molecular level. The promise of which was emphasized by Wang et al. [9] as well. That regardless of whether core genes exist in the definition they currently have, network analysis should elucidate more on the genotype - phenotype mechanisms.

The peripheral genes in the omnigenic model are more commonly accepted than the core gene concepts, even though their definitions are highly associated. Many papers shortly mention that their high number of variants found are in accordance with the omnigenic model and the high number of peripheral genes. An interesting part of the peripheral genes in the network is the specific peripheral master regulator.

The PMRs encountered here, were initially found as core genes under the previous definition of the

omnigenic model. However, with the current iteration of the definition high effect size, rare variants are simply more likely to be PMRs instead of core genes.

To conclude, the omnigenic model appears to be mostly accepted in literature, and core genes appear to exist. However, literature is not in full agreement on the definition of core genes and direct evidence that the core genes exist, that does not refer to the definition of PMRs, is scarce. My advice for future research would be to focus on the network centered research to elucidate more on the genotype - to - phenotype relationship in order to clearly state whether genes with a direct effect on a phenotype exist. This, since the high connectivity of the gene regulatory networks in organisms appears to be highly complex indeed.

Acknowledgements

Bibliography

[1]   E.A. Boyle, Y.I. Li, and J.K. Pritchard. "An Expanded View of Complex Traits: From Polygenic to Omnigenic". In: *Cell* 169 (2017), pp. 1177–1186. DOI: `10.1016/j.cell.2017.05.038`.

[2]   X. Liu, Y.I. Li, and J.K. Pritchard. "Trans Effects on Gene Expression Can Drive Omnigenic Inheritance". In: *Cell* 177 (4 2019). DOI: `10.1016/j.cell.2019.04.014`.

[3]   G. Gibson. "Rare and Common Vriants: Twenty arguments". In: *Nat Rev Genet.* 13(2) (2015), pp. 135–145. DOI: `10.1038/nrg3118`.

[4]   T. Horwitz et al. "A decade in psychiatric GWAS research". In: *Mol Psychiatry* 24,3 (2018), pp. 378–389. DOI: `10.1038/s41380-018-0055-z`.

[5]   van Rheenen et al. "Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis." In: *Nature Genetics* 48 (2016), pp. 1043–1048. DOI: `10.1038/ng.3622`.

[6]   P.M. Visscher et al. "10 years of GWAS discovery: Biology, Function, and Translation". In: *The American Journal of Human Genetics* 101 (2017), pp. 5–22. DOI: `10.1016/j.ajhg.2017.06.005`.

[7]   D. Botstein and N. Risch. "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease." In: *Nature Genetics* 33 (2003), pp. 228–237. DOI: `10.1038/ng1090`.

[8]   R.V. Broekema, O.B. Bakker, and I.H. Jonkers. "A practical view of fine-mapping and gene prioritization in the post-genome-wide association era". In: *Open Biol.* 10:190221 (2020). DOI: `10.1098/rsob190221`.

[9]   Y. Wang et al. "Comprehensive Cis-Regulation Analysis of Genetic Variants in Human Lymphoblastoid Cell Lines." In: *Front Genet* 10 (2019). DOI: `10.3389/fgene.2019.00806`.

[10] N.H. Barton, A.M. Etheridge, and A. Véber. "The infinitesimal model: Definition, derivation, and implications". In: *Theoretical Population Biology* 118 (2018), pp. 50–73. DOI: `10.1016/j.tpb.2017.06.001`.

[11] T.A. Manolio et al. "Finding the missing heritability of complex diseases". In: *Nature Genetics* 48 (2016), pp. 1043–1048. DOI: `10.1038/ng.3622`.

[12] R.A. Fischer. "The correlation between relatives on the supposition of Mendelian inheritance." In: *Philos. Trans. R. Soc. Edinb.* 52 (1918), pp. 399–433.

[13] A. Chakravarti and T.N Turner. "Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families". In: *Bio3ssays* 38 (2016), pp. 578–586. DOI: `10.1002/bies.201500203`.

[14] N.R. Wray et al. "Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model". In: *Cell* 173 (2018), pp. 1573–1580. DOI: `10.1016/j.cell.2018.05.051`.

[15] M.G. Valdés et al. "Pipeline design to identify key features and classify the chemotherapy response on lung cancer patients using large-scale genetic data." In: 12:97 (2018). DOI: `10.1186/s12918-018-0615-5`.

[16] M.F. Hughes et al. "Exploring Coronary Artery Disease GWAs Targets With Functional Links to Immunometabolism". In: *Front Cardiovasc Med.* 5:148 (2018). DOI: `10.3389/fcvm.2018.00148`.

[17] P. van der Harst and N. Verweij. "Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease." In: *Circ Res.* 122 (2018), pp. 433–43. DOI: `10.1161/CIRCRESAHA.117.312086`.

[18] C. Liu, S.V. Faraone, and J. Glatt. "Psychiatric Genetics, Epigenetics, and Cellular Models in Coming Years". In: *J Psychiatric Brain Sci.* 4 (2019). DOI: `10.20900/jpbs.20190012`.

[19] L.J. O'Connor et al. "Extreme Polygenicity of Complex Traits Is Explained by Negative Selection". In: *The American Journal of Human Genetics* 105 (2019), pp. 456–476. DOI: `10.1016/j.ajhg.2019.07.003`.

[20] C.M. Jakobson and D.F. Jarosz. "Molecular Origins of Complex Heritability in Natural Genotype-to-Phenotype Relationships". In: *Cell Systems* 8 (2019), pp. 363–379. DOI: `10.1016/j.cell.2019.04.002`.

[21] V. Pullabhatla et al. "De novo mutations implicate novel genes in systemic lupus erythematosus". In: *Hum Mol Genet.* 27(3) (2018), pp. 421–429. DOI: `10.1093/hmg/ddx407`.

[22] A.W. Charney et al. "On the diagnostic and neurobiological origins of bipolar disorder." In: *Transl Psychiatry* 10(1) (2020), p. 118. DOI: `10.1038/s41398-020-0796-8`.

[23] R.S. Singh and B.P. Gupta. "Genes and genomes and unnecessary complexity in precision medicine". In: *npj Genomic Medicine* 5:21 (2020). DOI: `10.1038/s41525-020-0128-1`.

[24] L.L. Colbran et al. "Inferred divergent gene regulation in archaic hominins reveals poteintial phenotypic differences". In: *Nat Ecol Evol.* 3(11) (2019), pp. 1598–1606. DOI: `10.1038/s41559-019-0996-x`.

[25] C.G. deBoer et al. "Deciphering eukaryotic gene-regulatory logic with 100 million random promoters." In: *Nat Biotechnol.* 38(1) (2020), pp. 56–65. DOI: `10.1038/s41587-019-0315-8`.

[26] D.M. Jordan, M. Verbanck, and R. Do. "HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases." In: *Genome Biol.* 20(1) (2019), p. 222. DOI: `10.1186/s13059-019-1844-7`.

[27] Z. Dai, N. Long, and W. Huang. "Influence of Genetic Interactions on Polygenic Prediction." In: *G3* 10(1) (2020), pp. 109–115. DOI: `10.1534/g3.119.400812`.

[28] S. Burgess and G. Davey Smit. "How humans can contribute to Mendelian randomization analyses." In: *Int J Epidemiol* 48(3) (2019), pp. 661–664. DOI: `10.1093/ije/dyz152`.

[29] S.J. Baumgart, E. Nevodomskaya, and B. Haendler. "Dysregulated Transcriptional Control in Prostate Cancer." In: *Int J Mol Sci* 20(12) (2019). DOI: `10.3390/ijms20122883`.

[30] H.A. Weisz et al. "MicroRNA sequencing of rat hippocampus and human biofluids identifies acute,chronic, focal and diffuse traumatic brain injuries." In: *Sci Rep* 10(1) (2020), pp. 1562–1569. DOI: `10.1038/s41598-020-60133-z`.

[31] S. Luts et al. "DNA variants affecting the expression of numerous genes in trans have diverse mechanisms of action and evolutionary histories." In: *PLoS Genet.* 15(11) (2020). DOI: `10.1371/journal.pgen.1008375`.

[32] B.P.H Metzger and P.J. Wittkopp. "Compensatory trans-regulatory alleles minimizing variation in TDH3 expression are common within Saccharomyces cerevisiae." In: *Evol. Lett.* 3(5) (2019), pp. 448–461. DOI: `10.1002/evl3.137`.

[33] L. Pineda-Cirera et al. "Exploring genetic variation that influences brain methylation in attention-deficit/hyperactivity disorder". In: *Transl Psychiatry.* 9(1) (2019), p. 242. DOI: `10.1038/s41398-019-0574-7`.

[34] Y. Qui et al. "Oligogenic Effects of 16p11.2 Copy-Number Variation on Craniofacial Development." In: *Cell Rep* 28(13) (2019), pp. 3320–3328. DOI: `10.1016/j.celrep.2019.08.071`.

[35] N. Schrode et al. "Synergistic effects of common schizophrenia risk variants." In: *Nat Gen* 51(10) (2019), pp. 1475–1485. DOI: `10.1038/s41588-019-0497-5`.

[36] H.E. Wheeler et al. "Imputed gene associations identify replicable trans-acting genes enriched in transcription pathways and complex traits." In: *Genet Epidemiol* 43(6) (2019), pp. 596–608. DOI: `10.1002/gepi.22205`.

Supplements

Prioritizing papers that cite the second iteration of the omnigenic model by Liu et al. [2]

Chapter 3 explores papers that cite either the initial omnigenic paper by Boyle et al. [1] or the iteration of the omnigenic model by Liu et al. [2] and explore their accordance with this definition of core genes.

I selected 4 out of 433 papers that cite Boyle et al., which contain the term 'core gene' in the title or abstract as a measure for the depth of their discussion of the omnigenic model core gene definition.

Additionally, I searched the terms 'core gene' and 'omnigenic' in papers that cite Liu et al.[2]. However, no papers contained these terms in the abstract or title and so I analysed the 20 papers that cite Liu et al. on the sentence preceding the citation and overall indications of discussing the omnigenic model and/or implying different models for networks or genotype to phenotype interactions.

Table S1 shows this subjective prioritization of the 20 papers that cite Liu et al. [2]. Interestingly, as can be seen in figure (based on table S1) there are 3 papers that do not discuss the omnigenic model to an extent that it adds to the discussion, there are 6 papers that discuss the omnigenic model in more detail, discuss either of the parts of the definition or propose different models than the omnigenic model. There are 4 papers that only briefly discuss that their data is in line with the omnigenic model and there are a total of 7 papers that discuss that the omnigenic model is valid with their data, but merely apply it to the trans-acting peripheral genes part of the omnigenic model, and do not specify on the existence of core genes.

This implies that the majority ($7 + 4 = 11$) of the 20 papers do not show a clear disagreement with the omnigenic model in general, nor denies the existence of core genes. The 5 papers that discuss the omnigenic model in more detail or present different models are reviewed below, in context of the core gene definition above and to see whether this definition holds within the new models.

Does not add to the discussion	[22] [23] [24]
Adds to discussion in depth	[8] [18] [9] [19] [20]
Discusses fitting with the omnigenic model briefly	[25] [26] [27] [28][29]
Only mentions being in line with trans-acting variants and large polygenicity	[30] [31] [32] [33] [34] [35] [36]

Table S1: Table of subjective classes and the papers that were assessed to belong to that class, as shown in figure S1
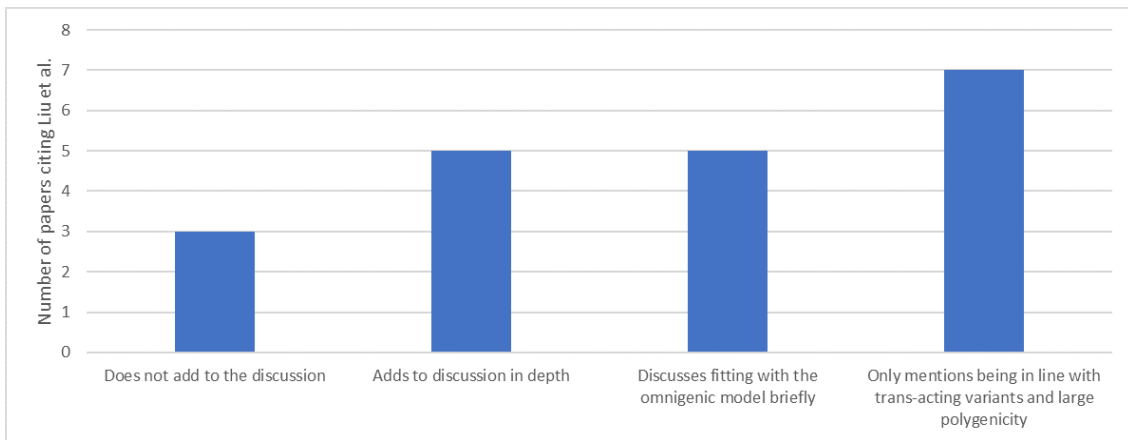
Figure S1: Bar graph of the subjective classification of 20 papers that cite the second iteration of the omnigenic model by Liu et al. [2]. From left to right: Papers that do not add to the discussion of the validity of the omnigenic model; Papers that add to the discussion by a more in depth analysis of the omnigenic model and its accordance with their data or model, or proposing different views on the model, or different models overall; Papers that briefly mention that their results are in line with the omnigenic model in general ; Papers that only mention that their results are in line with a high number of peripheral genes or a detection of trans acting variants