UNIVERSITEIT
GENT

FACULTY OF SCIENCES

Deciphering the evolutionary impact of gen(om)e duplications through mechanistic modeling of the genotype-phenotype map

Jayson Gutiérrez Betancur

Promotor: Prof. Dr. Ir. Steven Maere

Ghent University
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
VIB Department of Plant Systems Biology
Evolutionary Systems Biology Lab

Dissertation submitted in fulfillment of the requirements for the degree of Doctor (PhD) in Sciences, Biochemistry and Biotechnology.

Academic year: 2014-2015

Examination Committee

Prof. Dr. Geert De Jaeger (chair)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium

Prof. Dr. Ir. Steven Maere (promotor)

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium

Prof. Dr. Ir. Jan Fostier

Faculty of Engineering, Department of Information Technology, Ghent University, Belgium

Prof. Dr. Koen Geuten

Faculty of Science, Department of Biology, Katholieke Universiteit Leuven, Belgium

Prof. Dr. Olivier Martin

UMR de Geńet́ique Veǵet́ale, INRA, University of Paris, France

Dr. Rolf Lohaus

Faculty of Sciences, Department of Plant Biotechnology and Bioinformatics, Ghent University

Acknowledgements

Over the last few years, the journey I have been embarked on was possible thanks to a wonderful crew that provided me of guidance and support not only from an academic point of view but also from a more family-oriented perspective. Therefore, I am using this opportunity to express my deepest appreciation to all those who encouraged me and provided me the possibility to complete this thesis. First of all, I am deeply thankful to my Promotor, Professor Dr. Steven Maere for his aspiring guidance, invaluably constructive criticism and friendly advice, as well as for his truthful and illuminating views on a number of issues related to this project. Secondly, I'm greatly indebted to Prof. Dr. Geert De Jaeger, Prof. Dr. Jan Fostier, Prof. Dr. Koen Geuten, Dr. Rolf Lohaus and Prof. Dr. Olivier Martin for serving as my committee members, for your remarks, invaluable comments and positive criticism. I would also like to thank the impressive and efficient group of professionals behind the Vlaams Supercomputer Centrum (VSC), whose help has been critical for the realization of this work. A big thank you as well to Krzysztof Wabnik, a great colleague and friend from whom I received so much great feedback and advice.

A special thanks to my family. Words cannot express how grateful I am to my brother, my mother, my father and my sister, for all of the sacrifices that youve made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank my friends, Stephane Rombauts, Rahul Bhosale, Yadira Olvera, Mauricio Quimbaya and Lorena López-Galvis, David Velázquez, Jeison Fernández and Carlos Genes, who supported me in writing, and incented me to strive towards my goal. At the end, I would like to express my great appreciation to my partner Deborah Nicolescu, who spent sleepless nights with and was always my support in the moments when there was no one to answer my queries.

To all of you, a big thank you!

Table of Contents

List of Figures

List of Tables

"Nothing in biology makes sense except in the light of evolution ..."

Theodosius Dobzhansky

1

Research Purpose and Scope

1.1 Overview

Unlike human-made systems, biological systems are the outcome of a complex evolutionary process spanning millions of years. However, both engineered systems and evolved biological systems possess a given structural design that allow them to carry out particular functional tasks. In the life sciences, this structure-function relationship is better known as the genotype-phenotype map (GPM), and it is at the core of almost any biological problem known. What makes this biological mapping problem difficult to grasp is that it is intrinsically linked to the functioning of molecular interacting systems (*e.g.* the regulatory networks controlling the expression of cellular phenotypes). These molecular networks are complex because they operate in a non-linear manner, which makes the study of their functional and evolutionary properties by mere intuitive reasoning virtually impossible. The inter-disciplinary field of systems biology offers an ample range of quantitative tools to study the emergent properties of molecular interacting systems at different levels of granularity. In particular, systems biology-inspired modeling approaches have been increasingly used not only to decipher the inner workings of molecular networks, but also to shed light on the evolutionary origin of emergent systems properties such as evolvability, robustness and modularity. Although systems biology-inspired network models can often account for mechanistic details, an essential component that is still missing in these models is an explicit representation of the genetic encoding of molecular networks. In this sense, these models are of limited use to adequately study the evolutionary potential of molecular networks, essentially because the functional properties of GRNs are typically acquired through the gradual accumulation of discrete changes in their genotypic encoding over the course of evolution. This is one of the main motivations of this thesis, the design of an adequate and sufficiently detailed mechanistic modeling framework to simulate the GPM of a particular type of molecular networks, namely gene regulatory networks (GRNs). To develop this GPM model I built upon first principles to capture essential molecular mechanisms of transcriptional regulation, which allow to adequately accounting for the dosage sensitive nature characteristic of real GRNs. A critical step in the design of this model was the incorporation of a realistic genetic encoding so that the regulatory wiring of GRNs could be evolved *in silico* via the accumulation of point mutations in *cis*-regulatory regions (*i.e.* gene promoters) and *trans*-acting elements (*i.e.* the DNA binding domains of transcription factors).

Using the fine-grained, mechanistic GPM modeling framework outlined above, I concentrated on a specific biological problem that has remained largely understudied, and that is the immediate and long-term impact of gene and genome duplications on GRNs, both major sources of genetic novelty, and of special relevance to plant evolution. Concretely, a major aim was to investigate the impact of genome duplications on the evolvability of GRNs. Evolvability, the internal disposition to vary in the face of genetic perturbations, determines a systems's potential for future evolutionary change (adaptations). Evolvability is a defining feature of biological systems that has been the subject of much research over the last decades, the mechanistic basis and evolutionary origin of which remains quite controversial, mainly because a standard way to quantify it is lacking at this point. In fact, the definition and quantification of evolvability may be more system-specific. Therefore, the operational definition of evolvability adopted in this work refers to the capacity of artificial GRNs to evolve novel and increasingly better adapted expression phenotypes. Importantly, evolvability is quantified using dedicated fitness functions that assess the functional performance of the GRNs. To compare the evolvability of pre- versus post-duplication GRN system configurations, evolutionary explorations across sequence space were performed *in silico*, using a novel evolution protocol, to mimic the adaptation of GRNs. More precisely, GRNs were evolved toward newly imposed optima defined as oscillatory expression phenotypes with lower or higher frequency compared to the oscillatory expression phenotype of start (ancestral) configurations.

Next, the internal disposition of biological systems to respond to genetic perturbations is clearly epitomized by the dosage dependent functioning of transcriptional regulatory systems. Changes in the dosage balance among transcriptional regulators, achieved through gene duplication/deletion events, can substantially modulate their DNA occupancy profiles at the promoter region of target genes. In consequence, a typical outcome of dosage balance alterations is a considerable deviation in the expression patterns of key developmental genes compared to those observed under normal conditions. A clear quantitative understanding of the role of dosage balance alterations in the modulation of the expression dynamics of GRNs is currently lacking, mainly due to the fact that most present-day network models fail to capture essential mechanistic details of transcriptional regulation. Based on the mechanistic GPM modeling framework outlined above, the other major goal of this work was to investigate the proximate and ultimate consequences of dosage balance effects in GRNs. Concretely, I examined the impact of gene copy number variation, including single gene duplication and deletion, as well as amplification of regulatory gene copies, on the expression dynamics of GRNs. In addition, I simulated the evolution of GRNs carrying an extra copy of either a regulatory or an output gene, and examined the immediate fitness impact of dosage balance effects on the capacity of GRNs to evolve toward newly imposed optima.

Overall, the work presented here reveals an unanticipated complexity underlying the evolutionary potential of pre- and post-duplication GRNs. Specifically, we found a complex interplay between initial evolutionary conditions determined by genetic and non-genetic factors, such as the underlying structure of a start GRN genotype (genetic background), dosage balance effects, the nominal values of (partly) environmentally determined network control parameters, as well as quantitative aspects of the newly imposed phenotypic optima, which can severely constrain the adaptation of pre- and post-duplication GRN system configurations. The take home message of this work is that the evolvability of complex molecular networks possesses an intricate multifactorial basis that can be difficult to dissect through coarse-grained mathematical representations of the GPM.

Just as any endeavor promulgated by the emerging field of evolutionary systems biology, the computational work presented here aims toward a multidisciplinary approach to achieve a systems-level understanding of the evolution of biological systems, by emphasizing on the inner workings (*i.e.* operative rules) of molecular networks that mediate complex genotype-phenotype relationships at the cellular level. From a practical point of view, understanding how to reduce the mismatch between the current phenotypes expressed by a given biological system and those phenotypes that would be best suited for a given environment is directly linked to our ability to rationally design and optimize biological functions, the primary goals of many research disciplines in the life sciences, such as synthetic biology, plant biotechnology (*e.g.* applied to crop design), metabolic engineering, evolutionary medicine, microbiology, etc. Moreover, gaining insights into the mechanistic underpinnings of cellular information processing networks is crucial to understanding the origin of complex diseases such as cancer, which is itself the result of an intricate evolutionary process operating on somatic cells within tissues, whereby natural selection acts upon the phenotypic variation generated by the accumulation of genetic, genomic and epigenetic alterations, as well as upon phenotypic changes brought about by the inherent stochasticity of biochemical reaction systems (*i.e.* gene expression noise). Because of the complexity underlying the origin of cancer cell phenotypes, evolutionary systems biology approaches, together with molecular and cell biology experimental techniques could, for instance, aid in the design of effective drug therapies. Finally, from a basic research point of view, understanding the inner workings of molecular networks, and how evolution steers in them and shapes them at the same time, is crucial to address long-standing evolutionary questions, such as the origin of species diversity, the evolution of biological complexity, phenotypic innovation and survival of mass extinction events. In this sense, evolutionary systems biology offers the opportunity to recreate past evolutionary events, to reconstruct evolutionary trajectories and to assess their repeatability under similar conditions, as well as to assess how different starting conditions could impact on the outcome of evolution.

The structure of this thesis is as follows: Chapter 2 provides an overview on general aspects of gene regulation, genotype-phenotype mapping problems and GRNs, evolutionary aspects of GRNs, as well as features and the scope of conventional network modeling approaches. Chapter 3 provides an overview of current systems biology-inspired GPM modeling approaches used to study the evolutionary origin of emergent properties of gene regulatory networks (GRNs), discusses in detail their limitations to investigate GRN evolution under gene and genome duplication events with a special focus on plant evolution, and outlines the essential features that a GPM modeling approach should incorporate in order to adequately explore GRN evolutionary issues. Chapter 4 provides a detailed explanation of the mechanistic GPM modeling framework developed in this work, discusses the limitations and potential novel features that could lead to further model enhancements, as well as the methods used to simulate the evolution of GRNs across sequence space. Chapter 5 presents the research study that focuses on the impact of whole genome duplications on the evolvability of prototypical cascade-like GRNs with oscillatory expression phenotypes. The consequences of dosage balance effects on the evolution of GRNs toward a new phenotypic optimum, as well as the immediate impact of gene copy number variation, including single gene duplication/deletion and amplification of regulatory gene copies, on the expression dynamics and fitness of GRNs is addressed in chapter 6. Finally, chapter 7 closes with a general discussion and future perspective.

“The way to get good ideas is to get lots of ideas, and throw the bad ones away”

Linus Pauling

2

Introduction

2.1 On the structure of the genotype-phenotype map: regulatory networks and evolution

Complex genotype-phenotype mapping (GPM) problems, that is the association between the genetic blueprint (genotype) of a given biological system (*e.g.* a molecular network) and its phenotypic manifestation (*e.g.* the time varying concentration of gene products, or their biochemical activity), critically depend on the functioning of a cellular machinery constituted by multiple regulatory layers, wherein gene regulation plays a critical role. In particular, gene regulatory networks (GRNs) have been shown to be responsible for imparting precise control on the expression of the distinct transcriptional programs underlying, for instance, physiological responses to changing environments, as well as the spatio-temporal organization (development) of multicellular organisms. Because of the pivotal role of GRNs in the generative process of the phenotypes, changes in their genotypic structure have provided important raw material for evolution to act upon. In this chapter, I first present a brief history of the notion of the GPM problem, and then I will give a brief overview on general aspects of gene regulation, the structure and evolution of GRNs, as well as important conceptual and theoretical approaches, such as generic and system-specific mathematical network models, used to study a great variety of GPM problems in GRNs, including the evolutionary origin of emergent system properties. I will also briefly discuss several bottlenecks typically faced by network modeling studies.

Author contribution

All content within this chapter was written by myself and revised by professor Steven Maere

2.2 Genotype-phenotype mapping problems: a brief history

Although Darwin failed to adequately account for the laws of inheritance in his theory of evolution by natural selection[1], he nevertheless could anticipate the importance of understanding the generative processes underlying the variation of characters. The rediscovery of Mendel's rules of heredity by the botanist Hugo DeVries led the latter to suggest the concept of genes in his book *Intracellular Pangenesis* (1889). Then, at the beginning of the twentieth century, Wilhelm Johannsen made the distinction between the hereditary dispositions of organisms (their genotypes) and the ways in which those dispositions manifest themselves in the physical characteristics of organisms (their phenotypes)[2]. Few years later, the field of genetics was formally established upon the publication of the seminal book: *The Mechanism of Mendelian Heredity* by Morgan, Sturtevant and Bridges in 1915[3], which laid down the foundations for a chromosomal theory of heredity. After this great accomplishment, the generative process of biological functions begun to receive renewed attention. It was only after Conrad H. Waddington proposed his influential metaphor on the epigenetic landscape of cellular differentiation pathways[4] that a more systemic approach begun to emerge to explain the dynamics of developmental and evolutionary processes. Then, the seminal work on gene regulation by Jacob and Monod[5], perhaps the first rigorous attempt to describe a genotype-phenotype mapping (GPM) problem (*i.e.* the association between the genotype and its biological manifestation or phenotype) in mechanistic terms, inspired the famous theoretical biologist Stuart Kauffman to develop a Boolean model of large sets of interacting gene nets to explore the global dynamics of cellular differentiation pathways[6]. A few decades ago, the foundation for a dynamical systems approach to developmental mechanisms was laid down by Oster and Alberch[7,8], who aimed to provide explanatory principles of the evolvability (capacity to evolve) of multicellular organisms. More recent conceptual developments have formally incorporated environmental factors as key modulators of GPM problems in multicellular organisms [9,10]. Together, these seminal contributions brought more questions than answers on the structure and evolution of the generative processes of the phenotypic characteristics of organisms. With the advent of the omics revolutions, the collection of massive amounts of molecular data held the promise of deciphering the secrets of life[11]. Inspired on a rich tradition of systems approaches in engineering, network biology[12] and systems biology[13] have laid the foundations of powerful quantitative frameworks that allow us to interrogate the emergent properties of complex molecular interacting networks. Relying on both integrative data analysis and mechanistic modeling approaches, remarkable advances have been made over the last decade towards the elucidation of the inner workings of distinct GPM problems in many different organisms[14–17]. Not surprinsingly, what is at the heart of most GPM problems is gene regulation, the basis of which will be briefly described below.

2.3 Gene regulation: a defining feature of living organisms

Compared to prokaryotes, eukaryotic genomes typically have many more genes with complex internal structures (see Figure 2.1), which are distributed across multiple chromosomes, whereas genes in prokaryotes are organized into one single chromosome[18]. A distinctive feature of prokaryotes is that functionally related genes are generally organized into operons, which renders the regulation of the expression of prokaryotic genomes relatively less complex than in eukaryotes, where a greater variety of elaborated gene regulation mechanisms is necessary (see[18]). Gene regulation is a complex multifaceted process involving a dynamic interplay between the synthesis and the degradation of gene products[18,19].

Transcriptional control is an essential part of gene regulation, which is itself a multilayered process involving a suite of molecular complexes that bind to the promoter regions of the genes[20–22]. Transcriptional control in eukaryotes, as compared to prokaryotes, is known to be a much more elaborated process encompassing, for instance, structural aspects involving the remodeling of the chromatin structure (a tightly packed fiber composed of DNA and histone proteins), which renders access to the DNA by the RNA polymerase quite restrictive, as opposed to the non-restrictive accessibility of DNA in prokaryotes[21–24]. Furthermore, transcriptional control in eukaryotic cells involves several phases[25,26], the most critical ones being transcription initiation[22] and RNA polymerase translocation[26], which are driven by different classes of transcription factors (TFs) (see Figure 2.2) that act in a combinatorial fashion. For instance, the Pol II transcription-complex is composed of general TFs; another class of TFs is involved in DNA remodeling tasks; finally co-activators and co-repressors, another type of TFs, are essential in mediating the regulatory effects imparted by sequence-specific DNA binding TFs known as activators and repressors[20–22]. Several lines of evidence indicate that transcriptional activators and repressors are largely responsible for fine tuning the concentration levels of gene products across temporal and spatial scales. For instance, it has been shown that complex cascades of gene expression (transcriptional programs) are set in motion through the combined activity of activator and repressor TFs that bind in a combinatorial fashion to a collection of *cis*-regulatory sequences (DNA binding sites) scattered across the genome[27,28]. The resulting temporal progression of gene expression states is a major determinant of, for instance, physiological responses to environmental stresses in unicellular organisms[29], cell fate determination[30] and the formation of animal body plans[31,32], as well as the coordination of plant developmental switches[33]. In fact, this gene regulation-centered paradigm has proven instrumental in explaining a wide range of GPM problems[33–41], and has served as a solid foundation for the elaboration of enticing hypotheses that aim to explain the evolution of biological diversity and complexity in terms of changes in the structure of gene regulatory networks[32,42].

A distinctive feature of the components involved in transcriptional regulation is their modular architecture, that is, the presence of independent or individualized units (*e.g.* *cis*-regulatory elements, or proteins motifs) that perform specific regulatory tasks. Such modular organization provides flexibility in the way in which transcriptional regulation is achieved, for instance, individual modules can be replaced, added or deleted without affecting the proper function of the rest of the system[43]. A typical example of modular transcriptional regulation comes from the fruit fly, *Drosophila melanogaster*, where specific transcriptional readouts are controlled by dedicated enhancer sequences (typically spanning over 200-500 base pairs) located in the promoter regions of genes, which have the capacity to control particular phases (e.g. spatial-temporal) of gene expression[36,44]. Interestingly, the finding that enhancers contain binding sites that are usually distributed in a non-random manner has led to the proposition of a regulatory grammar[36,44], a set of quantitative parameters associated to the activity of DNA-bound activators and repressors (*e.g.* stoichiometry, affinity, spacing, and arrangement of binding sites) that together determine the functional properties of enhancer sequences[36,44]. Although insights into the extent of modularity of plant *cis*-acting elements are still limited, a few studies have shed light on what could be the regulatory grammar of stress-responsive *cis*-acting elements. For instance, due to their sessile life style, plants must display plastic physiological and/or morphological responses to stressful environments[45], which are achieved at the molecular level through extensive changes in the transcriptome (transcriptional reprogramming). Only a handful of *cis*-regulatory sequences had been linked to stress-induced transcriptional reprogramming[46,47], but a recent study has reported a series of putative *cis*-regulatory sequences in the *A. thaliana* genome[48]. Most interestingly, this study proposed a series of *cis*-regulatory codes,

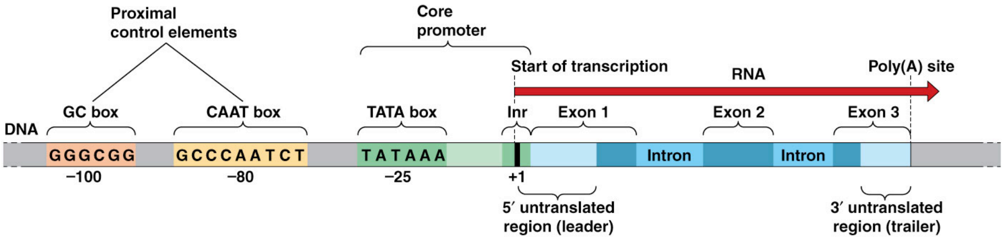


Figure 2.1: Schematic representation of a typical eukaryotic gene structure. *In a typical protein-coding eukaryotic gene, the mRNA is transcribed by RNA polymerase II. The core promoter is characterized by an initiator sequence surrounding the transcriptional startpoint and a sequence called a TATA box located about 25 bp upstream (to the 5 prime side) of the startpoint. The core promoter is where the general transcription factors and RNA polymerase assemble for the initiation of transcription. Within about 100 nucleotides upstream from the core promoter lie several proximal control elements, which stimulate transcription of the gene by interacting with regulatory transcription factors. The number, identity and location of the proximal elements vary from gene to gene. The transcription unit includes a 5 prime untranslated region (leader) and a 3 prime untranslated region (trailer) which are transcribed and included in the mRNA but do not contribute sequence information for the protein product. These untranslated regions may contain expression control sequences. In the primary transcript, at the end of the last exon is a site directing the cleavage of the RNA and poly(A) addition. Figure reproduced from: Principles of Cell Biology, Dr. Brian E. Staveley's Lectures, url: http://www.mun.ca/biology/desmid/brian/BIOL2060/CBhome.html.*

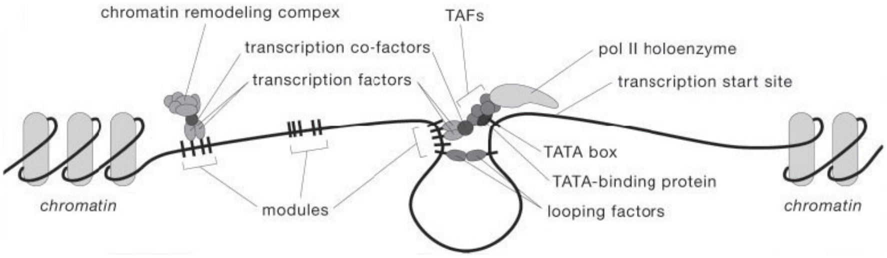


Figure 2.2: Eukaryotic promoter structure. *The figure illustrates an idealized gene promoter in operation. The Initiation of transcription requires several dozen different interacting proteins, including the RNA polymerase II holoenzyme complex (*$\sim$ 15 *proteins); TATA-binding protein (TBP); TAFs (TBP-associated factors, also known as general transcription factors;* $\sim$ 8 *proteins); transcription factors (the precise composition and number of which differs among target genes regulated, which may vary in space and time and according to environmental conditions); transcription cofactors (again, the precise composition of which is variable); and chromatin remodeling complexes (which can contain a dozen or more proteins). Figure reproduced from Wray et.al.*[22].

such as combinatorial relationships among *cis*-regulatory elements, location and copy number, by which stress-responsive gene expression could be modulated[48]. Although not strictly similar to the regulatory grammar of animal *cis*-regulatory regions, this study suggests that modular *cis*-regulatory activity is also common in plants.

Similar to *cis*-regulatory sequences, TFs possess a modular structure[43] consisting of several dedicated domains that carry out specific regulatory tasks. For instance, DNA binding domains (DBDs) are composed of amino acids that are either contiguous (e.g. MADS box) or dispersed (e.g. Zn-fingers) within the primary sequence[22]. The number of distinct DBDs is believed to range between 200-300, according to structural considerations[49]. These DBDs can be allocated into 8 broad structural classes distinguishable by their DNA binding mode[50], with Helix-Turn-Helix and Zinc-coordinating DBDs being the most abundant ones[49]. TFs have also been classified according to several lineage-specific DNA binding domain families (see Figure 2.4). In addition to the DBD, the presence of protein-protein interaction domains is an important structural requirement to mediate context-dependent interactions with other proteins, which are necessary to carry out specific regulatory tasks. Many eukaryotic TFs bind DNA only as homo or heterodimers[51]. An important class of TFs that make extensive use of protein-protein interaction domains to carry out crucial developmental tasks across eukaryotes, and specially plants, is the MADS-box gene family[52]. In particular, the MIKC-type proteins, a special class of MADS domain proteins only present in plants, are multi-domain proteins[53] able to bind DNA as homo- or heterodimers, or even as part of higher-order complexes. In this type of TFs dimerization and higher-order protein-protein interactions are established mainly through its so-called I and K-domains[53]. It has been hypothesized that the acquisition of this capacity to form multimers has conferred on the MIKC-type proteins the capacity to exercise more sophisticated transcriptional control of important developmental genes, and that this combinatorial feature might have facilitated the evolution of more complex developmental plant systems[53]. In addition, the regulatory status of TFs (activating vs. repressing regulatory activities) is thought to be specified by the presence of dedicated domains that mediate interactions with the RNAp II complex. Activation domains, which are usually composed of Gln, Pro, and/or Ser/Thr residues, are thought to exert their function by increasing the frequency with which the RNAp II complex initiates transcription[54]. Activation domains have also been found to mediate direct interactions with the TATA-binding protein (TBP), or indirect interactions via TAF (TBP-asssociated factors)[55]. In contrast, repression domains, many of which have been found to be composed of acidic residues, are thought to exert their function by decreasing the frequency at which the RNAp II complex initiates transcription[22]. Finally, despite the existence of distinct protein domains conferring on the TFs with a molecular identity, the fact that eukaryotic transcriptional control is achieved through complex combinatorial interactions usually renders the functional role of TFs highly context-dependent. For instance, the regulatory role associated to a given TF can vary as a function of the partners it interacts with, usually resulting in synergistic effects (i.e. combined effect is stronger than an individual effect) on the expression of a target gene (see Wray *et al.* [22] and references therein). Finally, a TF may act as a repressor if it masks the binding site of a transcriptional activator, an effect that does not require a specialized repressor domain [22].

Over the past few years, considerable efforts have been made toward deciphering quantitative aspects of protein-DNA binding interactions[57–59]. An important insight gained from these studies is that the interaction of a TF with a DNA site seems to be largely determined by a position-unspecific attraction and a specific interaction, whose energy values are thought to depend on the particular DNA binding sequence where the factor binds[59]. The unspecific part is the electrostatic interaction between the positively charged

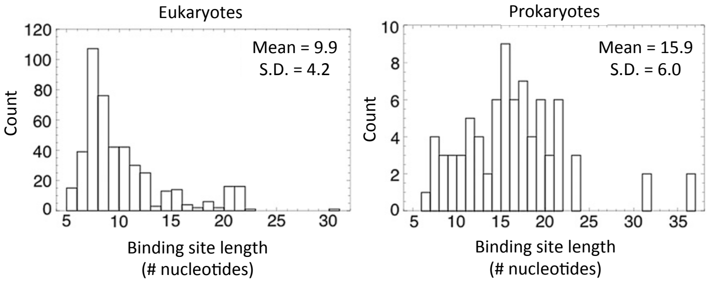
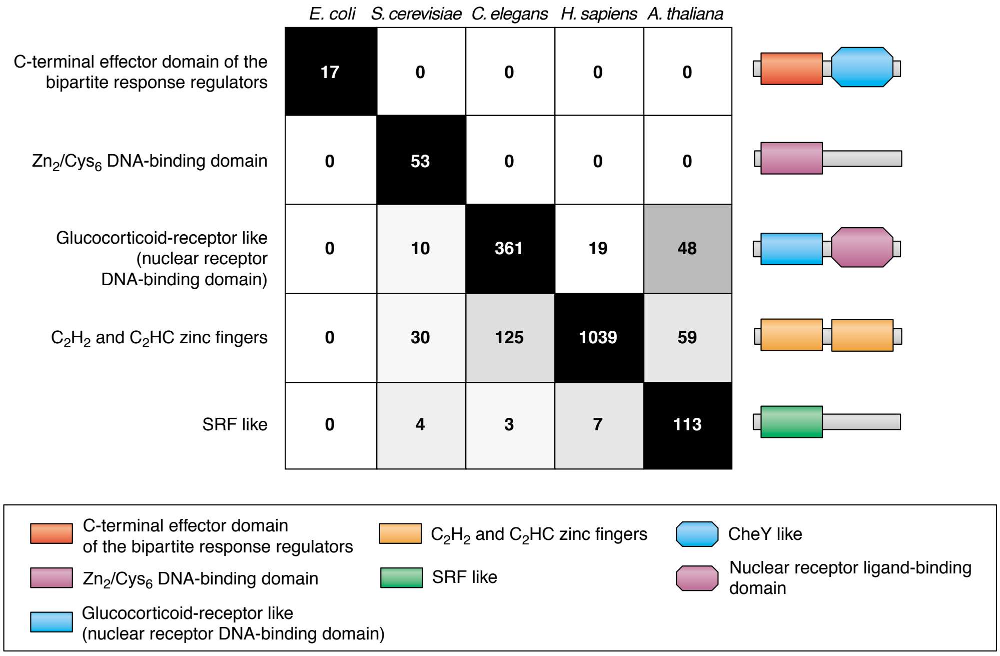


Figure 2.3: Distribution of the length of binding sites in eukaryotes and prokaryotes. *The lengths range from 5 to* ~ 30 *nucleotides in both eukaryotes (data shown for 454 curated DNA binding motifs) and prokaryotes (data shown for 79 curated DNA binding motifs). Figure reproduced from Stewart et.al.*[56].

protein and the negatively charged DNA backbone, while the specific part involves hydrogen bonds between the DBD of the TF and the nucleotides of the binding site. Recently, it has been hypothesized that animal transcription networks are likely to behave as highly connected, quantitative continua, in the sense that most TFs, by virtue of having high intracellular concentration levels (*e.g.* in the order of $10^4 - 3 * 10^5$ molecules per nucleus), are able to bind DNA sequences over a quantitative series of DNA occupancy profiles[60]. In other words, this implies that such high intracellular concentration levels would be sufficient to thermodynamically drive TFs to reside on DNA most of the time; at any instant, some molecules of each TF in vivo should be bound to any accessible DNA sequence by means of electrostatic, sequence-independent interactions (with $K_D \sim 10^{-6}$ M), and other molecules by sequence-specific interactions with many of tens of thousands of moderate and high-affinity recognition sites ($K_D < 10^{-8}$)[60]. Furthermore, it has been determined that several biophysical factors may impose some constraints on the length and information content of DNA binding motifs[57,58,61]. Interestingly, it has been determined that binding sites in both eukaryotic and prokaryotic genomes are typically 10 nucleotides long (see [56], and references therein), with lengths varying between 5 to $\sim$ 30 nucleotides (Figure 2.3). Moreover, it has been determined that the information content of binding sites, defined as the number of different bases that can occur at each nucleotide and still produce functional binding, may vary from a maximum of 2 bits per nucleotide (*i.e.* each nucleotide must assume a specific base to produce functional binding) to < 0.25 bits (*i.e.* each nucleotide can assume one of several bases and still produce functional binding) [62]. Next to biophysical constraints, natural selection may also place important constraints on the protein-DNA binding process due to the fact that TFs must correctly bind to some sites in the genome and avoid binding elsewhere in order to trigger appropriate transcriptional outputs ([63,64]). If binding sites are too short, TFs are predicted to bind too readily to non-desirable genomic sites, which may disrupt gene expression and deplete the pool of TFs available to bind where they are required. If binding sites are too long, on the other hand, sites where binding must be favored would tend to be too easily disrupted upon mutations[56,63,64].

2.3.1 Hierarchical organization of gene regulatory networks

As previously mentioned, transcriptional regulation is carried out by a complex gene regulatory network (GRN). At this level of organization, GRNs have been shown to exhibit an intrinsic hierarchical orga-



	E. coli	*S. cerevisiae*	*C. elegans*	*H. sapiens*	*A. thaliana*
C-terminal effector domain of the bipartite response regulators	17	0	0	0	0
Zn_2/Cys_6 DNA-binding domain	0	53	0	0	0
Glucocorticoid-receptor like (nuclear receptor DNA-binding domain)	0	10	361	19	48
C_2H_2 and C_2HC zinc fingers	0	30	125	1039	59
SRF like	0	4	3	7	113

C-terminal effector domain of the bipartite response regulators
C_2H_2 and C_2HC zinc fingers
CheY like
Zn_2/Cys_6 DNA-binding domain
SRF like
Nuclear receptor ligand-binding domain
Glucocorticoid-receptor like (nuclear receptor DNA-binding domain)

Figure 2.4: Examples of lineage-specific DNA binding domain families. *Figure reproduced from Babu et.al.*[65].

nization that allows their study at different levels of granularity[65]. For instance, using graph theoretical approaches interesting statistical regularities on the global structural features of the yeast transcriptional network have been identified[66]. Specifically, it has been found that the number of TFs regulating a target gene (incoming connectivity) follows an exponential distribution, indicating that most target genes tend to be regulated by small sets of TFs. The distribution of the outgoing connectivity, defined as the number of target genes regulated by each TF, which is indicative of a of a hub-containing network structure wherein only a few TFs take part in the regulation of a disproportionately large number of target genes. At an intermediate level, regulatory modules composed of a small number of interacting genes have been identified through graph theoretical analysis showing a high propensity for the nodes in *e.g.* the yeast GRN to form ’cliques’[66]. Although no consensus seems to exist as to the general properties of such modules, what seems to be clear is that GRNs are highly interconnected and very few modules tend to be entirely separable from the rest of the network. In fact, many identified modules are nested within each other in a hierarchical organization at differing levels of connectivities[65]. At the smallest scale, small recurring cross-regulatory patterns between TFs and target genes are thought to be the building blocks of the global GRN[67–69]. Network motifs have been attributed key roles in the dynamic behavior of GRNs, such as their potential to carry out specific information processing tasks related to the regulation of temporal gene expression patterns in response to external signals[70] (see Figure 2.5). Although the study of GRNs at different scales has proven instrumental in shedding light into the organization of complex GPM problems at the cellular level, we are still lacking a more detailed mechanistic understanding on how GRNs function. In particular, the description of eukaryotic GRNs as directed graphs provides only a rather abstract representation of their real regulatory structure, given that a simple edge-node interaction in a graph may be translated into multiple protein-DNA binding events whose regulatory activities can be context-dependent (*e.g.* DNA binding competition with other TFs).

2.4 Evolution of gene regulation: a major driver of phenotypic innovation, adaptation and biological complexity

Changes in gene expression achieved through alterations of transcriptional regulation patterns have long been considered as the major source of phenotypic evolution[22]. This is supported by a great body of evidence suggesting that much of the remarkable phenotypic diversity observed across different species has been achieved through the gradual accumulation of changes in gene regulation[38,42,72–75]. For instance, in a classic study, King and Wilson[76] compared the levels of morphological and protein divergence between humans and chimps and concluded that the level of protein divergence was too small to account for the anatomical differences between these two species. To reconcile the level of divergence between proteins and morphology, they proposed that morphological divergence was based mostly on changes in the mechanisms controlling gene expression and not changes in the protein-coding genes themselves. A similar observation was reached for plant systems a few years ago, where Doebley and Lukens[77] after summarizing a series of evidence concluded that the evolution of plant form is most readily accomplished by changes in the *cis*-regulatory regions of transcriptional regulators. Several hypotheses have been proposed to explain the pervasive role of regulatory evolution as a major source of the diversity of life[38,42,72–75]. An emerging consensus indicates that the inherent modular organization of transcriptional regulation endows GRNs with such a vast evolutionary potential that even small discrete changes may account for the relatively modest phenotypic differences observed among closely related species, as well as the markedly different phenotypes observed among groups at higher taxonomical levels[38,42,72–75]. This idea is further supported by simulation studies showing that binding sites can arise rapidly from random sequences[78] due to their low information content (*i.e.* because they are short, typically 10 bp long; see Figure 2.3), thus making them ideal as a source of genetic variation. Whether morphological and physiological traits have been assembled over evolution predominantly through changes in *cis*-regulatory sequences or substitutions in *trans*-acting elements has ignited an intense debate, with the argument in favor of the *cis*-regulatory hypothesis focusing mainly on the prediction of strong conservation of TF function[38,75]. Nevertheless, a growing body of evidence seems to suggest that changes in the coding sequences of TFs may have also contributed a significant fraction to the within and between species gene expression diversity and divergence[79–83].

Although a great deal of information exist regarding the association between changes in *cis* and *trans*-acting sequences and their effects on expression phenotypes, we still lack a comprehensive understanding of how such sequence changes may have impacted on the structure and the functioning of GRNs over evolutionary time scales. Nevertheless, promising advances have been made over the last decade toward deciphering potential evolutionary principles in model organisms. In particular, using a combination of global gene expression profiling, genome-wide chromatin immunoprecipitation (ChIP), DNA sequencing, and bioinformatics analyses, it has been possible to characterize GRNs involved in a great variety of biological processes in fungal species, such as ribosomal gene expression, galactose metabolism, amino acid biosynthesis, cell-cycle control, and cell-type control[84–87]. A general observation in these studies is that GRN rewiring has taken place via: 1) sequence turn-over of *cis*-regulatory regions, which create

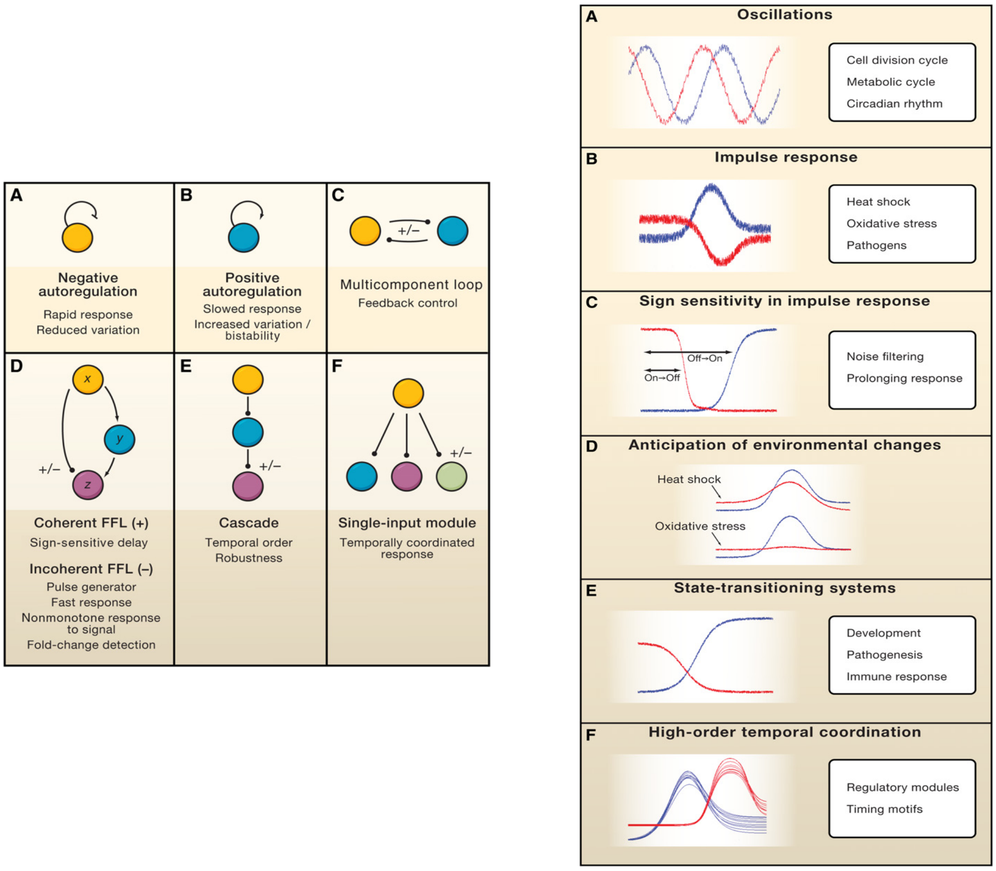


Figure 2.5: Network motifs and expression dynamics. *The panel on the left illustrates minimal regulatory schemes (network motifs) found in genome-wide transcriptional regulatory networks*[67,70]*. The schemes represent in a very abstract manner transcriptional regulators (shown as nodes) that control the expression of target genes by binding to their promoter regions. Regulatory interactions can be either activating (arrows) or repressing (blunt arrows). Circle-ending arrows are either activating or repressing. The panel on the right depicts prototypical patterns of temporal dynamics of gene expression generated during different cellular and developmental processes in response to stimuli. These temporal patterns of gene expression can be generated by some of the network motifs individually, or through their combined action. Figures reproduced from: Yosef and Regev*[71] *.*

and destroy binding sites; 2) protein-coding changes that either alter the binding specificity of a transcription factor or change its interaction with other co-factors; or 3) through the combined action of *cis* and *trans*-regulatory changes[85,86] (see Figure 2.6). Although the relative contributions of these principles to the rewiring of GRNs have been extensively characterized in fungal regulatory circuits, these are likely applicable to most, if not all, GRNs. For instance, as previously mentioned, the regulatory connectivity of plant GRNs composed of MIKC-type proteins may have evolved mainly through sequence divergence (*e.g.* protein-coding changes) between gene duplicates, by which TF paralogs may have partitioned their ancestral regulatory role (subfunctionalized) or acquired distinct functional roles (neofunctionalized) within the GRN context. In particular, divergence in protein-coding sequences between MIKC-type TF paralogs is thought to have represented a major driving force in the evolution of higher order molecular complexes by which more precise transcriptional regulation of target developmental genes may have been exercised[53]. Together, these principles of GRN rewiring underscore the importance of the modular and combinatorial nature of transcriptional control as major contributors to the evolutionary origin of biological complexity. What is yet not clear is the potential adaptive benefit of GRN rewiring, as the phenotypic output of many homologous GRNs compared so far have been found to be remarkably similar[84–87]. Given the lack of conclusive evidence on the potential fitness impact of GRN rewiring events, it is thus plausible that many of the cases reported so far have been the outcome of neutral changes[88], where evolutionary transitions may have usually taken place through intermediate states representing redundant regulatory programs with similar adaptive value[86]. In this sense, network modeling approaches may help clarify the potential adaptive role of GRN rewiring events achieved through *cis* and *trans*-regulatory changes, in addition to functional divergence between TF paralogs subsequent to gen(om)e duplication, which is the primary source of novel genetic material.

2.5 The role of gen(om)e duplications in evolution

Small and large-scale duplications, such as single gene duplication (SGD) and whole genome duplication (WGD), have long been recognized as a prominent factor in evolution[89–92], mainly because they provide novel genetic material for mutation, selection and drift to act upon. Over 40 years ago Susumu Ohno famously stated that without duplicated genes the creation of metazoans, vertebrates, and mammals from unicellular organisms would have been impossible, and that such big leaps in evolution required the creation of new gene loci with previously nonexistent functions[89]. In fact, it seems difficult to imagine, for instance, how the vertebrate adaptive immune system (with dozens of duplicated immunoglobulin genes) could have evolved without gene duplication[93].

Different types of duplication events can be achieved through distinct molecular mechanisms[93]: 1) tandem gene duplications (a variable number of duplicated genes that are linked in a chromosome) are usually generated through unequal crossing over; 2) duplicated genes that are usually unlinked to their original genes are generated via retrotransposition, which occurs when a messenger RNA is retrotranscribed to complementary DNA (cDNA) and then inserted into the genome, the activity (*i.e.* expression) of which may depend on the genomic context; and 3) chromosomal/genome duplications which presumably occur by a lack of disjunction among daughter chromosomes after DNA replication. At a larger scale, polyploidization, that is the increase in genome size caused by the inheritance of an additional set (or sets) of chromosomes, may originate from the same or a closely related individual (autopolyploid)

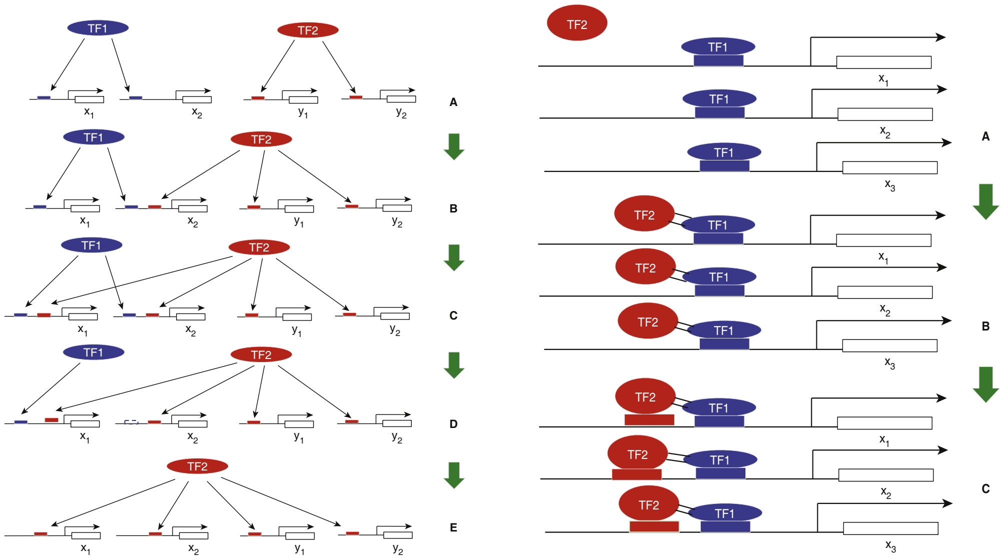


Figure 2.6: GRN rewiring events. *GRNs can undergo rewiring events over evolutionary time scales via turnover of cis and/or trans regulatory features. Panel on the left illustrates the Hand over of a regulon (collection of genes or operons under regulation by the same TF) from one TF to another. The regulation of x1 and x2 by TF1 in the ancestral circuit (A) has been taken over by a TF2 in the extant circuit (E). The rewiring may have occurred gene by gene, through intermediates with redundant regulation (BD). Panel on the right depicts the recruitment of a new TF to an existing regulon by the evolution of a new combinatorial interaction. The formation of a new interaction between TF1 and TF2 brings TF2 to the regulon controlled by TF1, effecting a concurrent rewiring of the full regulon (A,B). The new circuit can then be improved by step-wise cis-regulatory changes that stabilize the binding of TF2 to the promoters (C). Figures reproduced from: Li and Johnson*[86] .

or from the hybridization of two different species (allopolyploidy)[90]. When polyploidization involves duplicated sets of chromosomes that share homology but are sufficiently distinct due to their separate origins, these pairs of chromosomes are referred to as homeologs[90]. Moreover, segmental allopolyploids carry more than two partially differentiated genomes, which can lead to the formation of both bivalents and multivalents during chromosome pairing[94]. Major routes of polyploid formation are via gametic non-reduction and, to a lesser degree, somatic doubling[95]. In gametic non-reduction, fusion of two gametes, of which at least one contains a non-reduced, full somatic complement of chromosomes, can lead to polyploidy. Somatic doubling may occur in zygotic, embryonic or sporophytic tissue. Spontaneous genome duplication in those tissues can thus also produce viable polyploid offspring via gamete formation in the duplicated sectors[96]. Although auto- and allopolyploidy share the property of being duplicated genomes, differences in their origin and genomic compositions typically have notable consequences. For instance, in autopolyploids, chromosomes generally pair as multivalents during meiosis, while in allopolyploids bivalent pairing between chromosomes of the same original genome is prevalent[94], resulting largely in the maintenance of two separate genomes. However, the more closely genomes in the allopolyploid are related, the more likely it is for homoeologs to pair[97], resulting in chromosomal exchanges between the two genomes.

Long before whole-genome sequencing technologies were available, chromosome counts, studies of chromosome morphology, estimates of DNA content, and isozyme electrophoresis had made significant contributions to research on gene and genome duplication, mainly with regards to their prevalence across different organisms[91]. Then, the increasing availability of whole-genome sequence data over the last decade prompted a wave of comparative genomic studies and large-scale bioinformatics analyses aimed at revealing differences in genome structure and content among different species. The emerging consensus from these studies is that gen(om)e duplications have severely impacted genome architecture, driving the expansion of gene families in a great diversity of species, such as bacteria[98], yeast[99], fish[100], amphibians and reptiles[101], human[102], and plants[103]. Analyses of whole-genome sequence data from different species have also proven instrumental in shedding light into the evolutionary dynamics of gene duplicates. For instance, using whole-genome sequence data from *Arabidopsis*, rice, yeast, fly, worm, mouse and human, Lynch and Conery found that genes tend to undergo duplication at a rate comparable to the rate of nucleotide substitutions[104]. In addition, they found that the rate of duplication in humans and worms tends to be higher than in *Arabidopsis*, *Drosophila* and yeast. Relying on the pattern of nucleotide substitutions and on the frequency distribution of gene ages, an important conclusion by Lynch and Conery was that duplicates tend to experience a brief period of relaxed selection, with most duplicates becoming nonfunctional very quickly (*i.e.*, by the time silent sites have diverged by only a few percent[104]), which was inconsistent with previous analysis showing that in tetraploid species a large proportion of genes tend to be retained in duplicate. To explain this inconsistency, Lynch and Conery argued that selection might preferentially retain duplicates produced during whole-genome duplication events in order to maintain relative gene dosage[104]. Based on these observations, Lynch and Conery argued that the high rate of gene duplication may provide a substantial molecular substrate for the origin of evolutionary novelties, although the time window available for such evolutionary exploration by gene duplicates before they pseudogenize may be quite narrow. The authors also hypothesized that differential gene duplication and pseudogenization in geographically isolated populations might cause reproductive isolation and speciation[104].

Intriguingly, gen(om)e duplications have been found to be a particularly predominant force in the evolution of plant genome structure and content, especially in flowering plants where evidence suggests that genomes have undergone one and often several rounds of WGD early during their evolutionary history[105–107]. Although WGDs have received significant attention within the plant biology community, a growing body of evidence suggests that small-scale duplication events (*e.g.* segmental duplications) may have contributed equally to the repertoire of duplicated genes. For instance, it has been inferred that approximately 25% of the genes in *Arabidopsis*, the model eudicot species, are the product of ancient whole genome duplications[108], whereas nearly 16% of the genes tend to occur as tandem duplicates[109]. Interestingly, based on an evolutionary model of the duplication dynamics of genes applied to the *Arabidopsis* genome, Maere *et al.* estimated that the three WGDs this model species has presumably undergone over the last 350 My are responsible for approximately 90% of the increase in transcriptional regulators, signal transducers and developmental genes[110], which was congruent with previous observations indicating that duplicate genes with regulatory roles tend to be over-retained subsequent to WGD[111,112], presumably due to dosage balance constraints (see below).

That gen(om)e duplications have had an impact on the evolution of genome architecture does not imply, however, that they have contributed substantially to the evolution of biological diversity. In particular, the potential impact of genome duplications on patterns and rates of diversification, speciation, adaptation to novel environments and the evolution of biological complexity, remains quite controversial. In fact, diametrically opposing points of view exist as to the potential role of WGD in evolution. For instance, polyploidy has been assigned only a marginal role in progressive evolution and adaptation[113], and it has been considered as nothing more than an evolutionary dead-end[114], usually leading to extinction events[114,115]. Conversely, polyploidy has also been granted a primary creative role in evolution[116], and it has been associated with increased rates of adaptation[90], broader ecological tolerance[94], species diversification[117], and survival of mass extinction events[118]. Although WGDs have repeatedly been linked to the origin of evolutionary novelties in several organisms[119–122], it has recently been hypothesized that the evolutionary impact of WGDs may be linked rather to the elaboration of existing, primitive innovations[117,123]. Take for instance the invention of the flower, which has been followed by the elaboration of a huge variety of floral forms, specialized pollination syndromes and fruits. WGDs might not have been instrumental in developing the basic flower morphology, but rather its specialized derivations adapted to particular conditions/niches[123].

Most of the above hypotheses granting polyploids an edge in evolution over their diploid counterparts have been grounded on several biological observations linked to the immediate phenotypic/fitness effects of polyploidization, the major determinant of their successful establishment within populations. In other words, a polyploid lineage must survive long enough for evolution to act upon, and it will do so only if it is not immediately outcompeted by its diploid relatives[124]. In particular, three advantages of polyploids are often cited. Firstly, the increased number of alleles of a given gene in a polyploid may allow for the masking of deleterious recessive mutations[125], which may be advantageous under specific circumstances (*e.g.* under stable environments and thus strong stabilizing selection for an already well fitted phenotype). Secondly, the formation of allopolyploids and heterozygous autopolyploids can usually result in hybrid vigor (superior hybrid performance compared with the corresponding progenitor species), due to transgressively expressed phenotypes[126], which might confer hybrid polyploid organisms the ability to meet a broader range of environmental challenges than their progenitors. In this sense, hybrid polyploid organisms are usually thought of as being "pre-adapted" for survival in novel, often extreme, habitats, which might facilitate ecological speciation[127]. Importantly, contrary to diploid hybrids where hybrid vigor decays over subsequent generations due to homologous recombination, heterosis is stable in allopolyploids due to the predominant disomic pairing of identical homologous chromosomes[89,128]. The third major advantage of polyploids stems from the possibility that duplicated gene copies can evolve to assume new or slightly varied functions (neofunctionalization or subfunctionalization, see below), potentially allowing for ecological niche expansion or increased flexibility in the organisms responsiveness to environmental change[129]. Other often cited advantages attributed to polyploidization are higher selfing rates, reduced inbreeding depression (due to masking of deleterious alleles) and increased genetic diversity due to multiple formations of polyploids within populations[130]. Furthermore, WGDs have also been suggested to directly facilitate speciation by reciprocal gene loss, where different paralogues are lost in different populations, ultimately leading to genetic isolation and speciation of these populations[131]. By contrast, the increased number of chromosomes, and the greater complexity of their pairing and segregation interactions that can cause abnormalities (including aneuploidy) during meiosis and mitosis, is often cited as an important disadvantage that could lead to less vigor and a reduced adaptive capacity in polyploid species[128]. Additionally, the cell architecture in polyploids is altered because of generally increased cell size in polyploids, which alters the surface to volume ratio[97]. Finally, changes in polyploids that can be either advantageous or detrimental relates to altered transcriptomic profiles, genomic architecture and epigenetic factors, which can lead to gene silencing or activation[123,132,133].

2.5.1 Evolutionary fates of gene duplicates

Immediately after the formation of gene duplicates different evolutionary trajectories are possible[134], but most frequently one of the members of the newly formed duplicate gene pair is destined to be lost[135] given that the rate of deleterious mutations tend to be much higher than that of beneficial ones[136]. Intriguingly, however, many duplicate genes have been found to be retained over long evolutionary time periods[110,137]. Several models have been proposed to explain this pattern. For instance, the neofunctionalization model posits that one member of a duplicate gene pair is free to undergo changes without compromising the other gene's ancestral function[89,138], and eventually, the diverging gene copy may acquire a novel function, which can be achieved either neutrally[136] or adaptively (*i.e.* through positive selection). An alternative hypothesis, termed the subfunctionalization model, predicts that the retention of gene duplicates can be achieved through the partitioning of multiple ancestral functions among the paralogs[139,140], the outcome of which may largely depend on whether subfunctionalization takes place at the protein or expression level[140]. In a population genetic context, subfunctionalization represents a neutral form of duplicate gene evolution, as each copy accumulates mutations that may be reciprocally deleterious without interrupting the total function of their ancestral state. Based on this argument, it has thus been proposed that in organisms with small effective population sizes, subfunctionalization may be more relevant than neofunctionalization, given that it would be easier to lose an existing function than to gain a new one[141]. Further, the duplication, degeneration and complementation model (DDC), a type of subfunctionalization that may occur either at the gene expression or protein function levels, posits that duplicate genes will neutrally accumulate deleterious mutations on their *cis*-regulatory regions, and once sufficient mutation accumulation has significantly impaired the original expression pattern, each copy will tend to retain only a fraction of the ancestral phenotype and complement each other to cover the full spectrum of their ancestral expression pattern[140]. Recently, two additional models have been put forward that combine aspects of sub- and neofunctionalization, the "Innovation, Amplification, Divergence" (IAD) model[142,143] and the "Escape from Adaptive Conflict" model[144–146], the central idea of which is that secondary functions of an ancestral gene can get co-opted to a primary role in one of the gene duplicates[134]. In other words, the “new” function in the duplicated gene does not arise de novo but is already present in a seminal form in the ancestral gene, but the primary and secondary functions in the ancestral gene may be subject to pleiotropic constraints, precluding optimization or elaboration of both functions simultaneously. Gene duplication offers the opportunity to escape from such adaptive conflicts, allowing natural selection to optimize the primary and secondary function independently in different copies.

Alternative hypotheses exist to explain the preferential retention of duplicate gene pairs, such as the functional buffering model[147], which suggests that genes involved in essential cellular processes tend to be retained in order to ensure the maintenance of core cellular functions. In fact, several lines of evidence seem to support this hypothesis. Particularly interesting is the idea that molecular networks tend to be insensitive to single gene deletions. For instance, it has been shown that fewer than 20% of yeast genes are essential, and deletion of genes very often has little or no phenotypic effect, at least under rich media conditions[148]. A similar trend has been reported for plants and *C. elegans*[149]. Genetic robustness against null mutations has been commonly associated with the presence of backup copies (*i.e.* closely related paralogs) with functionally redundant roles[125,150]; although it has been suggested that completely redundant duplicates are most likely evolutionarily unstable, because either one of the copies can be deleted without phenotypic consequences, thus becoming invisible to selection[151–153]. In addition to the buffering capacity as a plausible mechanism to explain the reason why functionally identical duplicates may be retained over time, increases in gene dosage following duplication events may be selectively advantageous under certain conditions. Seoighe and Wolfe[112] noticed that highly expressed genes, such as ribosomal genes, were retained preferentially in duplicate after the WGD in yeasts. More recently, Conant and Wolfe[154] hypothesized that retention of specific glycolytic genes after the WGD in yeasts has caused an increased glycolytic flux that gave post-WGD yeast species a growth advantage by increasing their glucose fermentation speed. Although many of these models provide biologically plausible frameworks to explore the evolutionary fate of gene duplicates, they fail to give an account of this process under the constraints imposed by the network context in which genes operate. Thus, network-level features, such as dosage balance constraints (see below) may be a more determining factor of the feasibility of evolutionary trajectories followed by gene duplicates.

2.5.2 Gene dosage balance constraints

Based on comparative genomics data, the gene dosage balance hypothesis (GDBH) provides a set of principles to explain the dosage dependent functioning of molecular interacting systems[155–157]. From the perspective of GRNs, the GDBH posits that abnormal expression phenotypes would result from dosage balance alterations that impact on the DNA occupancy profiles of transcriptional regulators at the promoter region of target genes[158,159]. Similarly, changes in the stoichiometric relationships among the components of macromolecular complexes may induce drastic reductions of the assembled complex, thus producing unassembled intermediates and free subunits[158], which may have detrimental effects[155,160,161]. In this sense, the GDBH predicts that components with greater protein connectivity would tend to have increased chances of producing unassembled intermediates when over-expressed[158], an idea that is consistent with the finding that dosage-sensitivity is influenced by the size (*i.e.* number of interactors) of a molecular complex[161]. Moreover, it should be noted that many transcriptional regulators, such as the helix-loop-helix TFs, operate in multi-subunit complexes. Thus, changes in the stoichiometry of individual TFs are expected to impact on the activity of the complex as a whole, usually leading to altered expression patterns in a battery of target genes[162]. Similarly, in hierarchies of transcriptional regulators dosage effects tend to be pervasive. For instance, in complex regulatory cascades that affect the expression of many downstream target genes changes in the dosage of any one regulator on top of the hierarchy would be expected to propagate across the entire system, ultimately modulating the expression of the downstream genes. This situation is commonly observed in GRNs acting during early developmental phases of the fruit fly *Drosophila*, where several dosage dependent regulators have been found to control other transcription factors as targets, which in turn determine the amount and spectrum of the ultimate target genes expressed at any one time and place[163,164].

Changes in gene dosage are pervasive. For example, DNA replication during the cell cycle[165] tends to double gene dosage on a genomic scale, which can result in gene promoters displaying increased transcriptional activity during the G2 phase of the cell cycle as compared to G1[166]. Similarly, organisms such as yeast that switch between haploid and diploid life forms[167] must cope with the global increase in gene dosage. Global noise in gene expression[168,169] may also lead to significant variations in the concentration of molecular species. Moreover, such changes can have significant effects on the cellular phenotypes[170]. For example, in multicellular organisms, widespread dosage changes can be fatal[171]. Several mechanisms have been put forward to explain how dosage compensation could be achieved during both the formation of macromolecular complexes and transcriptional regulation[158], which is required to ensure proper cellular functions and homeostasis. For instance, it has been hypothesized that dosage compensation can be achieved at the transcriptional level via inverse dosage effects on a target gene being regulated by the balanced activity of an activator and a repressor encoded by a gene linked to the target gene, a mechanism that has been termed local dosage compensation[158]. Similarly, the loss or gain of genes encoding subunits of a complex may be compensated by either the inverse change in gene expression from the alternate copies of the gene or equivalent changes in gene expression from all of the other genes within the complex to maintain proper balance[158]. Therefore, if one partner in a complex is over-expressed, then the overproduction of its partner(s) is needed to maintain proper stoichiometric balance, which may be accompanied by changes in mRNA degradation rates[158].

Given that changes in gene dosage can actively modulate the dynamical behavior of GRNs, increasing attention has been paid to dissecting network-level mechanisms responsible for dosage compensation[171]. In a recent study, it has been demonstrated that the galactose signaling pathway (GAL pathway) in *Saccharomyces cerevisiae* is dosage compensated at the network level, in the sense that the activity of the inducible network showed no significant change when the dosage of the entire system was halved in diploid cells[172]. By mathematically and computationally analyzing 2-component networks, the authors were able to demonstrate that such compensation effect could arise solely as a topological feature of the network as long as the following criteria were met: 1) the two components had to have different regulatory signs; 2) they had to interact with a 1:1 stoichiometry; and 3) the effects of one of the two components had to be indirect and exerted its effects on transcription through action on the other component[172]. More recently, building upon the study by Acar *et al.*[172] on network dosage compensation, it has been shown computationally that necessary conditions are required in N-component networks to achieve dosage compensation, such as the existence of a 2-component subnetwork with an activator and an inhibitor[173].

The aforementioned mechanisms of dosage-compensation may have evolved to aid in both equalizing gene expression and alleviating the toxicity caused by free unbound components in macromolecular complexes. These observations thus underscore the pressing need of maintaining dosage balance as a requirement for proper network activity, and hence normal cellular functions. Several lines of evidence support this idea. For instance, Papp *et al.*[161] showed that an imbalance in the concentration of the components of protein complexes in yeast generally leads to lower fitness, while Yang *et al.*[174] suggested that in humans, dosage sensitivity increases and subunit duplicability decreases with an increasing number of subunits in a complex. Moreover, in yeast, subunits of heterogeneous protein complexes are significantly less duplicable than homocomplex subunits, consistent with the dosage balance hypothesis[175]. It follows then that the patterns of duplicate gene retention of certain gene classes observed across several species would be severely constrained by the mode of gene duplication[110,137] (*e.g.* while network dosage remains balanced following a WGD event, smaller-scale duplications result in imbalanced stoichiometric relationship among network components), and thus by dosage balance effects. In line with this idea, Li *et al.* [176] observed that in protein-protein interaction networks gene duplicability was negatively correlated with protein connectivity, indicating that gene retention after smaller scale duplications would preferentially occur to poorly connected genes, while genes retained in duplicate post-WGD tend to be allocated in more connected parts of the network. Interestingly, these observations are also consistent with the finding that genes duplicated through smaller scale duplications represent different gene classes than those retained from WGDs[110,137]. Importantly, such a reciprocal pattern in duplicate retention (*i.e.* significant over-retention following WGDs and under-retention following smaller-scale duplications) for certain functional gene categories is one of the predictions made by the GDBH[177]. Therefore, compared to the neofunctionalization and subfunctionalization models, it has been argued that the GDBH better explains the gene content data from several sequenced genomes across eukaryotic lineages that have undergone both small and large-scale duplications[178,179], and that neofunctionalization and subfunctionalization may have occurred, instead, once genes have been retained in duplicate as a result of dosage balance constraints[137,178,179].

2.6 Shaping the evolutionary potential of GRNs through *cis/trans* regulatory changes and gen(om)e duplications

On the basis of the different lines of evidence discussed above, it is evident that the combined force of *cis/trans* regulatory changes and gen(om)e duplications have played a decisive role in shaping the evolutionary potential of GRNs. Through the gradual accumulation of changes in *cis/trans* regulatory elements in combination with gen(om)e duplications a vast space of possibilities is provided for evolution to tinker with the structure of GRNs. WGD events, in particular, are thought to create “regulatory spandrels” as a result of dosage balance constraints that prevent the rapid loss of regulatory genes after WGD[123,137,177,180]. However, when the constraints imposed by dosage balance effects relax over time, due to network dosage compensation mechanisms (*e.g.* buffering, feedback and feedforward mechanisms[173,181]) or to changes in selective pressures, GRNs would be free to undergo sub- and/or neofunctionalizing changes in their structural connectivity (rewiring events), either neutrally or adaptively, via sequence turnover mainly at *cis*-regulatory regions (promoters), and less frequently at *trans*-acting elements (DNA binding domains). Ultimately, this would endow GRNs with an enormous evolutionary potential that may manifest itself under the appropriate conditions, which may explain why the evolution of GRNs have frequently been associated to the origin of morphological novelties, the evolvability and adaptability of organisms, and the evolution of biological complexity in general[117]. For instance, the implications of regulatory evolution upon gene duplication are of great interest in evolutionary studies focused on the origin of morphological novelties in plants[182,183]. In particular, in the MADS box gene family, a major determinant of floral organ identity, extensive and complex patterns of gene duplication have been documented. What we still don't know is how these events have impacted on the structure and function of the GRNs that determine the diversity of floral morphology observed across the angiosperms. Supported on a great deal of experimental evidence, Rosin and Kramer discussed plausible scenarios where regulatory changes, duplication and co-option occurring in the floral homeotic ABC gene system could have given rise to novel expression domains, and thus novel floral morphologies[183]. Similarly, Geuten *et.al.* provided an overview of the different ways in which evolvability and robustness could be favored in the floral homeotic B-class GRN[184]. The emerging consensus in these studies is that both regulatory subfunctionalization and neofunctionalization, achieved through *cis/trans* regulatory changes, may play a major role in the evolvability and robustness of morphological traits, which underscores the importance of concerted changes in the structure of GRNs in evolution. However, as pointed out by Geuten *et.al.* [184], fresh insights into GRN evolution may only be accessible through quantitative modeling approaches, which have indeed gained increasing attention over the last decade as tools to investigate the evolutionary origin of emergent system properties. In the next section a brief overview is provided on current modeling approaches to study, in particular, the evolution of GRNs.

2.7 An *in silico* approach to GPM problems

Perhaps the most appropriate way to study a complex GPM problem (*e.g* the association between the genetic blueprint–genotype– of a regulatory network and its phenotypic manifestation at the gene expression level) is by constructing a mathematical model able to capture essential features of the underlying molecular network (*e.g.* their regulatory wiring or network topology), and then interrogate the model systematically via computer simulations with the aim of identifying statistical regularities in the system's behavior under a wide range of perturbations. This *in silico* approach to GPM problems has received increasing attention since Kauffman developed the NK model to describe interacting gene nets with the aim of exploring the global dynamics of cellular differentiation pathways[6]. Despite of being an overly simplified representation of a real GRN, the Kauffman model has proven instrumental in the understanding of generic properties of complex interacting systems[6]. Technically speaking, Kauffman's Boolean GRN model was intended to give an account of the genome of an organism composed by a set of N genes each being represented as a binary variable describing two accessible gene-expression states: expressed (1) or not expressed (0). Since the expression of a gene is controlled by the expression of some other genes, Kauffman assumed the genome to be a directed network in which a link from a given gene X to another gene Y means that X controls the expression of Y. Given the complexity of real GRNs, Kauffman made three simplifying assumptions: i) every gene is controlled by (is connected to) exactly K other genes; ii) the K genes to which every gene is connected are chosen randomly with uniform probability from the entire system; iii) each gene is expressed with probability p and is not expressed with probability $1 - p$, depending upon the configurations of its K controlling genes. The computational cost required to simulate this type of GRN models is minimal, which has enabled the statistical analysis of ensembles of random networks with different structural features aimed at revealing universal properties of GRNs. For instance, one of the major insights derived from the work by Kauffman was that the randomly generated GRN models tend to exhibit "ordered behavior", as opposed to chaotic behavior, given some constraints in the amount of connectivity in the GRNs, a phenomenon that Kauffman termed "order for free" [6]. Surprisingly, relying on his analyses of random GRNs, Kauffman was able to predict the number of cell types in a species, given the number of genes that the species possessed[6].

Presently, a large family of network models exist that offer different levels of resolution on the molecular mechanisms by which GRNs operate. In this section I will provide a brief overview of the different GRN modeling approaches used to investigate GPM problems across different disciplines.

2.7.1 Network modeling approaches: features and scope

Network modeling approaches lie at the heart of many application domains of systems biology[11,13]. One case in point is the nascent field of evolutionary systems biology, whose major aim is to reveal the evolutionary origin of general emergent system properties, such as robustness, evolvability, modularity and phenotypic plasticity[185]. One of the pioneers in evolutionary systems biology studies is Andreas Wagner, who a few decades ago introduced an influential network model[186,187] that has served as a workhorse to investigate a great variety of evolutionary issues. In essence, the Wagner's network model provides an abstract representation of real GRNs where the genotype is conceived as an interconnectivity matrix W (see Figure 3.2), whose elements w_{ij} denote the regulatory interactions among genes, such as the effect on gene i of the product of gene j, involving activation ($w_{ij} > 0$) or repression ($w_{ij} < 0$). Under this connectivity matrix approach, GRN dynamics are simulated using the following set of difference equations:

$$S_i(t+\tau) = \sigma \left[\sum_{j=1}^{N} w_{ij} S_j(t) \right] = \sigma[h_i(t)]$$

With σ representing a sign function $\sigma(x) = 1$ if $x > 0$, $\sigma(x) = -1$ if $x < 0$, and $\sigma(x) = 0$ if $x = 0$. $h_i(t)$ represents the sum of all regulatory effects of all the genes on gene i. In this way, any expression state $S_i(t)$, at a given time point, maps to the set $\{-1, 1\}$. Since its introduction, the model has been slightly modified over the years. For instance, the first modification intended to incorporate more biological realism at the gene regulation level was proposed by Siegal and Bergman [188] who replaced σ by a sigmoidal function:

$$f(x) = \frac{2}{1+e^{-(a*x)}} - 1$$

With the function $\sigma(x)$ implemented in the original Wagner's model being a special case of $f(x)$ when $a \to \infty$. Later on, Masel[189] implemented $\sigma(x)$ to give either 1 (if $x \geq 0$) or 0 (if $x < 0$), by which any expression state $S_i(t)$, at a given time point, maps to the set $\{0, 1\}$.

The original model and its extended versions provide a very abstract representation of the structure and dynamics of GRNs, mainly because these models fail to account for concentration dependent regulatory effects on gene expression dynamics, which is a critical aspect of most real GRNs, specially those involved in developmental processes that are responsive to morphogen gradients[190,191]. This is but one of the many bottlenecks of this type of network models that will be discussed in chapter 3. Despite the level of abstraction and the limitations in providing an adequate representation of the structure of GRNs, these models have been used to address a great variety of evolutionary questions (for an extended discussion see [191]). For instance, the initial implementation of the model by Wagner was used to address the impact of gene duplications on the phenotypic stability of GRNs [186], as well as to investigate the evolutionary origin of robustness to perturbations [187]. Later on, using the modification mentioned above, Siegal and Bergman examined the effect of selection for an expression phenotype on the robustness to perturbations [188] and the potential of GRNs to buffer genotypic variation under normal conditions and to reveal it phenotypically under particular conditions [192]. Similarly, Azevedo and colleagues [193] used the Siegal and Bergman implementation to investigate the interplay between robustness, sexual reproduction and epistasis, whereas MacCarthy and Bergman used the a similar model to study conditions under which asexual reproduction could be favored, results that were later complemented by Lohaus *et. al.* [194] while addressing the long term competition between sexually and asexually reproducing individuals. Using this model, Martin and Wagner [195] focused on the effects of recombination on robustness, population variability and offspring viability. Draghi and Wagner [196] assessed the impact of sexual and asexual reproduction on the evolvability of GRNs, that is, adaptation to a new optimum phenotype. Ciliberti *et al.* used the model to study the relationship between innovation and robustness based on the structure of a metagraph of GRNs [197], while Masel used the model to study the evolutionary origin of genetic assimilation [189].

Building upon the influential connectivity matrix modeling approach described above, promising advances have been made towards increasingly biologically realistic GRN models. In particular, the use of network modeling approaches that account for concentration dependent regulatory effects on gene expression dynamics (*e.g.* based on ordinary differential equations–ODEs) have proven instrumental in shedding light on the evolution of developmental mechanisms and emergent system properties. For instance, in a series of studies, Salazar-Ciudad *et al.* developed a generic network modeling framework, building upon the connectionist model used to simulate the *Drosophila* GAP GRN [198–200], to perform evolutionary explorations of the space of regulatory networks with developmental pattern formation capabilities [201,202]. Interestingly, the modeling approach proved successful at revealing general design principles of developmental regulatory networks, as well as the complex genotype-phenotype relationships they mediate over the course of evolution [201,202]. Similar approaches have been used to shed light on the possible evolutionary paths that have led to the different segmentation modes observed in metazoans [203,204]. A common feature among these studies is that the basis for performing evolutionary explorations is the continuous parameter space defined by the network models. Recently, ten Tusscher and Hogeweg have developed a series of ODE-based GRN models that are built upon artificial genomes [205,206]. In this way, ten Tusscher and Hogeweg were able to simulate the evolution of GRNs using different types of mutation operators, such as duplications and deletions, to address several questions regarding, for instance, the impact of sexual reproduction on phenotypic diversity, as well as the robustness, modularity and evolvability of developmental pattern forming networks [205,206]. Despite providing a first glimpse into possible evolutionary properties of GRNs, these sequence-based models still fail at adequately accounting for the inherent dosage sensitive nature of transcriptional regulatory systems [156,207], mainly because essential mechanistic molecular details (*e.g.* competitive DNA binding between non-divergent paralogous TFs) are simply overlooked.

ODE-based network modeling approaches have also become instrumental in other research areas where a solid understanding of the inner workings of GPM problems is essential. For instance, in quantitative genetics research, these network models have been used to investigate how allelic variation in GRNs is transformed into patterns of phenotypic variation at the level of gene expression outputs [208–213].

A key assumption in systems inspired quantitative genetics approaches is the existence of an unequivocal relationship between allelic variation, which is discrete by definition, and changes in the kinetic parameters of a dynamical system model, which vary on a continuum[209–211,213]. Moreover, in this research field an ODEs-based network modeling framework is not only an efficient tool to recreate *in silico* large ensembles of GPMs, but it also provides a solid framework to interpret a great variety of biological data[212,213]. Furthermore, ODE-based network modeling approaches have also become the primary tool to guide the design of synthetic regulatory circuits capable of accomplishing predetermined regulatory tasks, by which cells are able to process information in complex ways[214–216]. More specifically, researchers in this field face the challenge of how to assemble from a set of genetically-encoded molecular components minimal regulatory schemes that can respond in an appropriate manner to chemical/physical stimuli (*e.g.* hormones or light), which can be seen as an extended version of a GPM problem due to the pressing need of considering the cellular environment in which these circuits are embedded. Using an ODE-based network model, the exploration of the design space of minimal regulatory schemes capable of accomplishing a given task can be efficiently performed by simulating large ensembles of structurally different circuits[214–216].

A notable study that highlights the importance of a systems biology-inspired network modeling approach to gain mechanistic insight on the design principles of complex biological systems focuses on a comparative analysis of what is perhaps the best understood developmental patterning system, the gap GRN, between two distantly related dipteran species, the moth midge *Clogmia albipunctata* and the fuit fly *Drosophila melanogaster*[217]. Inspired on the classical ODE-based connectionist model referred above[198–200], this novel study uses gene circuit models fitted to quantitative spatio-temporal gene expression data for four key gap genes (hunchback, Kruppel, giant, and knirps) in *Clogmia*, to compare the computationally inferred regulatory wirings with the reverse-engineered *Drosophila* GAP GRN. Interestingly, this study reveals that contrary to the single regulatory wiring found in the fruit fly, the *Clogmia* GAP developmental system seems to be operative under four distinct wiring configurations, which share some common features with the fruit fly circuit. This study demonstrates that a the network modeling approach can make testable predictions on core regulatory principles underlying complex developmental processes in two distantly related organisms. Another impressive example of the predictive power of a network modeling approach is the work by Chau *et.al.* who successfully engineered simple molecular circuits that reliably execute spatial self-organized programs, such as the asymmetric distribution of key molecules within the cell (polarization), using a coarse-grained computational model to explore a wide range of regulatory schemes [218]. Although not exactly being a GRN case study, this work demonstrates that the inner workings of complex regulatory circuits can be adequately captured in mathematical models that account for some degree of biological realism. More along this line is the GRN-centered study by Cotterell and Sharpe who built upon the previously mentioned connectionist modeling approach[198–200] to perform an exhaustive enumeration of the distinct regulatory topologies that can be generated from only three TF-encoding genes in order to map out an atlas of morphogen interpretation mechanisms[190]. Although the authors didn't prove experimentally their numerical predictions, the insights gained from this study may provide a valuable guideline for future synthetic biology studies focused on the design of complex developmental GRNs.

Despite being a mere abstraction, a mathematical model able to capture essential components/features of a real world system can provide some clues as to cause-effects relationships when the problem at hand is too intricate that can not be understood by intuitive reasoning alone. In the context of biological problems, abstract mathematical representations (*e.g.* network models) of either generic or relatively well experimentally characterized systems become indispensable tools to narrow down the number of plausible hypotheses that could explain, for instance, the evolutionary origin and/or mechanistic underpinnings of emergent system properties. Obviously, any mathematical model is always open at the top toward further generalization, and at the bottom toward further elaboration and enrichment[219]. A fundamental limitation comes from the fact that any mathematical model leaves out of its scope a vast universe of unmodeled realities, which can introduce unavoidable artifacts into any theoretical description of the real world system, and thus uncontrollable distortions into the set of hypotheses generated that aim to explain the system's behavior under particular conditions. Arguably, the most critical aspect in any attempt to model a given biological system is the lack of sufficient quantitative information on the system's parameterization, which may be prohibitively time- and labor-consuming, if feasible at all. Hence, in the vast majority of models such a parameterization may be simply unavailable. In addition, there is no guarantee that in vitro measurements of, for instance kinetic rates or thermodynamic constants, can be truly representative of the in vivo system's behavior[219]. Nevertheless, with the increasing availability of information on the wiring (structure) and phenotypic behaviors (function) of biological systems under a great variety of conditions, the time is ripe for the development of mechanistically detailed and biologically realistic GPM models to gain a systems-level understanding of complex biological phenomena.

In the next chapter, we will discuss several shortcomings that are present in most currently available GPM modeling approaches, which must be solved in order to adequately study the short and long-term impact of WGD and SGD events on the dynamic behavior of GRNs. In particular, we advocate the development of mechanistic GPM modeling approaches that explicitly take into account the genetic encoding of GRNs in order to shed light on their evolutionary properties. Arguably, this is a critical step toward the development of predictive frameworks on the evolution of emergent system properties, as well as for the rational design of synthetic circuits and for studying the etiology of complex diseases such as cancer.

“(...) I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics, for men thus endowed seem to have an extra sense”

Charles Darwin

3

Modeling the evolution of biological systems from a systems perspective

Abstract

Systems biology-inspired genotype–phenotype mapping models are increasingly being used to study the evolutionary properties of molecular biological systems, and in particular to study general emergent properties of evolving systems such as modularity, robustness and evolvability. However, the level of abstraction at which many of these models operate might not be sufficient to capture all relevant intricacies of biological evolution in sufficient detail. We argue that in particular gene and genome duplications, both evolutionary mechanisms of potentially major importance for the evolution of molecular systems and of special relevance to plant evolution, are not adequately accounted for in most GPM modeling frameworks, and that more fine-grained mechanistic models may significantly advance our understanding of how gen(om)e duplication impacts molecular system evolution.

Author contribution

Gutíerrez J. and Maere S.

Published in Trends in Plant Science, 2014, 19(5):292-303.

3.1 Introduction

Living organisms are extremely complex systems of interacting molecular components. A detailed insight into how molecular biological systems are structured and how they function can have far-reaching consequences in many life science application domains, from improving crop yields and optimizing industrial microbiological processes to fighting disease. The aim of molecular biology has always been to unravel the workings of molecular systems in living organisms. Although there have been countless attempts since the 1950s to characterize the function of individual genes, scaling such efforts from individual genes to integrated molecular systems was hampered by the lack of adequate system-scale data generation capabilities. However, in the past 15 years, technological developments such as DNA sequencing platforms, functional genomics techniques and bioinformatics data analysis methods have brought molecular biology into the era of systems biology.

Although systems biology studies have generated valuable data about the wiring of biological systems, attaining mechanistic insight into why these wirings are observed and how they function has been more challenging. Inspired by a rich tradition of technological systems design, many systems biologists tend to have an anthropocentric engineering perspective on molecular systems. A lot of studies have focused on analyzing the topological properties of biological networks: for example, in terms of motif content[67] and global topology[12]. Although such work has generated interesting parallels between technological and biological systems (but see e.g.[220] for critical remarks), and in some instances has led to mechanistic insight into the biological rationale for certain design features of biological systems (see e.g.[70]), in many instances such mechanistic explanations are not obvious. This is at least partly because biological systems are the result of billions of years of evolution rather than products of engineering. Consequently, a molecular biological system may exhibit unexpected design features that can only be understood properly by taking into account the system's evolutionary history. Indeed, paraphrasing a famous quote from T.G. Dobzhansky[221], one could argue that nothing in systems biology makes sense except in the light of evolution.

Vice versa, the evolution of genes is best approached in the context of the systems in which they function, since genetic components wired together in a molecular system do not evolve independently. With the exception of experimental evolution studies (recently reviewed in[222,223]) and particular branches of population genetics such as quantitative genetics[224], the field of molecular evolution (in the narrow sense, i.e. research concerned with the inference of evolutionary properties from sequence information) has until recently largely focused on the evolution of individual genes and gene families, or genome architectural properties[129]. Analogous to (and fueled by) the transition to increasingly mechanistic systems approaches in molecular biology, there is a growing interest in attaining a more mechanistic perspective on the evolution of biological systems.

Evolutionary systems biologists study the evolution of biological systems by integrating aspects of two of the most quantitative and mathematically formalized areas of research in modern-day biology: molecular evolution and computational systems biology[225]. The combination of the two could significantly advance our understanding of how biological systems work and evolve. Giving an all-encompassing definition of what is and isn't evolutionary systems biology is impossible, just as it has proven to be for at least one of its parent fields, systems biology, but one of the distinguishing features of evolutionary systems biology in our minds is that it aims to study the evolution of biological systems from a more mechanistic perspective than has been the case thus far in comparative genomics, quantitative genetics and theoretical population genetics studies (e.g.[226,227]). Although some of the first ideas along these lines have crystallized several decades ago[7,213,228,229], evolutionary systems biology can still be considered a field of research in its infancy.

Here, we discuss the use of systems biology-inspired mathematical models to study the evolution of molecular systems, and in particular to study the emergence of properties such as modularity, robustness and evolvability in gene regulatory networks. Although such modeling studies have forwarded interesting hypotheses on the evolutionary origin of emerging properties, they may not capture all relevant aspects of regulatory network evolution. For instance, gene regulatory network evolution is heavily impacted by gene and genome duplications, especially in plants. We argue that many of the existing models do not adequately capture the evolutionary effects of gene duplication and divergence, and we advocate the development of more fine-grained sequence-based models to address this issue. In the process, we discuss how the use of appropriately mechanistic evolutionary models might help address some of the major unsolved questions in plant evolution.

3.2 Genotype-phenotype maps

Evolution can be viewed as the process by which biological systems navigate genotype space in search of an optimal adaptive peak on the fitness landscape (see Glossary), a metaphor first introduced in 1932[230]. In order to study in mechanistic detail how evolving systems negotiate the fitness landscape, one needs to bridge the gap between the genotype, phenotype and ultimately fitness of a system[213] (see Figure 3.1). To this end, several mathematical models have been developed to describe the effect of genetic mutations on the phenotype of a system, thereby defining a genotype-phenotype map (GPM). The fitness of a particular genotype in a particular environment is then expressed as a function of the system's phenotype. Here, we only discuss GPM models for gene regulatory networks (GRNs), which essentially derive the temporal gene or protein expression pattern (phenotype) of a system from specific regulatory rules (network wiring and parameters) encoded in the genotype of the system.

A range of different models have been developed to represent the GPM of gene regulatory networks at different levels of abstraction, framed in distinct mathematical formalisms with particular simplifying assumptions, and aimed at unraveling the inner workings and principles of GRN evolution at different levels of granularity[191]. As with all modeling approaches, the adequacy of a particular GPM modeling formalism crucially depends on the question to be addressed. In the most abstracted (and most widely used) models, genotypes are essentially weight matrices describing the effect of certain gene products on the time-dependent or steady-state expression of other genes (see Figure 2). These models either describe the response of a target gene to input signals in Boolean terms (i.e. genes can only be on or off) or with near-Boolean sigmoidal response curves[186,188,192,193,205,233–237]. Conceptually similar models include models describing the gene regulatory network as a logical circuit[238,239] or a neural network[240]. At a higher level of mechanistic accuracy, a number of models use continuous differential equations to describe the time evolution of molecular species in the system[190,203,204,206,241–247]. While some of these models treat mRNA and protein species separately[241,242] or include time delays to account for intermediary steps in expressing proteins from mRNAs[203], most model only proteins. The most mechanistically detailed GPM models used to date for studying GRN evolution use a statistical thermodynamics approach to describe

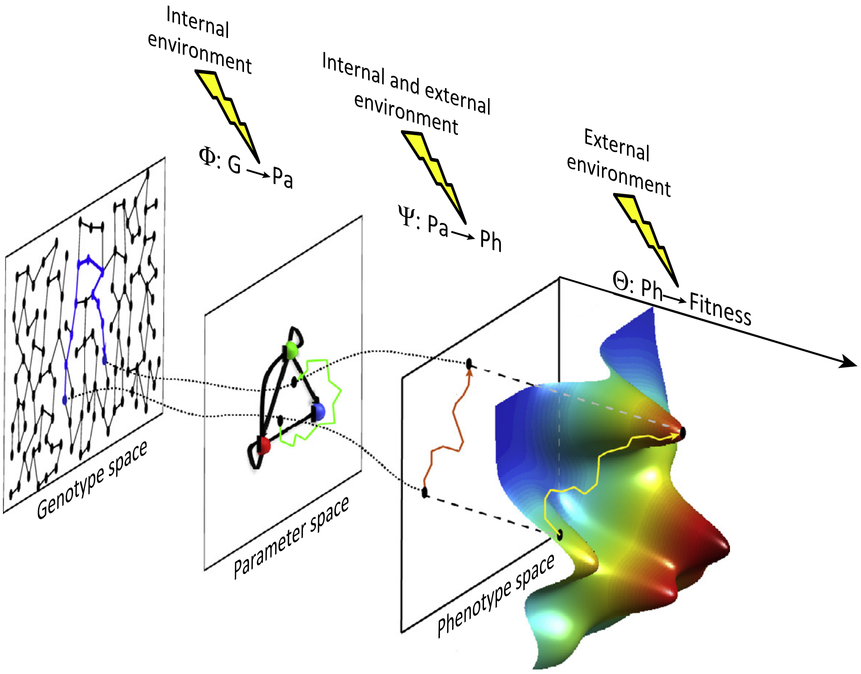


Figure 3.1: Predicting the fitness of a biological system from its genotype involves several intermediate mapping steps. *First, genotypes (G) are mapped to parameter space (Pa), describing the wiring and parameterization of the system. Pa is then mapped to phenotype space (Ph), in the case of gene regulatory network systems describing the dynamical expression pattern of the system. Finally, the fitness of the system is calculated from its phenotypic characteristics. All these mappings are dependent on internal and/or external environmental parameters, such as temperature, nutrients, and selection pressures, and, thus, may lead to different genotype-phenotype maps and fitness landscapes across different environmental conditions*[231,232]*. All spaces are represented as being 2D in this figure but are more highly dimensional in reality. Genotype space is discrete, with each dimension representing a nucleotide in the DNA sequence of a system. Pa is mostly continuous, with each dimension being a parameter in the mechanistic system model. Ph may include multiple dimensions related to the impact of the system on multiple phenotypic variables that are important for fitness.*

the binding of transcription factors to target promoter regions[248–252] (see Figure 3.2). The average promoter occupancies of various transcription factors are then used to calculate gene expression levels, either directly through sigmoidal response curves[251,252] or indirectly through differential equations, using transcriptional rates calculated from promoter occupancy profiles[248,250]. Statistical thermodynamics models can be considered analytical approximations of fully stochastic models[250,253,254].

Intriguingly, many of these mathematical GPM models were developed in the context of system design and optimization rather than system evolution[190,241–243,245]. Similarly, many of the *bona fide* systems evolution studies focus on modeling the evolution of GRN network designs rather than the evolution of the underlying genotypes, in part because the GPM models used were originally inspired on engineering methods for modeling regulatory circuits (reviewed in[255]). The use of engineering-inspired modeling frameworks has particular consequences for the adequacy of such GPM models for studying biological evolutionary processes. In particular, with few exceptions[249,251,252], the models described above do not use a biologically realistic genotype (in the present context defined in its most basic form as a string of discrete characters) as the starting point for modeling. Instead, most models study the evolutionary properties of biological systems in parameter space, i.e. they explore the effects of changes in the wiring of the system and parameterization on its dynamics, using the wiring parameters as the 'genotype'. In this respect, one might say that most models describe "parameter-phenotype maps" or PPMs rather than GPMs. One of the pitfalls of using PPMs to study the evolution of biological systems is that PPMs ignore the fact that genotype space, as opposed to parameter space, is discrete, and that there may be a complex mapping of discrete mutations in the genotype to changes in the parameters of the system. Only certain regions in parameter space may be accessible through sequence mutations, and sequence mutations may exhibit constraints, such as epistatic effects (see Figure 3.3) and trade-offs, that are not captured in parameter space. Furthermore, PPM models, by virtue of employing mutational operators on parameter values instead of sequences, generally do not adequately capture neutral mutational processes and the associated allelic variation and genetic drift, nor recombination effects on parameter values. The aforementioned factors may cause the evolutionary dynamics of PPM models to differ from those of real molecular systems, e.g. with respect to the origin of emergent properties in molecular networks, as discussed in the next section.

3.3 The use of GPM models to study the emergent properties of evolving biological systems

So far, most studies using GPM models in an evolutionary context have focused on studying emergent properties of evolved molecular networks, such as complexity[247], network motifs[238,258], modularity[235,238,240,259], robustness[193,233,252] and evolvability[226,231,260,261]. Evidence for the modular organization of gene regulatory networks and for their robustness against genetic and environmental perturbations is overwhelming[262]. However, how these emergent properties evolved and to what extent they are a general property of evolved molecular systems remains heavily debated. Although GPM-based evolutionary simulation studies have generated a wide variety of enticing hypotheses on how emergent properties might become established through evolution, some of these hypotheses are at odds with each other and with expectations from theoretical population genetics.

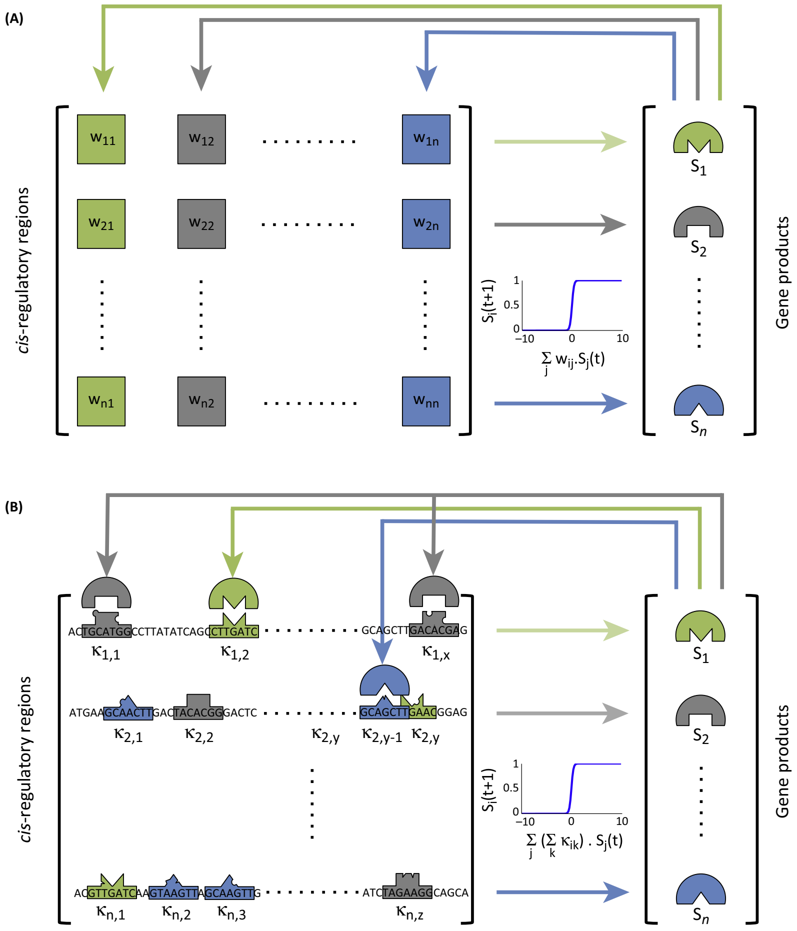


Figure 3.2: Coarse-versus fine-grained genotype–phenotype mapping models. *(A) Depiction of a weight-matrix based formalism to model the genotypephenotype map (GPM) of a gene regulatory network, inspired by*[188]*. The weight matrix W contains the parameter values describing the effect of a particular transcription factor (second index of W matrix) on a particular target gene (first index), and the phenotype vector S contains the gene expression levels at a given time step, which in the depicted model are calculated using a sigmoidal transfer function at each time step. (B) Example diagram of a GPM model based on statistical thermodynamics and incorporating an explicit sequence representation, inspired on*[252]*.* κ_{ij} *represents the binding affinity of the* $j-th$ *binding site in the promoter of gene* i *(note that different binding sites for the same transcription factor, represented by slightly modified pictograms, have different* κ *values depending on their distance from the consensus binding sequence). The regulatory input to gene* i *is calculated by summing up the regulatory effects of each binding site, taking into account the expression level of the binding transcription factor, and as in (A), the resulting input is translated to gene expression levels in the next time step using a sigmoidal transfer function (the index* k *in the sum on the* x*-axis of the sigmoidal transfer function only runs over the sites bound by transcription factor* j*). For both the (A) and (B) modeling types, the regulatory input represented on the left- hand side of the transfer functions is in some models translated to transcriptional rates rather than directly to gene expression values. These rates are then used in a differential equation formalism to calculate dynamic expression profiles. This approach is more biologically accurate and less discrete, but comes at the expense of increased computational complexity.*

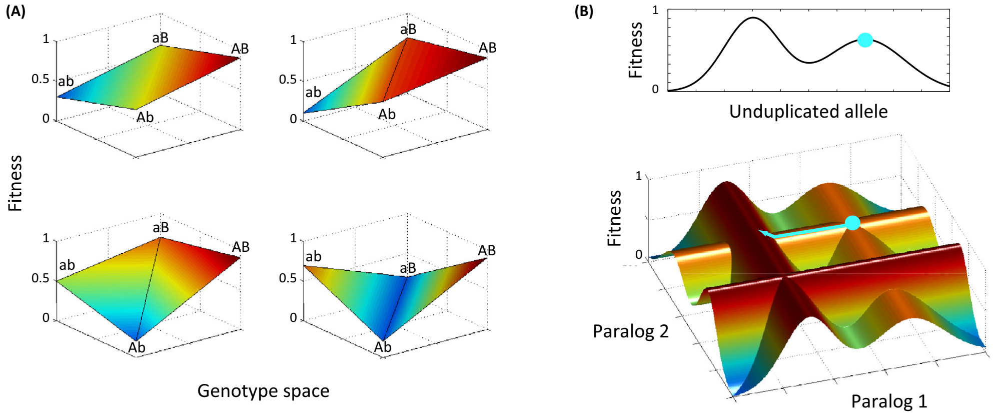


Figure 3.3: The impact of epistatic effects on fitness landscapes, and gene duplication as an extradimensional bypass mechanism. *(A) Plots representing possible epistatic effects among mutations at two loci a and b (inspired by[256]). The term epistasis refers to the fact that the fitness effect of a mutation depends on the genetic context in which it occurs. Shown are depictions of, in clockwise order starting from the top left, no epistasis, magnitude epistasis, sign epistasis, and reciprocal sign epistasis. In the case of no epistasis, the fitness effects of the $a \rightarrow A$ and $b \rightarrow B$ mutations are stable: that is, independent of the mutational trajectory followed from ab to AB, indicating that the genotype at b does not influence the fitness effect of the $a \rightarrow A$ mutation and vice versa. In the case of magnitude epistasis, the fitness effect of the $a \rightarrow A$ ($b \rightarrow B$) mutation differs in magnitude between both b (a) backgrounds but not in sign. Shown is a case of negative epistasis, where the first mutation on both trajectories has the biggest fitness effect, and the second mutation has a comparatively weaker fitness effect. Sign epistasis occurs when the fitness effect of at least one mutation changes from positive to negative between mutational trajectories. Reciprocal sign epistasis is a special case of sign epistasis in which both mutations exhibit sign-epistatic effects, creating a fitness valley between ab and AB. (B) An idealized single-gene example of an extradimensional bridge formed through gene duplication[257]. The original gene (upper panel) might reside on a suboptimal fitness peak (cyan-filled circle) from which it is unable to escape. The lower panel depicts the effects of gene duplication on the fitness landscape in case both duplicates are redundant and non-interacting (i.e., the fitness of the duplicated pair is the maximum of the fitness of both duplicates). One of the paralogs may in this case diverge (initially neutrally) while the other paralog buffers its function, causing the higher fitness peak to become accessible. However, in the primitive abstraction depicted here, only one divergence dimension is accessible per gene, restricting the possible paths of the system on the fitness landscape. In reality, neutral divergence scenarios such as the one depicted here most often lead to nonfunctionalization and loss of one of the duplicates.*

Modularity for instance has been hypothesized to emerge from modularly varying selection pressures on a system in time or in space[238,259,263]. A related hypothesis is that modularity evolves as a byproduct of selection for increased evolvability[231,260], while others have suggested it could be a byproduct of selection for increased robustness[264]. Some studies have hypothesized that in multicellular organisms such as plants, modularity emerges as a by-product of selection for tissue-specific specialization of expression patterns[235], which could, for instance, be accomplished through duplication and (adaptive) divergence of regulatory components[265,266]. Yet another alternative hypothesis states that modularity might emerge from selection pressure to reduce network connectivity costs[240]. The aforementioned studies all investigate possible adaptive scenarios leading to the evolution of modularity. However, it has been argued that modularity might more easily evolve through neutral mechanisms such as gene duplication, followed by neutral subfunctionalization and genetic drift of duplicate copies[267], although adaptive selection pressures might serve to augment the modular outcome of these neutral processes[266]. Neutral scenarios for the emergence of modular networks have thus far not been investigated using systems biology-inspired GPM models.

Another emergent property that has been intensively studied using systems biology-style GPM models is the genetic robustness of a system, defined as its degree of phenotypic invariance to genetic perturbations, i.e. mutation or recombination. GPM-based simulations have suggested that genetic robustness might be a direct effect of stabilizing selection[187,252], or an indirect effect of selection for developmental stability[188], in particular in sexually reproducing organisms[193]. Genetic robustness is generally assumed to be a distributed property of genetic networks, i.e. a property that cannot be pinpointed on particular genes, but rather emerges from the way systems are wired. However, an alternative mechanism through which molecular systems could acquire mutational robustness is through genetic redundancy caused by gene duplication[268]. Although purely redundant (i.e. selectively neutral) duplicates are generally evolutionarily unstable[104,137], redundant duplicates that are stabilized through (either neutral or adaptive) divergence may still buffer the function of their paralogous counterparts to some extent, and thereby contribute to genetic robustness. Robustness in this respect may to some extent be considered an inherent consequence of duplicate evolution rather than an evolutionarily selected feature. In contrast to mutational robustness, the evolution of environmental robustness in GRNs, i.e. robustness to macro-environmental changes or stochastic changes in the external or cellular micro-environment, has to our knowledge not been investigated thus far with mechanistic GPM models. Likely, one of the reasons is that (micro-)environmental perturbations are difficult to represent on the level of abstraction at which most present-day models operate, with the possible exception of stochastic models. However, especially for sessile organisms such as plants, environmental robustness is expected to be a crucial fitness-determining factor and therefore a directly selectable trait. Since environmental robustness may engender genetic robustness, direct selection for environmental robustness may be an alternative mechanism explaining the origin of genetic robustness[4,269]. This however remains to be investigated in mechanistic detail.

Evolvability, loosely defined as the ability of a lineage to generate novel phenotypes through mutation that could potentially be useful for adaptation to other environments than the current one, i.e. the capacity of a population to produce new selectable variation[267], is arguably the emergent property of molecular systems that is hardest to fathom. Most authors agree that the hypothesis that evolvability could be directly selected for because it provides future adaptability benefits is teleological in nature, and therefore not valid[267,270]. A more sensible hypothesis states that higher evolvability might be promoted by group selection in variable environments. Some GPM-based simulation studies have forwarded evidence for such a scenario[196,261,271], although it has been argued that the conditions for this kind of group selection to outweigh the selection pressures operating on the level of individuals are rarely fulfilled[267,270]. A third class of hypotheses consider evolvability to be a byproduct of other selective forces. Evolvability could for instance be a byproduct of modularity tenTusscher:2011gh, which, as outlined above, is itself hypothesized to be the result of either direct or indirect selection pressures (intriguingly, in a circular way including evolvability-related group selection[231,260]), or of neutral evolutionary processes, e.g. related to gene duplication and divergence[267,270]. Since evolvability is often perceived as being opposed to robustness[226,272], a number of GPM-based simulation studies have tried to reconcile the emergence of both in molecular networks, sometimes with contrasting results. Some studies found that evolvability could emerge as a by-product of indirect selection for genetic robustness in multicellular developmental networks[206], or more generally in complex GRNs[233]. Genotype network theory, in which genotypes are depicted as residing on a neutral network, i.e. a set of mutationally connected genotypes with the same phenotype, makes similar predictions. Since more genetically robust genotypes have larger neutral networks, they are also expected to have more off-network neighbor genotypes that may facilitate evolvability[272–274]. The results of other studies[275,276] moderate these claims, arguing that not all networks can simultaneously exhibit robustness and evolvability, but only those that operate close to the so-called critical regime, i.e. networks at the transition between ordered and chaotic dynamics. At the other end of the spectrum, mutational robustness was found to negatively correlate with evolvability in sexual[277] and asexual[196] populations evolving in environments that fluctuate on an evolutionary timescale, and assortative networks molded through frequent gene duplication were found to exhibit higher robustness but lower evolvability[278]. Although the latter scenario might be very relevant for plants in view of the frequent occurrence of gene and genome duplications in plant evolution, the evolvability of plant networks is in contrast thought to be partly engendered by gen(om)e duplications[279] (see also below).

As evidenced above, GPM model simulations have led to a multitude of alternative and sometimes competing adaptive hypotheses to explain the origin of emergent properties in molecular networks. Further research is needed to reconcile these disparate views and to gain more insight into the true evolutionary origins of the aforementioned emergent properties, in particular since some GPM modeling results contradict expectations from population genetics theory[267]. In this respect, potential neutral explanations for emergent properties have thus far largely been overlooked in systems biology-inspired GPM modeling studies. Although systems biology-inspired GPM models have been instrumental in investigating the evolution of molecular systems, and will continue to be so in the future, a crucial question is whether the level of abstraction at which most of the current GPM models operate is suited to study the emergent properties of molecular biological systems adequately, i.e. whether they reflect the essential features of molecular fitness landscapes and truthfully account for the evolutionary mechanisms and population genetic factors that impact the evolution of biological systems. In addition to the PPM versus GPM arguments presented above, we argue that this is not the case for at least one important factor impacting the evolution of molecular systems, namely gene duplication. First, we outline the importance of adequately modeling gene duplication and divergence for studying the evolution of molecular systems, and the particular importance of modeling the effects of gen(om)e duplications to understand plant evolution.

3.4 The impact of gene duplications on evolution

Gene duplication has since long been recognized as an important factor in the evolution of biological systems[280], and gene duplications have at some point been linked to all of the emergent system properties discussed in the previous section[265–267]. In addition, gene duplication has been forwarded as a means to change the properties of the fitness landscape itself. Fitness landscapes are to a large extent shaped by epistatic effects between alleles (see Figure 3.3)[256,281,282]. The evolutionary navigability of the fitness landscape is severely dependent on the nature of these effects and their prevalence across the fitness landscape. Reciprocal sign epistatic effects for instance may lead to rugged fitness landscapes featuring fitness valleys[283], and it is well-known that evolutionary trajectories have difficulties crossing such valleys (*i.e.*, there may exist a more optimal system configuration nearby in sequence space, but getting there involves several mutations that are deleterious on their own, even though they are beneficial when combined)[256,283] (see Figure 3.3). If the fitness valley is not too wide and can be bridged by a few mutations, deleterious mutations may act as stepping stones for adaptive evolution[284], but population genetics theory suggests that such a direct crossing mechanism is only efficient under particular conditions, either involving genetic drift in small populations or high double mutant frequencies in large populations, provided that the mutant loci are closely linked[285]. These conditions are rarely fulfilled for sexually reproducing metazoans with average effective population sizes, such as most higher plants. Crossing fitness valleys through point mutation trajectories is therefore often not feasible, and special evolutionary mechanisms need to be invoked to create passable ridges on the fitness landscape. In particular, the navigability of sequence space may be facilitated by introducing extra genotypic dimensions, enabling the circumvention of reciprocal sign-epistatic effects by rerouting the evolutionary trajectory, *i.e.* by making compensatory changes at other loci first, a mechanism that has been termed an 'extradimensional bypass'[286–288]. Increased navigability has been proposed to be a general feature of high-dimensional fitness landscapes[289,290], but it has in particular been linked to gene duplication, since a duplicated gene copy may in some cases buffer deleterious mutations in the other copy[257,287,288] (see Figure 3.3).

Duplications have also been studied in the context of alleviating another important factor constraining the evolution of multifunctional molecular systems, namely evolutionary trade-offs between multiple functions of a molecular system, which may either refer to clearly identifiable subfunctions of a system or to the functioning of the system in different environments[291,292]. Evolutionary trade-offs have been studied intensively for single proteins, and in particular for multifunctional enzymes. Simultaneous optimization of two protein subfunctions is often constrained by adaptive conflicts between the subfunctions, and one way to escape such adaptive conflicts is to duplicate the protein and optimize the different subfunctions in different paralogs[145,293–296]. A gene duplication strategy to circumvent adaptive trade-offs could also be perceived as an extradimensional bypass mechanism.

3.5 The importance of gene and genome duplications in the evolution of plant systems

An adequate inclusion of gene duplication mechanisms in systems evolution models is particularly important for studying plant evolution. Large paralogous gene families abound in higher plants, and *e.g.* developmental processes in plants have been heavily impacted by gene duplication. For instance, the

Aux/IAA and ARF gene families, major players in auxin signaling responses, contain 29 and 23 members in Arabidopsis, respectively[297–299]. Subtle differences in the ARF family DNA-binding domains[297] and differences in their status as a repressor or an activator of transcription[300], together with a complex pattern of overlap and diversification in their potential interactions with various Aux/IAA family members[298] lead to a bewildering variety of possible combinatorial control functions governing auxin responses, which have thus far been elucidated only partially[301]. A similar picture arises for the cyclins and CDKs regulating cell cycle processes in plants, the duplication and functional divergence of which caused plants to have a considerably more complex cell cycle regulation structure than yeasts or animals[302–304]. Gene duplication has also played a major role in the evolution and diversification of angiosperm flowers. Various subclasses of the diverse Type II (MIKC-type) MADS-box transcription factor family are essential for floral organ specification[305]. Although most of these subclasses likely originated before the invention of the flower in ancestral seed plants[306], and although the basic ABC(E) model of floral whorl specification appears largely conserved across angiosperms[307], duplication and divergence of members of the different MIKC-type gene subclasses have played major roles in establishing the variety of elaborate flower morphologies observed in present-day angiosperms[184,305,308–315]. Besides their role in flower development, several MIKC-type subfamilies are also involved in fruit, embryo, pollen and (lateral) root development, and many MIKC-type MADS-box genes functioning in flower development have been coopted to or from functions in other developmental processes[305].

Although small-scale duplication processes likely played a crucial role in establishing the diversity of developmental gene families in plants[316], several studies suggest that the expansion of many regulatory gene families in angiosperms, *e.g.* Aux-IAAs and various MADS-box subfamilies, may to a large extent be due to genome duplications[299,308,309,315,317]. Whole-genome duplication (WGD) is particularly common in higher plants. All extant seed plants are believed to have a polyploid ancestry[318], and in many lineages there is evidence of additional rounds of genome duplication, with some angiosperm genomes carrying remnants of up to six WGDs[319,320]. It has been hypothesized that the preferential expansion of regulatory gene families through genome duplication is caused by dosage balance effects, *i.e.* quantitative effects on the expression of target genes as a consequence of disturbing the stoichiometric balance between regulators and targets[159,321]. Since WGDs, as opposed to small-scale duplications, conserve the relative dosage of *e.g.* transcription factors and targets, and since loss of duplicated regulators after WGD would disrupt this balance, dosage balance effects are thought to promote the retention of regulatory duplicates specifically after WGD[110,137,177]. Although the same argument could be made for transcription factor targets, the loss of duplicated targets on average arguably leads to less pleiotropic effects than loss of duplicated regulators. Thus, it has been argued that WGDs leave behind a regulatory spandrel in diploidized polyploids, a reservoir of evolutionary potential that may manifest itself if the conditions are right (*e.g.* related to niche availablility)[117,123]. In this respect, ancient WGDs have repeatedly been linked to the invention of evolutionary novelties, such as flowers in angiosperms[117,134,322,323]. However, rather than facilitating de novo innovation, the power of genome duplications may lie more in their ability to elaborate on primitive versions of innovative features and fully exploit their potential[117,123], for example by lifting pleiotropic constraints on multifunctional genes and facilitating their co-option for specialized purposes[134]. To what extent genome duplications facilitate the evolvability of plant systems remains an unanswered question, and evidence to support or disprove a causal link between WGDs and evolvability is scarce and circumstantial. Another debated feature of WGDs is their potential to generate immediate or short-term adaptive benefits. Such benefits have been hypothesized based on the prevalence of present-day polyploids in stressful habitats, and on the observation that many of the more recent paleopolyploidizations in angiosperms did not occur randomly in time, but that they appear to be clustered around the K-Pg mass extinction event, 66 Mya[118,324]. The latter suggests that compared to their diploid progenitors, polyploid plants might have been better able to avoid extinction and adjust to the changed environment. Although various adaptive hypotheses have been forwarded to explain the increased occurrence of paleopolyploids around the K-Pg boundary, most notably related to heterosis effects in (allo)polyploids Vanneste:2014,Chen:2013bj, and although a recent study on the salt stress tolerance of various Arabidopsis accessions and cytotypes showed that even autopolyploids may exhibit pre-adaptation to environmental stress factors[325], neutral scenarios, such as increased unreduced gamete production in stressful environments, may also explain the increased occurrence of WGDs around the K-Pg extinction[324].

Without question, gene and genome duplications have had a considerable impact on the structure and complexity of plant developmental systems. More generally, evolution after gen(om)e duplication in regulatory systems is a complex story of dosage effects, neutral and adaptive divergence, partial redundancy and co-option of duplicates to other functions, of which we've only written the introduction at this point. A crucial question is to what extent duplication of regulatory genes, either through small-scale duplication or WGD, may contribute to true evolutionary innovations or to elaborations of existing systems. In this respect, mechanistic GPM-based simulation of WGD effects on the evolvability of molecular systems would prove very helpful. Although evolution after genome duplication has already been studied to some extent for metabolic networks[154,326], mechanistic GPM-based models have thus far not been used to study the evolution of gene regulatory networks after genome duplication. GPM-based simulations might also shed more light on more immediate effects of allo- and autopolyploidization on the phenotype of molecular systems in plants, such as heterosis and dosage balance effects.

3.6 Modeling duplications in GPM models

Despite the potentially major importance of gen(om)e duplication processes in systems evolution, ranging from their potential roles in establishing emergent properties such as robustness, modularity and evolvability to their impact on the topology of fitness landscapes and their hypothesized roles in adaptability, evolutionary innovation and elaboration of existing molecular systems, which are particularly heavily debated in the plant evolution field, the majority of the GPM-based GRN evolutionary simulations performed thus far have not included gene duplication as an evolutionary mechanism. More importantly, as outlined below, most of the studies that do consider gene duplication exhibit important flaws in the modeling of duplicate divergence, which could in some cases considerably impact the evolutionary simulations performed and the conclusions drawn.

A substantial number of neutral and adaptive mechanisms have been forwarded in the past decades to explain the divergence of duplicate genes, mostly from a theoretical viewpoint[137,280,327]. However, in a gene regulatory network context, the duplication and divergence of transcriptional regulators is subject to particular constraints that have been understudied so far on the mechanistic level. One of these constraints is the aforementioned dosage balance effect[159,321], which is thought to severely impact the chance of duplicate fixation in regulatory networks subject to stabilizing selection. On the other hand, dosage balance effects may prove beneficial under directional selection. In either case, insufficiently mechanistic models of dosage balance effects may lead to unrealistic gene duplication dynamics in GPM-based simulations. In GPM models based on a weight matrix formalism with Boolean or sigmoidal transfer functions, dosage balance effects are very hard to capture given the quasi-discrete gene expression dynamics used. In this respect, differential equation-based formalisms are better suited to study dosage balance effects.

Another characteristic of duplicated transcriptional regulators often overlooked in systems biology-inspired GPM models is the fact that they do not diverge independently. In the absence of allelic differences, duplicated transcription factors initially bind to the same target sites in the genome. Divergence of their target repertoire is expected to be a gradual process involving changes in both the promoter regions of the target genes (*cis* changes) and the DNA binding domains of the duplicates (*trans* changes). Any *trans*-regulatory change in the DNA binding domain of one of the transcription factors is expected to generate effects on the expression of many target genes simultaneously. On the other hand, a *cis*-change in a target gene is expected to impact the effect of both duplicated transcription factors simultaneously. In contrast, many studies assume that regulatory linkages evolve independently across target genes and across duplicated transcription factors[204,236,254,268,271,328], which is not a realistic assumption for either *cis*- or *trans*-evolution of duplicated regulatory links. A couple of studies do allow *trans*-mutations to affect multiple targets simultaneously, but in a very discrete way, in the sense that a diverging transcription factor immediately looses all regulatory links to previous targets and gains a new set of targets[244,329,330]. Some studies assume a similarly discrete functional divergence of duplicates immediately upon duplication[275,276,278], which rather mirrors the addition to the GRN of a novel transcription factor unrelated to the existing ones, as in[247]. In other models, duplicated transcription factors are exclusively allowed to diverge on the *cis*-regulatory level, *i.e.* their target preference remains identical over the course of evolution, but they might become differentially regulated in time or in space[205,206,261]. Although this strategy avoids the pitfall of independent divergence of duplicates and their target lists, and can be considered reasonable under the assumption that GRN evolution is mainly cis-regulatory in nature[75,331] (but see[81]), many intriguing aspects of evolution after gene duplication cannot be studied within this framework.

As a result of the fact that the relevant constraints are not captured accurately in many GPM modeling frameworks, few studies have thus far adequately addressed GRN evolution through gene duplication. A notable exception is a study on a duplicated two-component system based on experimentally derived fitness landscapes[256,332]. However, since obtaining comprehensive experimental fitness landscapes for larger systems is an extremely challenging task, fully unraveling the effects of gene duplication in complex GRNs will largely depend on the development of adequate GPM models. In this respect, one recent study used a statistical thermodynamics modeling framework to simulate the evolution of a duplicated autoregulatory activator[250]. Although the modeling framework in this study did not yet incorporate an explicit genotype representation, we believe that models at this level of mechanistic granularity are the way forward for studying the evolutionary effects of gene duplication, and more generally for studying the evolution of molecular systems and emergent properties.

3.7 The way forward

As outlined in the previous sections, the failure to take into account the explicit genotypic structure of biological systems, together with the fact that transcriptional regulatory interactions, gene duplication and duplicate divergence are often not modeled realistically, limits the use of many of the present-day GPM models for studying the evolutionary properties of gene regulatory networks. From that perspective, the time is ripe to use more mechanistic models to simulate systems evolution, such as the statistical thermodynamics models with explicit sequence representations that have been developed to study transcriptional regulation processes[58,333–335]. Although these models are simplified in their own right (e.g. with respect to modeling the effect of transcription factor binding on RNA polymerase recruitment and expression), they may reflect the true nature of the genotype–phenotype map of gene regulatory networks to a sufficient extent to significantly advance our knowledge on GRN evolution. A number of recent studies have already used statistical thermodynamics models based on explicit genome representations in an evolutionary context, e.g. to study the evolution of a single enhancer[251], or in an engineering context, e.g. for the custom design of enhancers[335], or to investigate the relation between network motifs and function[258]. To our knowledge, only one study thus far[252] has performed evolutionary simulations using a statistical thermodynamics modeling framework on larger systems (10 genes) in a population context. Indeed, population-based evolutionary simulations at this level of granularity are extremely challenging in view of the massive amount of computing involved. However, given the recent advances in high-performance computing, such simulations are expected to become increasingly feasible in the near future. Even these more mechanistically accurate analytical representations might not be sufficient for some systems. One study found that stochastic simulation of the phenotypic effects of gene duplication in a simple toggle switch GRN produced qualitatively different results from analytical approximations[253], suggesting that the use of even more fine-grained evolutionary GPM models may be required for some systems.

In addition to being crucial for characterizing the fundamental evolutionary properties of molecular biological systems in detail, such as the structure of fitness landscapes and the emergent properties of evolved systems, more realistic GPM models will also become increasingly useful for addressing evolutionary questions of practical importance, e.g. with regard to plant breeding, carcinogenesis or the development of antibiotic resistance, all of which are essentially evolutionary processes. Addressing the major unsolved questions regarding the impact of genome duplications on the adaptability and long-term evolvability of plants may also prove useful from a practical perspective. It might for instance give us more insight into the evolutionary potential of crop species, many of which are polyploids, and reveal to what extent WGD might help present-day plant species to adapt to current climate change and other human-caused environmental upheaval[118,336]. Furthermore, mechanistic GPM-based evolutionary simulations may in the future prove useful to assess the evolutionary potential of genetically engineered systems in crops or microorganisms, and the impact thereof on natural ecosystems[225]. Fine-grained GPM models are already employed in the design of synthetic systems (e.g.[335,337]), but thus far not to study their evolutionary aftermath.

"All models are false, but some models are useful"

George E. P. Box

4

A mechanistic GPM model to study the evolutionary potential of GRNs

Author contribution

All content within this chapter was written by myself and revised by professor Steven Maere. It contains the materials and methods of the work presented in chapter 5, which is a paper under preparation.

4.1 Introduction

In this chapter I describe a novel sequence-based model of GRNs where the underlying genotypic encoding can be linked with dynamic expression phenotypes in mechanistic detail using core principles of transcriptional regulation. Building upon standard thermodynamic models[58,333,338–340] and a connectionist network modeling approach[198–200], and inspired by the regulatory genome paradigm[32], this mechanistic GPM modeling framework is intended to simulate the evolution of GRNs over an explicitly defined genotype space. Most importantly, this modeling framework addresses the shortcomings discussed in chapter 3, which make most present-day GPM models inadequate for simulating the impact of gene and genome duplications on the evolution of GRNs.

4.1.1 Computing protein-DNA binding propensities via a low-level molecular mapping

In order to link GRN genotypes with dynamic expression phenotypes, a low-level molecular mapping is implemented to allow the quantitative assessment of sequence-encoded protein-DNA binding propensities (see Figure 4.2, next chapter). To achieve this, we build upon a previously introduced empirical mapping approach[341], which has proven successful in the evolutionary design of promoter regions with distinct signal integration properties[337]. Such a low-level molecular mapping assumes that whenever a TF *j* binds to a DNA motif *x*, each amino acid in the DNA binding domain interacts with exactly one base pair[341]. Accordingly, individual AA-base interaction scores are then calculated based on the log-odds between the frequency of observed AA-base contacts (in crystallographically resolved TF-DNA structures) and the expected frequency of such contacts under the assumption that there are no specific AA-base binding preferences (see table 4.1). The total TF-DNA interaction score was then calculated as the sum of scores over all amino acid-base contacts. As did in[337], we followed this rationale to model the binding free energy, $\Delta G_{j,x}$, between a given TF j with an array of n amino acids and a binding site x of $n = 10$ nucleotides long (a typical length for binding sites found in both eukaryotic and prokaryotic genomes, see [56] and references therein), both indexed by i, as follows:

$$\Delta G_{j,x} = -\sum_{i=1}^{n} U_{j_i,x_i} \tag{4.1}$$

Where U is a 20 x 4 matrix containing the binding propensities of amino acid-base contacts as inferred from crystallographically resolved protein-DNA complexes[341]. These binding propensities were found to be roughly proportional to binding free energies[341]. In our simulations, the proportionality constant was set to -1 (the exact scale of the binding free energy values is of minor importance in the present artificial system context), implying that a high score is associated with a negative $\Delta G_{j,x}$ (or strong binding). The binding free energies computed using (eq. 4.1) are then used to calculate protein-DNA binding affinities (association constants) according to the following relationship:

$$K^{Assoc}_{j,x(i;n,m)} = \exp(-\Delta G_{j,x(i;n,m)}/(R * T)) \tag{4.2}$$

With $K^{Assoc}_{j,x(i;n,m)}$ being the association constant of TF j bound to a DNA motif x with sequence coordinates (n, m) along the promoter region of gene i; $\Delta G_{j,x(i;n,m)}$ representing the estimated binding free energy, $R = 8.314 J * mol^{-1} * K^{-1}$ denoting the Gas constant, and the T the temperature in K units (300 K), being a scaling factor for the free energy of change upon binding[57,58]. To assess all potential protein-DNA binding interactions operating in the GRN system, gene promoters were scanned (using a simple sliding window of length $n = 10$ bases, as mentioned above) for TF binding sites. We assume that in order to influence transcription, TFs should bind in a specific orientation with respect to the transcription start site. Although such an orientation bias is observed for only a minority of TFs in reality[342], implementing this restriction allowed us to screen promoter sequences unidirectionally, leading to substantial computing time savings. The algorithm further assumes that a given TF is only able to recognize sequences up to three mutations away from its highest-affinity binding site (*i.e.* the consensus sequence). Allowable (near-optimal) binding sequences for a given TF are defined as sequences in the 3-mutational neighborhood of the sequence whose binding energy differs from that of the reference sequence by at most 10*$R * T$, which seems to be a biologically plausible energy threshold[57,59]. The previous biologically reasonable restrictions were mainly imposed to avoid the computational overhead caused by the combinatorial explosion of possible (mostly ineffective) binding sites.

4.1.2 Assessing the wiring of GRNs based on individual protein-DNA binding events

Upon evaluation of all active TF binding sites in the genome and the associated binding affinities using the aforementioned procedure, a thermodynamic modeling approach is used to assess the regulatory wiring of a GRN. To this end, we use a similar formalism as in previously published models[58,333,338–340] to compute the aggregated regulatory input to a given gene. First, for a given TF binding to a permissible DNA motif on the promoter region of a target gene i, we calculate its fractional occupancy as follows:

$$f_{[j,x(i;n,m)]} = \frac{K^{Assoc}_{j,x(i;n,m)} * [P_j]}{1 + K^{Assoc}_{j,x(i;n,m)} * [P_j] + CF_{x(i;n,m)}} \tag{4.3}$$

With $[P_i]$ and $K^{Assoc}_{j,x(i;n,m)}$ representing the concentration of TF j and its binding affinity for site $x_{(i;n,m)}$, respectively, and $CF_{x(i;n,m)}$ denoting a binding competition factor defined as:

$$CF_{x(i;n,m)} = \sum\nolimits_l \sum\nolimits_{\forall z \in Z_l} K^{Assoc}_{l,z} * [P_l] \tag{4.4}$$

	G	A	T	C
Gly	–3.93	–3.93	–3.93	–3.93
Ala	–3.93	–3.93	0.66	–3.72
Val	–3.93	–3.93	–0.17	–3.57
Ile	–3.93	–3.93	0.65	–3.44
Leu	–3.93	–3.93	–0.94	–3.93
Phe	–3.93	–3.93	–0.81	–0.12
Trp	–1.96	–3.93	–1.96	–3.93
Tyr	–2.87	–2.87	0.54	0.13
Met	–2.58	–0.28	0.42	–0.28
Cys	–2.23	0.07	–2.23	0.07
Thr	–3.46	–0.06	–0.06	–1.16
Ser	0.42	–0.68	–0.28	–0.68
Gln	–0.09	1.16	0.31	–3.09
Asn	0.48	1.93	0.71	0.71
Glu	–3.93	–1.24	–3.93	0.55
Asp	–3.93	–3.37	–3.93	1.01
His	1.56	0.46	0.87	–0.23
Arg	2.74	0.34	1.25	–3.93
Lys	2.16	–0.08	0.21	–3.93
Pro	–3.93	–3.93	–0.30	–3.29

Table 4.1: Binding propensity scores for amino acid–DNA base pairs. *Table of scores reproduced from*[341]. *Scores were calculated using the formula:* $ln[f_{ij}/(f_i x 0.25)]$, *where* f_{ij} *is the frequency of the pair between amino acid* (i) *and DNA base* (j). f_i *is the frequency of amino acid i in the SWISSPROT database of protein sequences and 0.25 is the equal probability assumed for each of the DNA bases.*

With l indexing competitive binding TFs, $[P_l]$ denoting the concentration of a competitive binding TF l, and Z_l representing the set of sites recognized by TF l that overlap with site $x_{(i;n,m)}$, which is bound by TF j. We assume that the set of binding sites recognized by a given TF on the promoter of a given target gene make independent contributions to the transcriptional regulation of the gene. This leads us to the formulation of the following expression to account for the aggregated regulatory input F contributed by TF j to control the expression of gene i via a set of binding sites X_i:

$$F_{[j,i]} = \sum\nolimits_{\forall x \in X_i} f_{[j,x(i;n,m)]} \tag{4.5}$$

Furthermore, it is assumed that activating and repressing regulatory signals contributed by TFs act independently to elicit quantitative effects on the transcriptional regulation of target genes. Accordingly, the aggregated regulatory input $u^i(t)$ on a target gene i at time t is modeled as follows:

$$u^i(t) = \sum\nolimits_j W_{ij} * F_{[j,i]} \tag{4.6}$$

With W_{ij} describing the nature of the regulatory influence of (activating or repressing) of TF j on a target gene i, where W_{ij} = 1 (-1) denotes a transcriptional activator (repressor). We have deliberately assumed that TFs can only impart either activating or repressing transcriptional control on target genes, but not both (*i.e.* a TF j cannot activate the expression of target gene i and inhibit that of another gene k). Therefore, TFs with variable regulatory roles were not accounted for (see discussion below).

4.1.3 Deriving dynamic expression phenotypes from the thermodynamically assessed transcriptional regulatory logic of GRNs

Lastly, the thermodynamically assessed regulatory wiring obtained through (eq. 4.1 - eq. 4.6) is plugged into a kinetic model of gene regulation to derive the dynamic expression phenotype of a GRN. To do this, we rely on an expanded version of a classic ODE-based neural-like network model that has been used to study developmental pattern formation in the Drosophila embryo[198–200]. Unlike the original formulation of the model[198–200], our model explicitly accounts for the expression dynamics of mRNA species as well as protein species (see Figure 4.1). Although this entails increased mathematical complexity and additional computational load, accounting for this allows for a more realistic description of the multiple regulatory layers underlying the dynamical behavior of GRNs, as well as more flexibility toward future model enhancements. The following expressions describe the structure of the ODE model:

$$\frac{d[m_i]}{dt}(t) = T^i_{\max} * \varphi(u^i(t)) - k^i_{m\deg} * [m_i](t) \tag{4.7}$$

$$\frac{d[P_i]}{dt}(t) = k^i_{psynt} * m_i(t) - k^i_{p\deg} * [P_i](t) \tag{4.8}$$

With $[m_i]$ and $[P_i]$ denoting the concentration levels of the mRNA and protein species encoded by gene i, respectively. The expressions $k^i_{psynt} * m_i(t)$ and $k^i_{p\deg} * [P_i](t)$ from eq. 4.8 provide the rates at which the protein encoded by gene i is being synthesized and degraded. Protein synthesis rate constants (k^i_{psynt}) were sampled from the biologically plausible range $[0.5, 20]$ (min^{-1})[343]. Plausible values for protein half-lives, from which one can derive the protein degradation rate constant ($k^i_{p\deg}$), were sampled from the range $[500, 5500]$ (min). The expressions $T^i_{\max} * \varphi(u^i(t))$ and $k^i_{m\deg} * [m_i](t)$ from eq. 4.7 represent the rates at which the mRNA encoded by gene i is being produced and degraded, respectively. The term $k^i_{m\deg}$ gives the mRNA degradation, which is derived from the mRNA half-life and sampled from the biologically plausible range $[10, 100]$ (min) (see [344,345]). The term $T^i_{\max}$ denotes the maximal achievable transcriptional rate, which has been shown to be a temperature-dependent and diffusion-limited parameter (see [25,26]). Transcriptional rates are considered in molar terms (M/min), and operational values for $T^i_{\max}$ have been sampled from the biologically plausible range $[10^{-4}, 10^{-1}]$ (see [343,346]). Further, $\varphi(u^i(t))$ represents a sigmoid-like function that determines the transcription rate of a gene as a function of the protein-DNA binding interactions realized in the promoter region:

$$\varphi(u^i(t)) = \frac{1}{1 + \exp(-\vartheta(\theta^i_o + u^i(t)))} \tag{4.9}$$

With ϑ being the slope of the sigmoidal function, and denoting a threshold for the transcriptional activity of gene i, which is a proxy for a basal transcriptional rate. In biochemical terms, ϑ can be regarded as the transcriptional responsiveness of a target gene to incoming regulatory signals. Upon initial explorations of values from the range $[5, 15]$ we observed that the spectrum of achievable expression profiles was qualitatively consistent. Therefore, unless otherwise mentioned, simulations have been performed with $\vartheta = 10$. In addition, to determine suitable values for θ^i_o it was assumed that in the absence of any regulatory signal, the basal transcriptional rate of a target gene falls within the range $[0.01 * T_{\max}, 0.1 * T_{\max}]$. Unless otherwise mentioned, simulations have been performed with

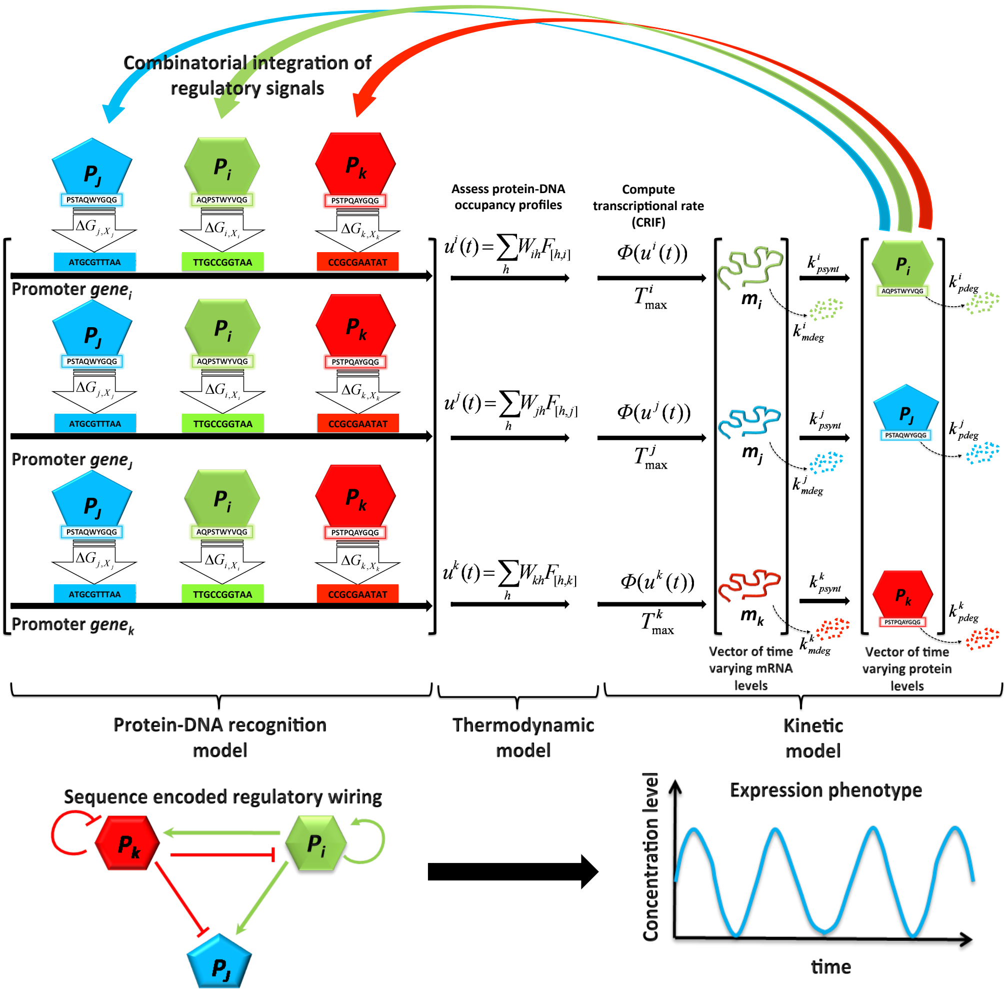


Figure 4.1: Schematic representation of the the different features incorporated in the GPM modeling framework. *The GPM modeling framework is composed of 1) a protein-DNA recognition model (low-level molecular mapping model) used to scan promoter regions in search for DNA motifs bound by TFs with variable affinities; 2) a thermodynamic model used to compute protein-DNA occupancy profiles, and correspondingly the transcriptional rate of a target gene; 3) a kinetic model used to simulate the time varying concentration levels of mRNA and protein species.*

$T_{\max} = -log(19)/10$, which gives a basal transcriptional rate of $0.05 * T_{\max}$ in the absence of regulatory signals. Operationally, $\varphi(u^i(t))$ represents a multidimensional *cis*-regulatory input function (CRIF), the evaluation of which can be thought of as giving the frequency with which the RNA polymerase bound to the core promoter enters an energetically favorable state[198,347]. Importantly, the term $T^i_{\max} * \varphi(u^i(t))$ must be scaled in order to properly account for genome dosage balance and compensation of network dynamics (see below). ODE simulations were performed with MATLAB (version R2011a), using the ode stiff solver ode15s. In all simulations, the initial mRNA and protein concentrations were set to 0, and (2000 time steps) were simulated to assess the expression phenotype of a model GRN. Computing times are quite variable, depending on the parameter settings used, but they typically range on $[10, 60]$ s.

It is worth noting that the value of particular network control parameters such as $T_{\max}$, $k_{m\,\deg}$, k_{psynt} and $k_{p\,\deg}$ can be largely determined by the intracellular milieu (*i.e.* PH, temperature, ionic strength, etc.). It should be noted, however, that such kinetic parameters may also be in part genetically determined. For instance, kinetic rate parameters can be quantitatively fine-tuned through point mutations targeting dedicated genetic elements (e.g. a protein's synthesis rate is influenced by the strength of its ribosome binding site). Genetic factors influencing kinetic rate parameters are not taken into account in the current version of the GPM model. Future model extensions will have then to incorporate extra DNA sequence templates (*i.e.* degradation sequence tags (DST), Kozak elements and degron elements) from which one can assess the effect of genetic variation on mRNA and protein degradation rates and protein synthesis rates. In the simulations presented here (see next chapter), rate constants for particular GRN genes were sampled in advance from the aforementioned biologically plausible ranges using a latin hypercube sampling scheme.

Although the GPM modeling approach described above fails to capture non-linear interactions between DNA-bound TFs and the basal transcriptional complex (*e.g.* synergistic interactions), it can nevertheless account for combinatorial protein-DNA interactions resembling the so-called billboard-like regulatory modules[36], which are believed to play a critical information processing role during the establishment of spatio-temporal expression domains in developing embryos. As a final note, one should bear in mind that our model deliberately disregards key regulatory layers, such as the chromatin structure, alternative splicing or post-translational modifications. For instance, we have assumed that every gene in a network is always found in a transcriptionally favorable state, where the promoter sequence always remains readily accessible for the transcription factors to bind their respective target sites. Likewise, neither time delays nor stochastic reaction processes were accounted for in our modeling framework. We have focused, instead, on continuous deterministic expression dynamics. This implies that the GRNs operate under a regime of sufficiently high intracellular mRNA and protein concentrations.

4.1.4 Genome dosage balance and compensation of network dynamics

Evidence suggests that to keep the concentration of transcription factors (TFs) relatively invariant upon changes in ploidy levels, cells implement a volume-mediated compensation mechanism[348,349]. In this sense, our modeling framework assumes that every time a GRN duplicates its full repertoire of genetic sequences (*i.e.* through a WGD) its immediate impact on the dynamic expression phenotype being evaluated is compensated for by a concomitant doubling in cell volume. In other words, invariant expression dynamics are effectively achieved every time a GRN undergoes a WGD as a result of all the internal regulatory fluxes being automatically rescaled. Therefore, although the duplication of a GRN system configuration results in a much more complex regulatory wiring (*i.e.* the number of regulatory interactions quadruples upon the duplication event), at the outset, no improvement in functional performance is achieved. Nevertheless, the newly acquired regulatory layer may in principle confer other advantages such as enhanced robustness to genetic and/or environmental perturbations, or it could also bring about considerable functional advantages following sequence divergence over the course of evolution. To implement ploidy-invariant GRN dynamics (*i.e.* genome dosage balance) the rate equation describing the time varying concentration levels of an mRNA species is rescaled as follows:

$$\frac{d[m_i]}{dt}(t) = \frac{T^i_{\max}}{NG} * \varphi(u^i(t)) - k^i_{m\,\deg} * [m_i](t) \tag{4.10}$$

With $T^i_{\max}$ representing the maximal transcription rate achievable from gene i, and NG denoting a cell volume-related factor that scales the rate at which the mRNA encoded by gene i is being synthesized as a function of the number of copies of a haploid GRN system configuration. In other words, NG = 1, 2, corresponds to a haploid and a diploid GRN system configuration, respectively. Note that re-scaling this rate equation results in the thermodynamic equations used for modeling fractional occupancies being rescaled automatically, as follows: suppose that for the haploid case there is one copy of the TF encoding gene j, which recognizes a binding sequence located on the promoter of target gene i, along with one copy for the TF encoding gene l, which is a competitive binding factor that recognizes the binding sequence that overlaps with the previous binding sequence. Accordingly, we get the following modified fractional occupancy term for a haploid case:

$$f_{[j,x(i;n,m)]} = \frac{K^{Assoc}_{j,x(i;n,m)} * [P_j]}{1 + K^{Assoc}_{j,x(i;n,m)} * [P_j] + K^{Assoc}_{l,z(i;r,s)} * [P_l]} \tag{4.11}$$

Then, upon duplication of the (haploid) GRN system configuration the above fractional occupancy term is reformulated in order to properly account for the presence of the resulting gene copies j^1 and j^2, and l^1 and l^1, coding for TF j and TF l, respectively. The full form of the fractional occupancy for a given gene copy j^1 is now defined as:

$$f_{[j^1,x(i^1;n,m)]} = \frac{K^{Assoc}_{j^1,x(i^1;n,m)} * [P_{j^1}]}{1 + K^{Assoc}_{j^1,x(i^1;n,m)} * [P_{j^1}] + K^{Assoc}_{j^2,x(i^2;n,m)} * [P_{j^2}] + K^{Assoc}_{l^1,z(i^1;r,s)} * [P_{l^1}] + K^{Assoc}_{l^2,z(i^2;r,s)} * [P_{l^2}]} \tag{4.12}$$

Then, right after duplication of the GRN system configuration, the fractional occupancy of both copies of a given gene j are equal, and equal the pre-duplication gene, resulting in: $[P_{j^1}] = [P_{j^1}] = [P_j]/2$, which entails that the fractional occupancies for each copy are now:

$$f_{[j^1,x(i^1;n,m)]} = f_{[j^2,x(i^1;n,m)]} = \frac{f_{[j,x(i;n,m)]}}{2} \tag{4.13}$$

This renders the $u^i(t)$ term in the aforementioned rate equation essentially the same for both copies, which together with the halving of T_{max}, leads to halving of all the mRNA and protein rate equations, thus resulting in genome dosage balance and compensation of GRN dynamics. It should be noted, however, that this only applies for cases immediately after duplication of the (haploid) GRN system configuration. As duplicated GRNs undergo quantitative diversification via sequence divergence of the gene copies over the course of evolution, the above equations get automatically reformulated accordingly. Also note that the assumption of genome dosage invariance is only biologically reasonable for strictly intracellular systems, and not *e.g.* for transmembrane components, given that cell surface area scales differently compared to volume.

4.1.5 Shortcomings and future GPM model extensions

Like any other model, the one developed in this work is characterized by a set of assumptions, which makes it an incomplete representation of the type of biological systems under study. It is important to bear in mind that any attempt to model a biological system will be far from being "exact", in the sense of being able to fully account for all possible intricacies. There will always exist features and processes that a model will fail to capture, because of either mathematical complexity, computational limits, or simply because of our lack of knowledge on how certain features of the system under study work together. Despite these inherent limitations in the modeling process, acknowledging the existence of bottlenecks is crucial towards future model enhancements. In what follows we will discuss the shortcomings of, and potential extensions to the model.

The assumption that the status of a given TF is either activator or repressor is an important shortcoming that limits the number of alternative wiring configurations that can be probed over the course of evolution, which may in turn negatively impact on the evolvability of GRNs. In actual GRNs, the regulatory status of a TF may be encoded in a dedicated domain (*i.e.* transactivation)[22,55]. However, these functional domains can exhibit a variety of amino acid sequences and structural features in different transcription factors[18]. Most critical, the status of a TF as a repressor or activator is usually highly context-dependent, largely determined by the TF's interacting partners (*i.e.* other TFs or cofactors)[22,51], the mechanism of which is not fully understood at this point. Moreover, it has been shown that the regulatory activity of key developmental TFs can be shifted depending on the signaling strength (*i.e.* concentration) of a morphogen. This is the case of the Hedgehog (Hh) gradient in *Drosophila*: when Hh signaling is low, the Hh effector Cubitus interruptus (Ci) acts as a transcriptional repressor; when Hh signaling is high, Ci acts as a transcriptional activator[350]. Due to these complex context-dependent effects, the regulatory status of TFs may be highly variable during an organism's lifespan, and it may readily evolve via sequence changes affecting, for instance, protein-protein interaction (PPI) domains[51]. From a modeling perspective, the problem then becomes: how to encode the regulatory status of TFs in small stretches of nucleotides. Perhaps the simplest way to do so is by assuming the existence of a direct relationship between a TF's regulatory status and the enrichment of certain amino acids in an effector domain-encoding sequence, an assumption that currently lacks experimental support. Another important fine-grained aspect neglected in this model regards the protein-DNA binding process itself, which can be extremely complex in real cellular systems (see[351]), involving structural features (both in the TF and the target DNA sequence) as well as environmental components (*e.g.* temperature), which can render the binding process rather stochastic. Moreover, the overall transcriptional rate of a gene may be contributed in a non-linear (synergistic) manner by a series of interacting TFs[352], which is not accounted for in the currently implemented CRIF. Additional mechanisms of gene regulation that are not being currently considered are short-range repression mechanisms[333,338] and nucleosome positioning[353], which may contribute significantly to fine tuning expression patterns across spatial and temporal scales[353].

A major bottleneck of ODE-based models of gene regulation is that they are valid only under the assumption that the intracellular environment represents a well-stirred biochemical reactor, in the sense that any molecular species produced anywhere in this reactor becomes immediately available for participating in any biochemical reaction. However, within the nuclear environment this assumption may be questionable because of slow delivery of TFs to DNA binding sites. This picture is intimately linked to the notion of molecular noise, *i.e.* the stochastic or inherently random nature of the biochemical reactions underlying, for instance, the regulation of gene expression, which seems to be widespread in most intracellular environments[169,354,355]. The fundamental limits of deterministic behavior at the molecular level suggest that biological systems have evolved to cope with and exploit stochastic behavior in gene expression[355]. Noise is thought to be dictated by fluctuations in mRNA levels, which may arise from fluctuations in promoter states or the random births and deaths of mRNAs themselves, and has also been shown to result from fluctuations in factors extrinsic to the genes themselves (including pathway specific and global factors of gene expression such as the levels of transcription factors, nucleic acid polymerases, and ribosomes)[169,354,355]. Given the presumably omnipresence of molecular noise, and their important role in the functioning of regulatory circuits, and thus in modulating genotype-phenotype mappings, future model extensions should ideally account for this important factor as well. Nevertheless, under high intracellular TF concentration levels (*i.e.* expressed at $10^4 - 3 * 10^5$ molecules per nucleus), which seems to be the rule rather than the exception in animal cells (see[60]), and perhaps in other eukaryotic systems as well, biochemical reaction events are expected to exhibit a more deterministic behavior. This, for instance, entails that under such a broad range of high concentration levels, TFs will most likely reside on the DNA most of the time[60].

Obviously, model enhancements allow for a more realistic description of the multiple regulatory layers underlying the dynamical behavior of GRNs, but it comes at the cost of a heavier computational load. In essence, increasingly more detailed and biologically realistic, in the mechanistic sense, GPM models of molecular interacting systems implies the exploration of increasingly more complex parameter spaces in order to find regions in such spaces where a desired system's behavior can be effectively reproduced. However, currently it is virtually impossible to constrain the modeling process of complex biochemical reaction networks based only on experimentally verified parameter values[356]. We can only attempt to probe computationally a model's parameter space in order to find working parameter settings (see for instance[357,358]). At this point, we must emphasize that the currently implemented GPM model does not attempt to recapitulate the dynamic behavior of a given real GRN, neither it is aimed at predicting the phenotypic effects of point mutations or any other type of genetic perturbations, such as gene duplication or deletion in a particular GRN. Rather, the present model is used as a tool to investigate generic evolutionary properties of idealized transcriptional regulatory systems. In spite of these shortcomings, the GPM model is nevertheless expected to capture the most essential features of transcriptional regulatory systems, and definitely better so than the coarse-grained models used so far.

Because of all the shortcomings discussed above, the GPM model presented in this work may be extended in several directions in future model versions. One obvious extension is the the incorporation of extra genetic elements coding for additional properties of the molecular species being simulated, by which one could extend the spectrum of mutable GRN features. One particular feature that is of great interest is the capacity of TFs to form dimers (*i.e.* homodimers or heterodimers), which is usually the form required for many TFs to bind the DNA in a cooperative manner[359] (but see[60] for arguments favoring accessibility of individual TFs to DNA over cooperative binding modes). The incorporation of this molecular feature into the GPM modeling framework thus implies dedicated genetic elements coding for protein-protein interaction domains that mediate the homodimerization and/or heterodimerization of TFs[51,359]. Extending the GPM modeling framework in this direction one could then simulate the effect of higher order synergistic effects on transcriptional regulation, and examine what type of binding modes (homo- or heterodimerization) tend to be favored under different selection pressures (*e.g.* stabilizing vs. directional selection) for particular GRN dynamics (*e.g.* bistability or oscillations) as the protein-protein interaction domains of duplicate TFs diverge. In addition, one could assess whether the acquisition of the different binding modes proceeds in a neutral or adaptive manner, as well as determine whether obligate heterodimerization readily evolves from homodimerization under different evolutionary scenarios. In this way, one could shed light on the different patterns of duplicate divergence observed across several gene families, such as the class-B floral homeotic TFs where divergence of protein-protein interaction domains has been shown to be critically involved in the diversification of the structural connectivity of the regulatory network. For instance, based on electrophoretic mobility shift assays and the yeast two-hybrid system Winter *et. al.*[360], it has been shown that obligate heterodimerization may have evolved from homodimerization in a class-B floral protein during the gymnosperm/angiosperm transition. Specifically, the GGM2-like gene products tend to form homodimers in gymnosperms, while the products of the duplicated homologs in eudicots, the DEF-like genes and the GLO-like genes, tend to form heterodimers. Interestingly, it has been found that the products of the DEF-like genes and the GLO-like genes in monocots can both homodimerize and heterodimerize, which is thought to represent the transition between the homo- and heterodimerized states[360,361]. Similarly, by extending the GPM modeling framework in this direction one could also assess the impact of divergence in protein-protein interaction domains on nonlinear degradation properties of multimeric proteins, a molecular feature that has recently been linked to the evolvability of regulatory circuits[362].

4.1.6 *In silico* evolution of GRNs

4.1.6.1 Proof of concept GRN model and artificial genome structure

For the purpose of this study, genomes are condensed to a minimal GRN form, consisting in the haploid case of a linear arrangement of an activator and a repressor encoding genes that control the expression of the downstream output gene (see figure 4.2). Each transcription factor (TF) encoding gene (see figure 4.2) possesses a promoter region spanning 200 nucleotides (a typical size for functional yeast gene promoters, see[363]) and a DNA binding domain (DBD) spanning 10 codons (a size assumed due to the 1 nc:1 aa correspondence in the low level molecular mapping used in our model, see subsection 4.1.1), corresponding to 30 nucleotides. In addition, the output gene, which lacks a coding region, is only encompassed by a promoter region of 200-nucleotides in length, which represents a mutational target in the model GRNs. In the simulations presented here, only the promoter and DBD sequences of the genes are allowed to evolve, entailing that the activating or repressing regulatory status of the TFs is kept fixed (to 1 and -1 for activator and a repressor TF, respectively) over the course of the simulated evolutionary process. Obviously, by enabling the activating/repressing regulatory status of the TFs to evolve the number of distinct regulatory schemes that can be explored over evolution will increase substantially, but at this point we rather remain conservative on this modeling aspect simply because we lack sufficient experimental support on how this protein feature is genetically encoded (see previous discussion). Therefore, for the sake of model consistency, we restricted ourselves to simulating single nucleotide substitutions at gene promoters and DBDs, while keeping fixed the regulatory status of TFs. On the other hand, diploid GRN system configurations are represented by a linear genome containing four TF encoding genes (two activators and two repressors) and two output genes. As for the haploids, only the promoter regions and the DNA-binding domains of the TFs are allowed to evolve, entailing that the duplicated output proteins are assumed to remain structurally identical.

4.1.6.2 Evolution protocol

An evolution protocol has been developed to simulate the evolutionary adaptation of GRNs toward a newly imposed phenotypic optimum. Unlike most theoretical investigations that use parameterized representations of the wiring of molecular interacting systems, or parameter-phenotype mappings (PPMs), as the basis for simulating evolution throughout a continuous parameter space (see discussion in chapter 3), our point of departure is an explicitly defined sequence space, which is inherently discrete and thus requires a different treatment. For instance, simulating evolution across sequence space involves a series of constraints that must be properly addressed if one is to reveal details underlying the navigability of the fitness landscape. Critically, allowable mutational moves in sequence space are always limited by the number of mutational states accessible from a given reference point in sequence space (*i.e.* the mutational neighborhood of a given sequence). As a consequence, the number of mutant GRN genotypes that can be sampled at a given time point along an evolutionary trajectory will always be constrained by the size of the mutational neighborhood associated to a given target sequence.

Furthermore, intrinsic biochemical constraints exist that can bias the nucleotide and codon substitution patterns in evolving sequences, which may therefore impact on the mutational moves performed over sequence space at a given time point. Addressing such constraints requires the use of formal models of DNA sequence evolution. For this purpose, we use the Kimura's two-parameter model (K80)[365] and the GY94 model[366]. We must emphasize that these models were not used in this case to deliberately impose selective pressures on the DNA level itself, but only as a means to derive instantaneous nucleotide and codon substitution probabilities conforming to a specific set of rules. More specifically, the K80 and GY94 models were used to simulate nucleotide substitutions in gene promoters and codon substitutions in the DBD encoding sequences, respectively. A key parameter in both models is the transition/transversion rate ratio, κ, which in our study has been set to 2. The GY94 model takes the ratio of non-synonymous and synonymous substitution rates, ω, as another parameter, which has been set to 1 (meaning neutral changes) in this study to avoid a bias in the codon substitution process (*i.e.* to rule out any selection pressure on the DNA level).

Based on the aforementioned mutation scheme, we simulated single-nucleotide substitution mutational pathways, using a Markov chain Monte Carlo (MCMC) algorithm, to investigate the evolutionary accessibility of newly imposed phenotypic optimum in haploid and diploid GRN system configurations. We used as starting points for the evolutionary simulations an ensemble of 50 non-correlated oscillatory GRN genotypes widely scattered over sequence space (see details on the engineering of start configura-

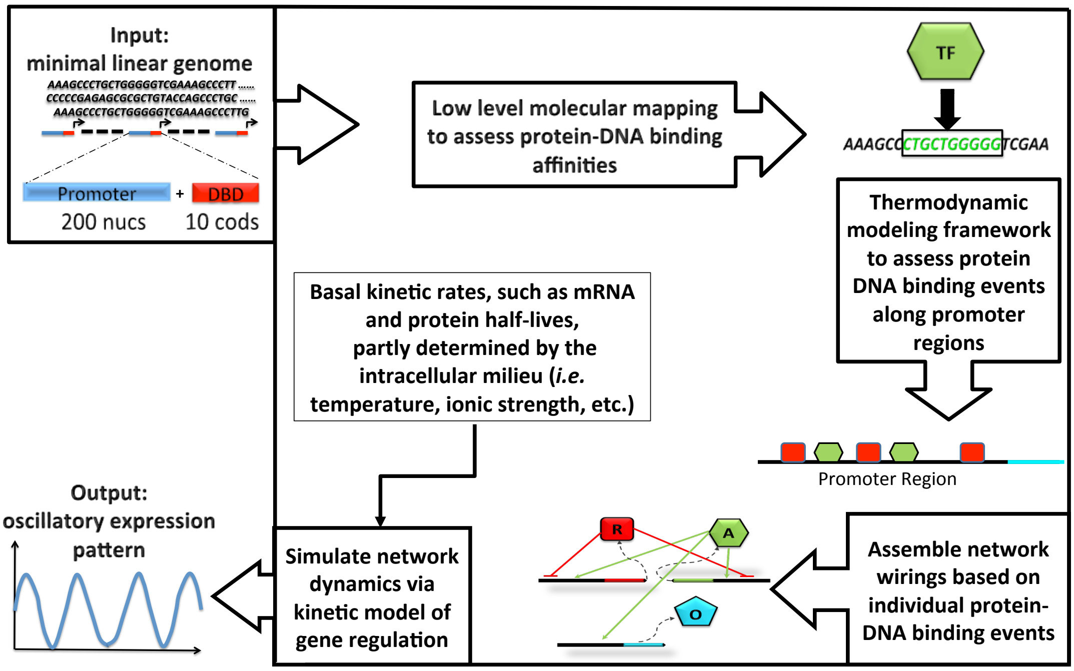

Input:
minimal linear genome
AAAGCCCTGCTGGGGGTCGAAAGCCCTT ......
CCCCCGAGAGCGCGCTGTACCAGCCCTGC ......
AAAGCCCTGCTGGGGGTCGAAAGCCCTTG
Promoter
+
DBD
200 nucs
10 cods
Low level molecular mapping to assess protein-DNA binding affinities
TF
AAAGCCCTGCTGGGGGTCGAA
Thermodynamic modeling framework to assess protein DNA binding events along promoter regions
Basal kinetic rates, such as mRNA and protein half-lives, partly determined by the intracellular milieu (i.e. temperature, ionic strength, etc.)
Promoter Region
Output:
oscillatory expression pattern
Simulate network dynamics via kinetic model of gene regulation
R
A
O
Assemble network wirings based on individual protein-DNA binding events

Figure 4.2: Proof of concept GRN model and artificial genome structure. *Workflow of the sequence-based dynamical modeling framework designed to link artificial DNA sequences with dynamic expression phenotypes. We start with minimal linear genomes containing a given number of transcription factor encoding genes (represented by blue –promoter– and red –DBD– domains, as displayed in the top-left subfigure) separated by non-mutable, non-functional stretches of DNA (30 nucleotides, black color-coded dashed lines). Every gene is assumed to encode for only transcriptionally related information, namely a promoter region (200 nucleotides) and a DNA-binding domain (DBD) (10 codons corresponding to 30 nucleotides). A low-level molecular mapping together with a thermodynamic modeling framework are used for assessing the wiring of a GRN, which is determined by microscopic molecular features such as individual protein-DNA binding events, their associated binding affinities, as well as competitive binding events among the DNA binding proteins. These microscopic features then enter into an ODE based model of transcriptional regulation to parameterize the cross-regulatory interactions among genes (e.g. a Smolen-like topology, see*[364]*), which together with network control parameters (i.e. basal kinetic rates) and initial concentrations, fully determine the expression phenotype of a GRN*

tions in subsection 4.1.6.4 below). Note that the use of several uncorrelated start GRN genotypes allows us to gather meaningful statistics on several features of the adaptive walks, thus affording an unbiased assessment of the impact of different starting conditions on the navigability of the fitness landscape. Importantly, our evolution protocol differs from traditional adaptive walk and steepest ascent algorithms in several respects. Specifically, our evolution protocol does not strictly conform to the strong selection-weak mutation (SSWM) regime, as has been traditionally used in theoretical studies of adaptation[367,368]. Under the SSWM regime, only beneficial mutations are allowed to drive evolution towards a new optimum. Thus, under this restrictive regime the role of neutral evolution is entirely neglected, which severely constrains the navigability of the fitness landscape. In our study, adaptive walks simulated under the SSWM regime (see description below) consistently lead to only small net fitness increments (*i.e.* adaptive walks are rather destined to end up entrapped prematurely on local fitness peaks), thus demonstrating the restrictive nature of this algorithm to navigate the fitness landscape (see results in chapter 5). By contrast, enabling neutral evolution significantly improves the navigability of the fitness landscape through diffusion across neutral networks of genotypes, which have been forwarded as a key organizational principle of genotype space intimately linked to the robustness and evolvability of biological systems[272,369]. Motivated by these ideas, our evolution protocol, referred to as NEA (neutral evolution allowed), has been designed to navigate the fitness landscape through the combined force of beneficial (*i.e.* fitness increasing) and neutral (*i.e.* fitness invariant) mutations, with slightly deleterious mutations being sporadically accepted (see below). Doing so enables us not only to appreciate the impact of evolution across neutral networks on the accessibility of newly imposed phenotypic optima, but also to examine the evolutionary consequences of gen(om)e duplication-mediated expansion of the neutral networks of genotypes associated to GRNs. We must emphasize that this algorithm is not intended to simulate population genetics processes. Rather, this algorithm is used as a tool to interrogate the impact that WGD/SGD have on the navigability of the fitness landscape, which is essentially achieved by assessing the difficulty of accessing high fitness scoring solutions from sub-optimal configurations via mutational pathways across sequence space.

4.1.6.3 Multi-objective fitness function

A critical aspect in any evolutionary process is the assessment of a fitness score, whose mathematical definition is problem-specific. In our case, the fitness of a GRN is a numerical value (distributed over $[0, 1]$) that provides information on how well the expression of the downstream output genes conform to a predefined dynamical pattern, being this either a low frequency or a high frequency oscillation. Assessing this type of periodic phenotypic signals requires the implementation of a multi-objective fitness function that considers several key quantitative features of the periodic signal. For the purpose of this study, we found the following fitness function to be suitable enough:

$$F(P) = OP * \alpha * \beta * \sigma \tag{4.14}$$

Where OP denotes the oscillatory potential of the time series expression output, P, with values ranging on $[0, 1]$, evaluated by the spectral analysis of the phenotypic signal. This is achieved via a conventional Fourier transform method aimed at discriminating oscillatory signals based on a statistical signature, the so-called g-statistic, g_{stat}, which is computed following the expressions below:

$$I(\omega) = \frac{1}{N} \left\| \sum_{n=1}^{N} x_n e^{-(i\omega n)} \right\|^2 ,\omega \in [0, \pi] \tag{4.15}$$

With N denoting the length of the time series expression output (1800 time steps). The periodogram ($I(\omega)$) is then evaluated at its normalized harmonic frequencies:

$$\omega_l = \frac{2\pi l}{N} ,l = 0, 1, \ldots, a;a = \frac{(N-1)}{2} \tag{4.16}$$

The spectral estimator used here for detecting dominant periodic components in the time series expression output of GRNs uses a formula introduced in[370]:

$$\tilde{S}(\omega) = \sum_{k=-L}^{L} \tilde{\rho}(k) e^{-(i\omega k)} \tag{4.17}$$

With $\tilde{S}(\omega)$ being a correlogram spectral estimator, which is equivalent to the periodogram $I(\omega)$, and $\tilde{\rho}(k)$ denoting a rank-based autocorrelation estimator (see[370] for further details). Then, using this spectral estimator we evaluate the g-statistic, g_{stat}, for each time series spectral estimate as follows:

$$g_{stat} = \frac{\max_{1 \le l \le a} \left| \tilde{S}(\omega_l) \right|}{\sum_{l=1}^{a} \left| \tilde{S}(\omega_l) \right|} \tag{4.18}$$

Put into words, g_{stat} denotes the maximum periodogram ordinate divided by the sum over all periodogram ordinates $l = 1, ..., a$, reflecting the dominance of the primary periodic component, and hence the single-frequency oscillatory nature of the signal. We restricted ourselves to a relatively small range of periodogram ordinates spanning over $[9, 20]$, with 9 denoting a low frequency signal, $[14, 15]$ denoting intermediate frequency signals (for the start GRN configurations, see subsection 4.1.6.4 below), and 20 a high frequency one. This operative range of periodogram ordinates was defined upon noting that lower or higher frequencies were rarely (less than 0.05%) found in thousands of oscillatory phenotypic signals corresponding to initially sampled GRNs (see subsection 4.1.6.4 below). As a criterion for the identification of a dominant periodic component in the time series expression output, the following function, referred to as the oscillatory potential, was found to be suitable enough:

$$OP = \frac{g_{stat}^{10}}{g_{stat}^{10} + 0.15^{10}} \tag{4.19}$$

This function was designed in a way that a GRN could be deemed oscillatory if $OP \geq 0.9$. An additional objective was evaluated in order to further discriminate oscillatory expression phenotypes according to their amplitude. To achieve this, we evaluated the maximum (X_{max}^{mut}) and minimum (X_{min}^{mut})

values for the time series expression output generated by a given mutant (X^{mut}) GRN. Then, the amplitude ($X^{mut}_{Amp} = X^{mut}_{max} - X^{mut}_{min}$) value was treated as a log-normal distributed random variable, set by the following parameters $\mu = X^{ref}_{Amp}$, and $SD = \sqrt{\log{(2)}/2}$, with $X^{ref}_{Amp} = X^{ref}_{max} - X^{ref}_{min}$ being the amplitude of the time series expression output (X^{ref}) of a reference GRN used as starting point for evolutionary simulations. All time series expression outputs were evaluated after time step $t \geq 200$. Using these parameters the following log-normal distribution was implemented to assess deviations in amplitude between a mutant and a start GRN:

$$\alpha = \frac{\mathit{normpdf}(\log{(X^{mut}_{Amp})}, \log{(X^{ref}_{Amp})}, \sqrt{\log{(2)}/2})}{\mathit{normpdf}(\log{(X^{ref}_{Amp})}, \log{(X^{ref}_{Amp})}, \sqrt{\log{(2)}/2})} \tag{4.20}$$

It should be noticed that by setting $SD = \sqrt{\log{(2)}/2}$, this objective α ensures that the fitness scores assigned to mutant phenotypic signals whose amplitudes are half and twice the amplitude of the original (unperturbed) signal, all else being equal (*e.g.* period), are effectively half the value of the original fitness score. Taking this into account is specially critical in the assessment of fitness scores of dosage imbalanced GRN system configurations (*i.e.* upon duplication and deletion of the output gene in haploid and diploid GRNs, respectively). Importantly, the fitness function implementing the above objective α as a log-normal distribution is referred to as FF1 in chapter 5 to distinguish it from another fitness function, referred to as FF2, which implements α as a Gaussian distribution for the amplitude requirements in oscillatory phenotypic signals, described as follows:

$$\alpha = \frac{\mathit{normpdf}(X^{mut}_{Amp}, X^{ref}_{Amp}, SD)}{\mathit{normpdf}(X^{ref}_{Amp}, X^{ref}_{Amp}, SD)} \tag{4.21}$$

With $SD = 0.2 * X^{ref}_{Amp}$. Further, a third objective was defined to assess the offset expression level (*i.e.* the magnitude of the off-phase within a period) of the time series expression output generated by a given mutant (X^{mut}) GRN configuration, using the following expression:

$$\beta_{t \geq 200} = \frac{X^{mut}_{max} - X^{mut}_{min}}{X^{mut}_{max}}, \lim_{X^{mut}_{min} \to 0} \beta = 1 \tag{4.22}$$

With β being evaluated after time step $t \geq 200$. Note that our search task requires β to be maximized, which occurs when $X^{mut}_{min} \to 0$. Finally, we defined the following function to drive the evolutionary optimization of GRNs across a particular frequency range in the space of oscillatory expression phenotypes:

$$\sigma = 1 - \left[\frac{|T_{PO} - C_{PO}|}{N_{POs}} \right] \tag{4.23}$$

With T_{PO} denoting the dominant periodogram component of the oscillatory phenotypic signal set as evolutionary target (phenotypic goal), C_{PO} representing the dominant periodogram component of the current phenotypic signal (X^{mut}) being evaluated, and N_{POs} denoting the number of periodogram ordinates being considered $N_{POs} = 13$. In this study, the dominant periodogram component associated to the phenotypic signal of start GRN configurations is $C_{PO} = 15$, and the dominant periodogram component (T_{PO}) for the LF-type and HF-type target phenotypic signals was set to 9 and 21, respectively.

In summary, the multi-objective fitness function $F(P)$ assesses the functional performance of GRNs distributed across sequence space. Evolution was simulated using a stringent selective criterion mimicking directional selection, generally forcing evolving GRNs to climb up in the fitness landscape. At every MCMC step along a simulated mutational pathway the fitness difference ΔF between the current genotype and the previous one is evaluated, and the following Metropolis criterion is used to decide whether the current genotype is accepted:

$$\left\{ \begin{array}{cc} \Delta F \geq 0 & accept \\ \Delta F < 0 & accept if rnd(0,1) \leq \exp(\Delta F/\kappa) \end{array} \right\} \tag{4.24}$$

With $rnd(0,1)$ being a random number uniformly drawn from $[0,1]$, and κ denoting a scaling factor used to control the magnitude of the fitness effect of a deleterious mutation for which $\Delta F < 0$. Accordingly, the lower the value for κ is the more stringent the selection criterion becomes, in the sense that even slightly deleterious mutations tend to have a substantial impact on fitness. For the purpose of this study we set $\kappa = 0.001$, entailing that the probability of accepting mutants with even very subtle negative impacts on the oscillatory expression dynamics would be, on average, rather low, thereby forcing mutational trajectories to ascend the fitness landscape. Although mutational pathways take place mainly via substitutions with neutral ($\Delta F = 0$) or beneficial effects ($\Delta F > 0$), the above criterion does not rule out the sporadic acceptance of slightly deleterious mutations, which are not allowed under the restrictive SSWM regime[367,368].

4.1.6.4 Engineering of start GRN configurations

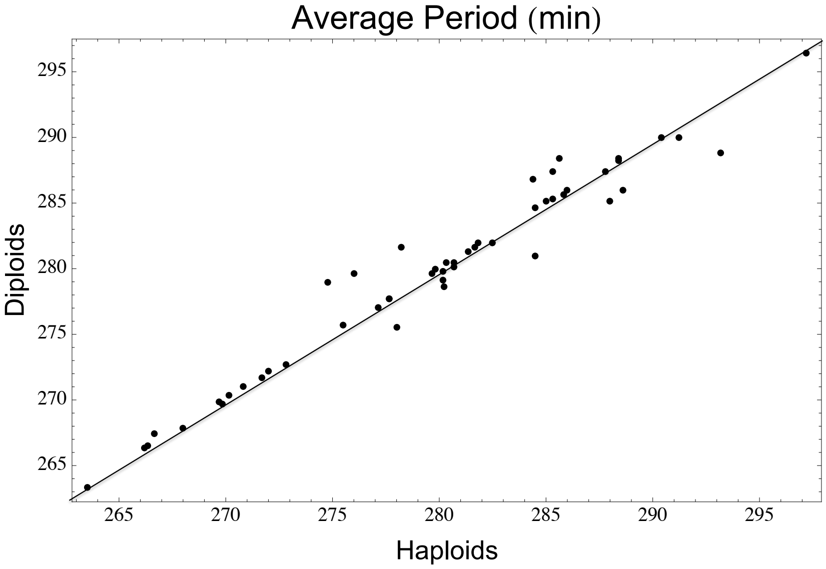
Engineering of the start (IF) GRN configurations initially required the random generation of thousands of minimal genomes (GRN genotypes), which were used to sample sequence space in search for particular configurations. Sequence space was probed using an MCMC sampling technique equipped with a fitness function that relied only on the objective OP (see eq. 4.19) previously described. Our search task involved the identification of GRN genotypes encoding for oscillatory expression phenotypes with amplitudes > 15 (a.u. of concentration), and periods falling on the rage $[250, 300]$ mins (see Figure 4.3), corresponding to signals for which OP scores ≥ 0.9 were associated to periodogram ordinates between 14 and 15 (see Figures 4.4 and 4.5). Using a 1000 previously sampled GRN configurations we conducted MCMC sampling of sequence space for 10000 steps (a single mutation per step) in total. Each sampling round was performed with a different set of network control parameters (*e.g.* mRNA and protein half-lives, see description above). Among the set of recovered solutions we noticed that approximately 90% of the oscillatory configurations coded for Smolen-like topologies (see Figure 4.2 [364]). From this set of recovered solutions we chose 50 distinct start configurations (see Figure C.1 for distributions of network control parameters) for conducting the evolutionary simulations discussed in chapters 5 and 6. In summary, the start GRN system configurations represent particular combinations of genotypes (encoding for Smolen-like topologies [364]) and network control parameters that display oscillatory expression phenotypes falling within a pre-specified range of frequencies (IF). Finally, it should be noticed that the fitness score of the start (IF) configurations under, for instance, FF1 is far from being the maximum achievable ($F = 1$) simply because the offset expression level (*i.e.* the magnitude of the off-phase within a period, which is assessed by the objective β) is relatively large (see Figure 4.6)

4.1.6.5 Implementation of the SSWM

Adaptive walks under the SSWM regime were simulated using standard procedures[367,368]. Essentially, the SSWM regime assumes the evolution of monomorphic lineages that adapt to a new environment by only fixing beneficial mutations sequentially. Then, if selection is strong relative to random genetic drift, where neutral and deleterious mutations get lost, the probability of fixation of a mutant with a coefficient of selection (s) is given by[368,371] :

$$\pi(s) \sim 1 - e^{-2s} \tag{4.25}$$

Accordingly, strongly selected mutants are relatively more likely to undergo fixation than those weakly selected, being mutants with very large selective advantages undergoing fixation with high probability. According to Gillespie the probability that a mutation with a given s sweeps through a population is proportional to $\pi(s)$[367]. Therefore, if a population is fixed at a given sequence i with fitness $f(i)$, then the probability that a mutant j with fitness $f(j) > f(i)$, substitutes i is then proportional to the fixation probability $\pi(s)$, with s defined as: $s = \frac{f(j)-f(i)}{f(i)}$



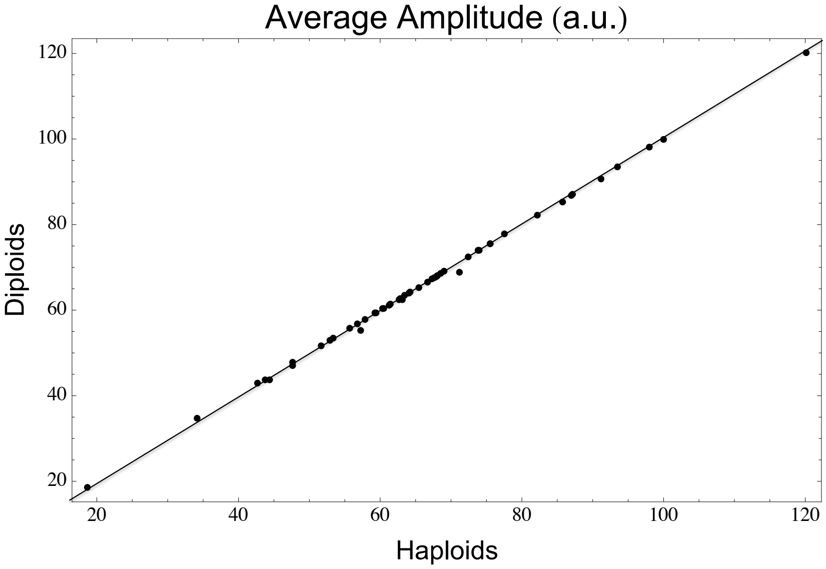


Figure 4.3: Periods and amplitudes of start GRN configurations. *Scatterplots for period and amplitude values of the oscillatory phenotypic signal in start haploid and diploid configurations. Each point in the plots describes the associated value in the pre- and post-duplication GRN system configuration.*

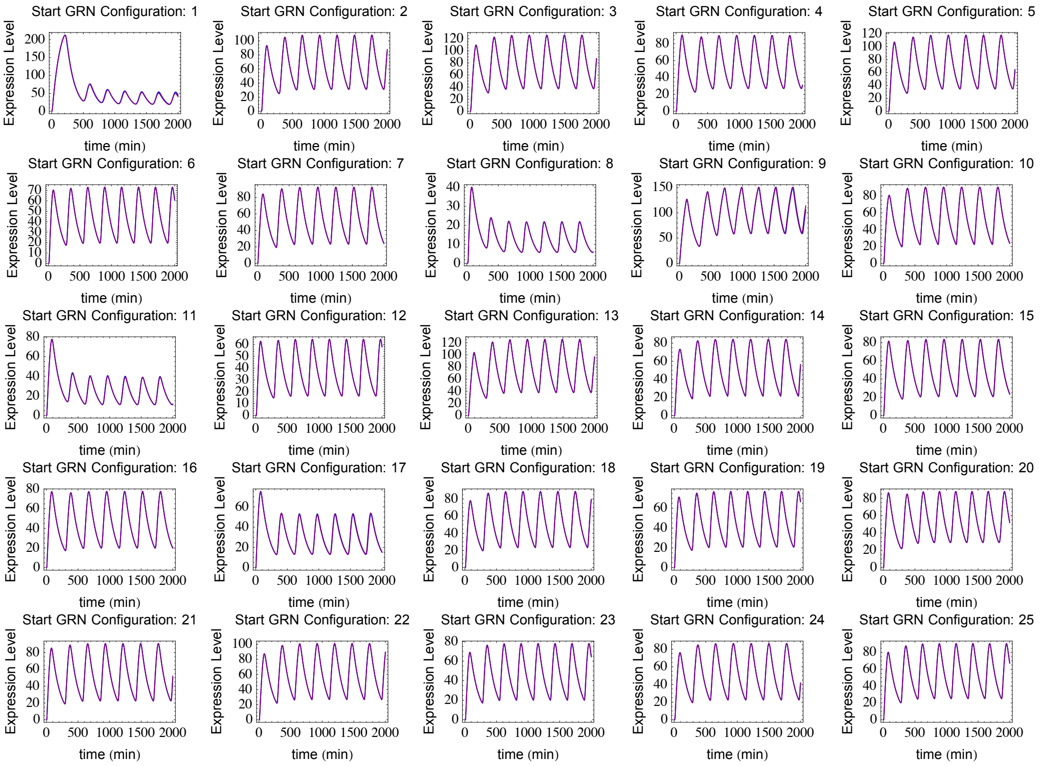


Figure 4.4: Expression profiles of start GRN configurations 1:25. *Phenotypic time series for start haploid (blue) and diploid (red) configurations. Note that haploid and diploid system configurations encode for virtually the same (IF) oscillatory expression phenotype, with only small discrepancies being noticed for a few cases that arise due to numerical integration errors (see Figure 4.5)*

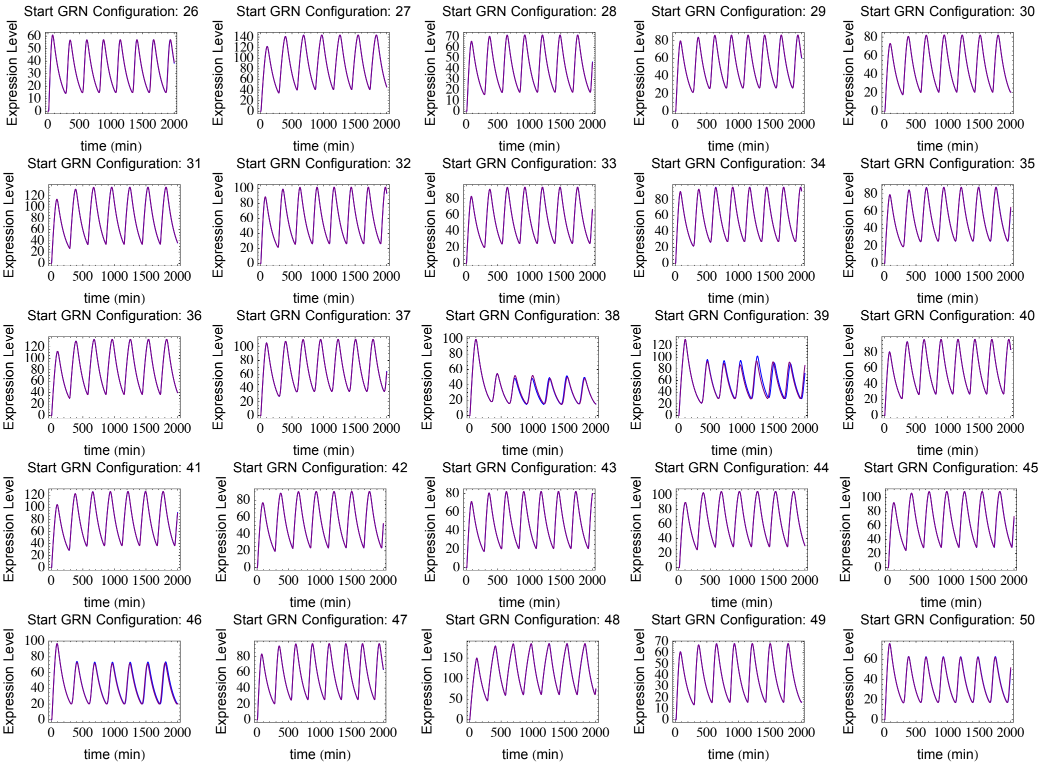


Figure 4.5: Expression profiles of start GRN configurations 26:50. *Phenotypic time series for start haploid (blue) and diploid (red) configurations. Note that haploid and diploid system configurations encode for virtually the same (IF) oscillatory expression phenotype, with only small discrepancies being noticed for a few cases that arise due to numerical integration errors*

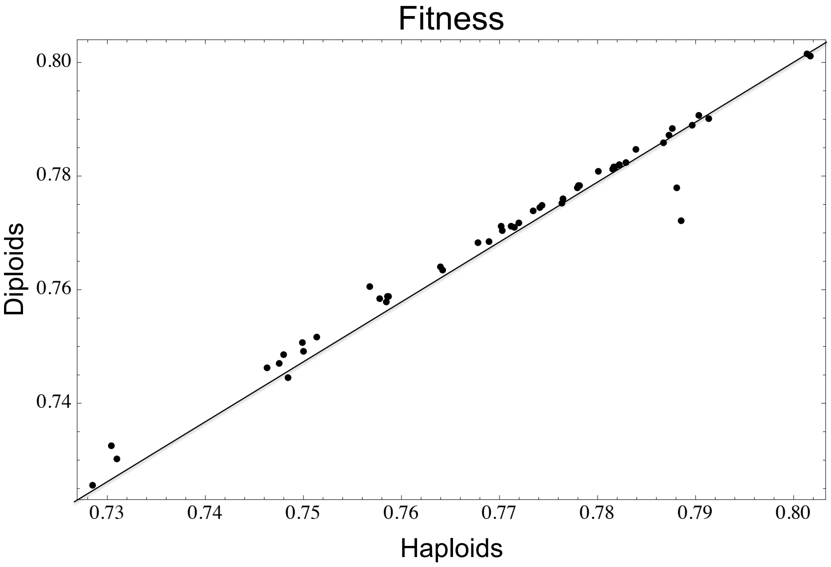


Figure 4.6: Fitness scores of start (IF) pre- and post-duplication GRN configurations. *Scatterplot for fitness scores associated to start haploid and diploid configurations. Scores were evaluated using the multi-objective fitness function FF1. Each point in the plot describes the associated value in the pre- and post-duplication GRN system configuration.*

“If you are faced by a difficulty or a controversy in science, an ounce of algebra is worth a ton of verbal argument”

J. B. S. Haldane

5

Evolvability of GRNs before and after whole genome duplications

Abstract

Whole genome duplications (WGDs) have played a prominent role in the expansion and diversification of gene families across different species, specially in flowering plants. This finding has raised many speculations as to the impact of WGDs on the evolutionary potential of biological systems. In particular, the link between WGDs and the evolvability (capacity to evolve novel phenotypes) of molecular systems, such as gene regulatory networks (GRNs), remains at this stage rather circumstantial, mainly due to our fragmentary knowledge on the mechanistic underpinnings of the genotype-phenotype map (GPM). Here we use the fine-grained, mechanistic GPM modeling framework described in chapter 4 to investigate the impact of WGD on the evolvability of GRNs with oscillatory expression phenotypes. We performed evolutionary explorations across genotype space mimicking the adaptation of individual haploid GRNs and their duplicated versions (resembling a WGD) toward newly imposed phenotypic optima (*e.g.* higher and lower frequency oscillatory expression phenotypes). Based on this, we compared the efficiency of GRNs to attain high fitness levels (our operational definition of evolvability) before and after WGD. Our simulation results reveal that duplicated GRN system configurations do not necessarily adapt faster than their haploid counterparts. We also found that the evolutionary accessibility of newly imposed phenotypic optima after a WGD is frequently, but not always, improved. Moreover, by monitoring mutational pathways that attain high fitness levels we were able to describe the dynamics of sequence substitution at the *cis* (*i.e.* gene promoters) and *trans*-regulatory regions (DNA binding domains) of evolving GRNs, and characterize their associated distribution of fitness effects. Overall, our results demonstrate that the impact of a WGD on the evolvability of oscillatory GRNs depends on a complex interplay between initial evolutionary conditions determined by genetic and non-genetic factors, such as the underlying structure of a start GRN genotype, the nominal values of (partly) environmentally determined network control parameters, as well as quantitative aspects of the newly imposed phenotypic optimum.

Author contribution

All content within this chapter was written by myself. It contains the results of a research paper, currently under preparation, which has been designed by me and by professor Steven Maere.

5.1 Introduction

Whole genome duplications (WGDs) have long been appreciated as a major driving force of the evolutionary process (Ohno, 1970). The impact of WGDs on genome architecture and gene content has been extensively studied across the eukaryote phylogeny[372]. In plants, in particular, genome analyses have identified a plethora of ancient WGD events[107,319,324,373,374], which have been linked to the expansion and diversification of many regulatory gene families[315,375–377]. Over the last decade, increasing attention has been paid to the issue of how WGDs have impacted on the structural connectivity of transcriptional regulatory circuits[154,378], protein-protein interaction networks[53,379] and metabolic pathways[154], and promising advances have been made toward deciphering the rules by which molecular networks may have evolved subsequent to WGD events[53,309,380–383].

In stark contrast, evidence is scarce, circumstantial, and contradictory so as to conclusively assign a direct role to WGDs in promoting the evolvability (capacity to generate novel genetically determined phenotypes/functions/responses) of molecular interacting systems[94,184,266,384–386]. In addition, several other hypothesized roles have been attributed to WGD. In particular, WGDs have been associated to an increased adaptive potential of organisms in the short and the long term[90,92,386], they have been linked to an increased capacity to invade new ecological niches[387], increased speciation rates[90,320,388] and survival of mass extinction events[118,389], and they have also been attributed a role in the evolution of phenotypic innovations and elaborations of complex developmental systems[117,322,323,389,390]. Likewise, gen(om)e duplications have been assigned a role in the origin of other defining emergent properties of molecular systems, such as modularity and robustness[125,150,266]. Moreover, it has been hypothesized, based on theoretical arguments, that gen(om)e duplications may drastically change the properties of fitness landscapes[257], for instance, by creating extra-dimensional bypasses[286,391] that could effectively facilitate the navigability of the fitness landscape along neutral ridges connecting high fitness peaks[289], thus circumventing low fitness valleys[392].

Although thought provoking, most of the above hypothesized links remain at this stage quite elusive, due to the lack of conclusive evidence, and most critically, because of our yet fragmentary knowledge on the mechanistic underpinnings of the genotype-phenotype map (GPM) of molecular networks[229,393,394]. In essence, WGDs in molecular interacting systems, such as gene regulatory networks (GRNs), set the stage for a complex evolutionary scenario involving dosage balance effects, differential gene loss and retention, epistatic interactions, neutral and adaptive sequence divergence, subfunctionalization and neofunctionalization, functional redundancy and intricate network rewiring events[179,382,383,395], among others. To shed some light into this complex scenario from a mechanistic point of view (*i.e.* by pinpointing cause-effect relationships among systems variables, rather than through mere correlations) mathematical modeling of the GPM of GRNs becomes an indispensable tool (see extended discussion in chapter 3)

Over the last decade, a series of coarse-grained GPM models have been developed to investigate, in particular, the evolutionary origin of emergent system properties of GRNs[186,192,193,197]. Despite being overly simplified representations of real regulatory systems, extensive computational interrogation of the GPM models have led to the accumulation of a great body of knowledge on presumably universal network properties, such as robustness and evolvability. One particular shortcoming of most generic models, which is also prevalent across most systems biology-inspired network models[396,397], is their inability to account for an unambiguous genetic representation of the system under study. To the best of our knowledge, only a handful of GPM models exist that use explicit genetic encodings as their basis for simulating changes in the wiring of the system[252,258,261]. Nevertheless, several important missing aspects of GRNs (*e.g.* their dosage sensitive nature) cannot be adequately accounted for in these sequence-based network models due to their limited resolution on the mechanistic details underlying transcriptional regulation (*e.g.* protein-DNA binding process and combinatorial regulation).

In this study, we use the fine-grained, mechanistic GPM modeling framework described in chapter 4 to investigate the impact of WGD on the evolvability of GRNs with oscillatory expression phenotypes. To achieve this, we conducted evolutionary explorations across genotype space mimicking the adaptation of individual haploid GRNs and their duplicated versions (resembling a WGD) toward newly imposed phenotypic optima, which in the context of the present study are represented by oscillatory expression phenotypes with higher and lower frequency (denoted by HF and LF, respectively). Based on this, we compared the efficiency of GRNs to attain high fitness levels (our operational definition of evolvability) before and after WGD. Our simulation results reveal that duplicated GRN system configurations do not necessarily adapt faster than their haploid counterparts. In addition, we found that the evolutionary accessibility of newly imposed phenotypic optima after a WGD is frequently, but not always, improved. Moreover, our *in silico* evolution experiments shed important light on the sequence divergence process driving the evolution of GRNs toward a newly imposed phenotypic optimum, as well as on the distribution of fitness effects associated to substitutions in the the *cis* (*i.e.* gene promoters) and *trans*-regulatory regions (DNA binding domains) of evolving GRNs. Together, our results reveal the existence of several non-intuitive limiting conditions on the evolvability of pre- and post-duplication GRN system configurations. Specifically, we found a complex interplay between initial evolutionary conditions determined by genetic and non-genetic factors, such as the underlying structure of a start GRN genotype (genetic background), the nominal values of (partly) environmentally determined network control parameters, as well as quantitative aspects of the newly imposed phenotypic optimum, which can severely constrain the adaptation of GRNs.

5.2 Results

5.2.1 *In silico* evolution

Using the evolution protocol described in the previous chapter, our intent was to recreate *in silico* a simple evolutionary scenario (see figure 5.1) encompassing: 1) start (ancestral) GRNs with intermediate frequency (IF) oscillatory expression dynamics (see figure 5.1), which are assumed to be initially well adapted to the current environment (*i.e.* occupying a local fitness peaks); 2) upon a sudden environmental change these configurations become appreciably less fit (now situated on a lower fitness valley or local fitness peak); and 3) are thus required to undergo a phase of re-adaptation by evolving towards a newly imposed phenotypic optimum (either HF or LF-type oscillatory expression dynamics) (see figure 5.1). For the sake of simplicity, the new environment was assumed to remain stable over the simulated evolutionary time window. In essence, our evolutionary simulations entail a systematic exploration of sequence space in search for the newly imposed phenotypic optimum. To do this, we used the MCMC-like evolutionary algorithm described in subsection 4.1.6.2 (see also Figure 5.2) to simulate mutational pathways involving single-nucleotide substitutions mimicking adaptive walks for individual GRNs. Simulations were performed under the two multi-objective fitness functions (FF1 and FF2) previously described (see

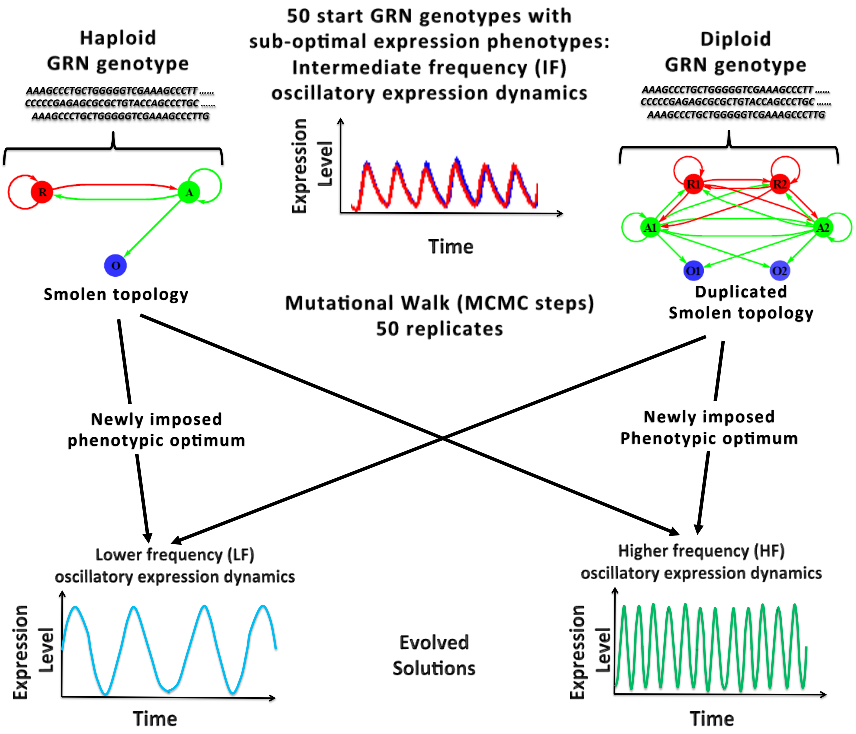

Haploid
GRN genotype
AAAGCCCTGCTGGGGGTCGAAAGCCCTT ......
CCCCCGAGAGCGCGCTGTACCAGCCCTGC ......
AAAGCCCTGCTGGGGGTCGAAAGCCCTTG
R
A
O
Smolen topology
50 start GRN genotypes with
sub-optimal expression phenotypes:
Intermediate frequency (IF)
oscillatory expression dynamics
Expression Level
Time
Diploid
GRN genotype
AAAGCCCTGCTGGGGGTCGAAAGCCCTT ......
CCCCCGAGAGCGCGCTGTACCAGCCCTGC ......
AAAGCCCTGCTGGGGGTCGAAAGCCCTTG
R1
R2
A1
A2
O1
O2
Duplicated
Smolen topology
Mutational Walk (MCMC steps)
50 replicates
Newly imposed
phenotypic optimum
Newly imposed
Phenotypic optimum
Lower frequency (LF)
oscillatory expression dynamics
Expression Level
Time
Evolved
Solutions
Higher frequency (HF)
oscillatory expression dynamics
Expression Level
Time

Figure 5.1: Evolutionary Scenario. *Our point of departure is an ensemble of 50 start GRN configurations composed of minimal linear genomes and network control parameters, such as mRNA and protein half-lives. Although GRN genotypes are widely spread out across sequence space they all encode for Smolen-like topologies with IF-type oscillatory expression dynamics. A diploid start GRN configuration represents a duplicated version of a haploid one, mimicking in this way the outcome of a WGD event. Importantly, due to the genome dosage invariant property of the GPM model (see subsection 4.1.4), a haploid GRN (blue color-coded expression profile) and its duplicated version (red color-coded expression profile) encode for virtually the same IF-type oscillatory expression phenotype at the start of the evolutionary simulations, with only small discrepancies being noticed for a few cases that arise due to numerical integration errors. Therefore, regulatory balance is preserved in the GRNs despite the fact that after a WGD the number of regulatory interactions quadruples (e.g. 5 interactions in the start haploid GRN configuration vs 20 interactions in the start diploid GRN configuration). The start haploid and diploid GRN configurations are then evolved toward either a LF-type or HF-type oscillatory expression dynamics, using the evolution protocol described in subsection 4.1.6.2. Repressor and activator transcriptional regulators are shown in red and green, respectively. The output genes, shown in blue, provide the phenotypic readout of the system. It should be noticed that the time series expression profile of the diploid GRNs is taken as the sum of the expression profiles of both output gene copies.*

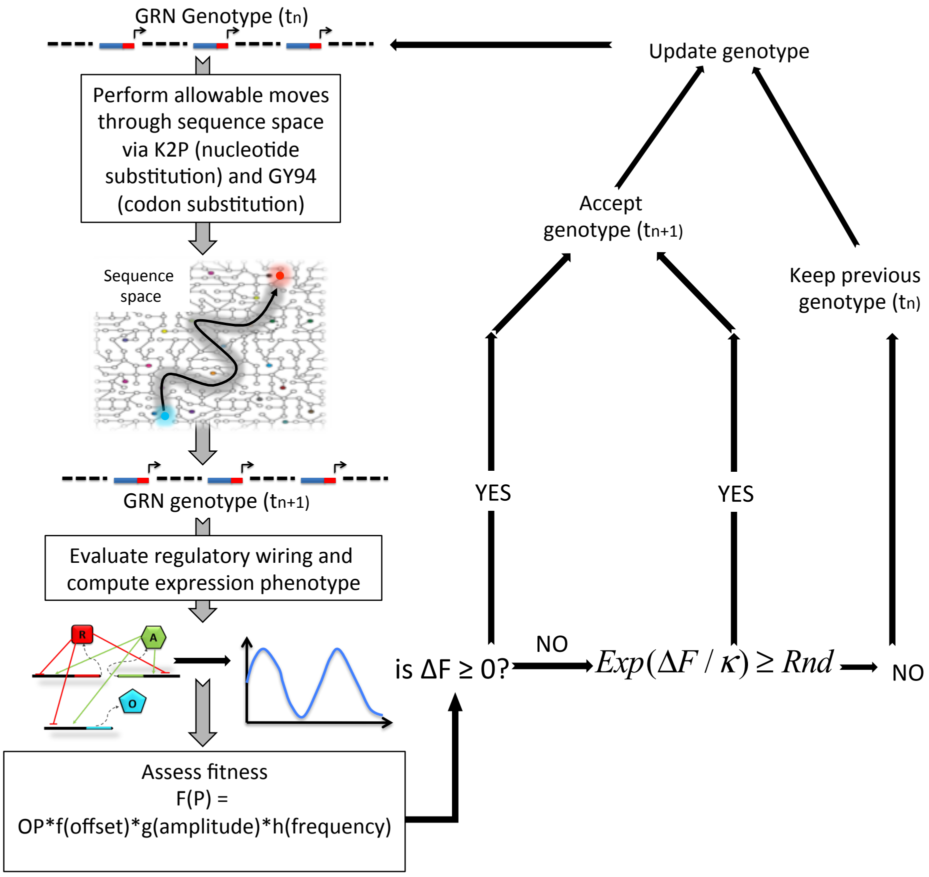

GRN Genotype (t_n)
Perform allowable moves through sequence space via K2P (nucleotide substitution) and GY94 (codon substitution)
Sequence space
GRN genotype (t_{n+1})
Evaluate regulatory wiring and compute expression phenotype
R
A
O
Assess fitness
F(P) =
OP*f(offset)*g(amplitude)*h(frequency)
is ΔF ≥ 0?
NO
$Exp(\Delta F / \kappa) \geq Rnd$
NO
YES
YES
Accept genotype (t_{n+1})
Keep previous genotype (t_n)
Update genotype

Figure 5.2: Evolutionary algorithm. *Schematics of the MCMC-like evolutionary algorithm used to simulate the adaptation of GRNs toward a newly imposed phenotypic optimum. Essential components of the algorithm are 1) a multi-objective fitness function that assesses the functional performance of evolving GRNs based on several quantitative features of the expression phenotype; 2) a Metropolis decision rule; and 2) conventional models of DNA sequence evolution, i.e. the Kimura 2-parameter (K2P) model and the Goldman-Yang 1994 (GY94) model, to perform allowable mutational moves in sequence space. Blue and red points in sequence space indicate the starting (sub-optimal solution) and the ending point (optimized solution), respectively, in a typical evolutionary run.*

subsection 4.1.6.2), which assess different quantitative features of the oscillatory expression phenotypes (see subsection 4.1.6.2). It is worth mentioning that genome dosage invariant expression dynamics is accounted for by our GPM modeling framework (see subsection 4.1.4), which allows us to fairly compare the evolvability of a haploid GRN versus its duplicated system configuration by starting off the simulations from virtually the same fitness level. An important feature of our MCMC-like evolutionary algorithm is that it enables evolutionary pathways to proceed via single nucleotide substitutions that are mostly neutral, or associated with beneficial changes in fitness, with slightly deleterious changes being sporadically accepted (see detailed description in subsection 4.1.6.2). Throughout the manuscript we will refer to this dedicated algorithm as NEA (neutral evolution allowed) to distinguish it from the conventional adaptive walk algorithm conforming to the classical strong selection-weak mutation (SSWM) regime (see description in subsection 4.1.6.5), which simulates evolution toward a new optimum by allowing the sequential fixation of only beneficial substitutions[367,368]. Most importantly, the NEA algorithm allows us to assess the impact of evolution across neutral networks of genotypes[197,272] on the navigability of the fitness landscape (*i.e.* the difficulty of accessing high fitness scoring solutions from sub-optimal configurations via mutational pathways across sequence space) of pre-and post-duplication GRN system configurations.

5.2.2 Impact of WGD on the evolutionary accessibility of newly imposed phenotypic optima

We focus on the analysis of several interesting features of the adaptation process of individual GRNs, such as fitness gains at an early stage of the adaptation process, and the accessibility of newly imposed phenotypic optima (end point fitness values). This allowed us to compare the rate of the adaptation process as well as the efficiency of haploid GRNs vs. duplicated (diploid) GRN system configurations to attain high fitness levels over a longer evolutionary time scale. Figure 5.3(A) depicts the average fitness trajectories of the adaptation process started out from particular haploid and corresponding duplicated GRN configurations. Simulations were conducted using an implementation of the NEA algorithm with the multi-objective fitness function FF1 (see description in subsection 4.1.6.2). The results show that duplication of a start GRN system configuration can have variable impacts on the dynamics and the outcome of the adaptation process (all fitness trajectories simulated from the 50 distinct start GRN configurations are shown in supplementary figures C.2-C.6). Interestingly, a few cases were observed where (individual) haploid GRNs adapt significantly faster than the diploid counterparts, as evidenced by the appreciably long delays in the adaptation process experienced by the diploids compared to the haploids (see Figure 5.3(A), cases GRN Genotypes 1: IF $\rightarrow$ LF, and GRN Genotypes 39: IF $\rightarrow$ HF).

To gain a more general idea as to the prevalence of delays during early stages of the adaptation process we compared the fitness values attained at 5% of the total number of MCMC steps (evolutionary time window) simulated for the haploid GRNs and the corresponding diploid system configurations, both for the LF-type and the HF-type phenotypic optima, across the entire set of start GRN configurations considered (table 5.1). Overall, in a small proportion of the cases tested (3/50 and 10/50 for the LF-type and HF-type phenotypic optima, respectively) we found that diploids adapt significantly slower than haploids. These results indicate that, under specific conditions, haploid GRNs may have an effective advantage, in terms of higher fitness values being attained, over duplicated GRN system configurations during initial stages of the adaptation process. One of the factors that could explain these cases of slower adaptation dynamics in diploids compared to haploids is the presence of initially redundant gene copies in recently formed duplicated GRNs, which may render the system intrinsically buffered against the effect of any potentially advantageous mutation. Therefore, it is likely that for certain duplicated GRN system configurations to escape from the buffered state the initially redundant gene copies must diverge via neofunctionalizing changes in *cis* and/or *trans*-regulatory sequences. This sequence divergence process may imply longer waiting times for certain diploids compared to the haploid counterparts before considerably fitness gains can be achieved over the course of evolution. Another way to interpret this is in the context of genotype networks theory[197,272], which conceives genotype space as being organized into sets of mutationally interconnected genotypes that encode for distinctive phenotypes[197,272]. Accordingly, given that the sequence space of GRNs undergoes an automatic expansion upon a WGD, the extent of the neutral networks associated to particular phenotypes might tend to increase proportionally, decreasing the chances for the system to experience a significant change in fitness within a relatively small evolutionary time period.

Next, we concentrate on the analysis of the end-point fitness values attained. Figure 5.3(B) illustrates the apportionment over evenly sized ranges of the end point fitness values attained from individual start GRN configurations. Our analysis demonstrates that the accessibility of high fitness levels within the evolutionary time windows considered can heavily depend on the intricacies of the start GRN configuration, such as the interplay between current position in genotype space and the nominal values of network control parameters (*i.e.* basal kinetic rates), which altogether determine the dynamical behavior of a GRN. Moreover, the analysis reveals that the accessibility of high fitness scoring solutions becomes considerably more difficult for the HF-type than for the LF-type oscillatory expression dynamics, which applies both for the haploid GRNs and the duplicated system configurations. Statistical analysis (see table 5.1) indicate that in 76% (28/50) and 84% (47/50) of the start GRN configurations evolved toward LF and HF, respectively, significantly higher end-point fitness values were attained upon WGD. Our simulation results thus demonstrate that, under several distinct initial conditions, a WGD event generally, but not always, improve the evolvability of oscillatory GRNs.

To have a broader perspective on the impact of duplication of a GRN system configuration on the adaptation process, we have conducted evolutionary explorations across genotype space under fitness function FF2, which is more stringent than FF1 with respect to amplitude requirements in oscillatory expression dynamics (see details in subsection 4.1.6.2). Using this fitness function we simulated mutational pathways for the same ensemble of start GRN configurations. Examination of these results revealed that for particular start GRNs the average fitness trajectories can differ substantially under the two fitness functions. (see Figure C.7). Moreover, in agreement with our previous simulation results obtained with FF1, we observed a few cases where duplicated GRN system configuration adapt slower than the haploid counterparts (see fitness trajectories for GRN Genotypes 1: IF LF and GRN Genotypes 39: IF LF and IF HF in Figure 5.4(A). Overall, we found that under a more stringent fitness function for amplitude requirements in oscillatory expression dynamics a higher proportion of haploid GRNs seem to adapt significantly faster than the duplicated system configurations (see table 5.2). In addition, similar to our previous observations, we found that significantly higher end point fitness values are usually, but not always, attained by the duplicated GRN system configurations (see Figure 5.4(B)) (all the fitness trajectories simulated from the 50 start GRN configurations are shown in figures C.8-C.8).

Note again that the results discussed so far were simulated using the NEA algorithm, which enables mutational pathways to proceed via single-nucleotide substitutions with beneficial and mostly neutral fitness consequences, with slightly deleterious substitutions being accepted sporadically (see subsection

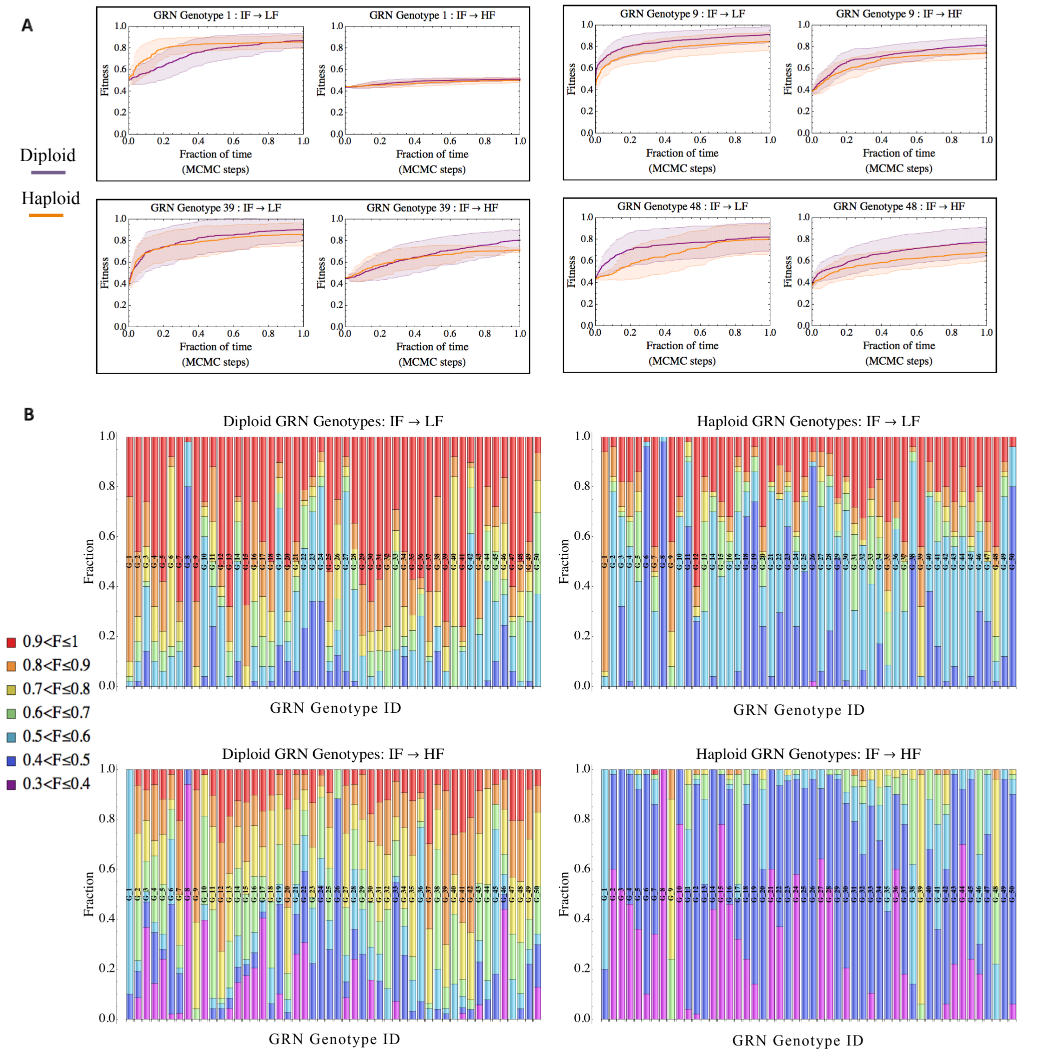
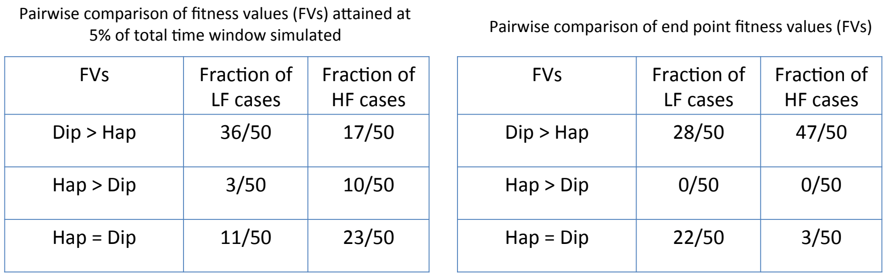


Figure 5.3: Average fitness trajectories: Fitness function FF1. *A, Temporal sequences of fitness values recorded from 50 independent simulation replicates using particular start GRN configurations were averaged out to display the general trend of the adaptation process toward a new phenotypic optimum. Error bars along the trajectories indicate standard deviations. In view of the fact that duplicated (diploid) GRN system configurations present a mutational target twice the size of haploid GRNs, the simulated evolutionary time window for diploids spans twice the number of MCMC steps considered for haploids. Specifically, the length of the simulated mutational pathways was set according to the total number of mutable sites per genome (effective genome size), as follows: haploid GRNs were evolved for 660 (haploid effective genome size) x 5 = 3300 MCMC steps, whereas diploid GRN configurations were evolved for 1320 (diploid effective genome size) x 5 = 6600 MCMC steps. The time (x) axis in the plots shown has been re-scaled to 1. B, distribution of end point fitness values recovered from the 50 evolutionary replicates performed per start GRN configuration. End point fitness values attained were allocated in predefined (color-coded) ranges as indicated by the legend shown on the left-hand side. Fitness scores were computed using the multi-objective fitness function FF1 described in section 4.1.6.2.*



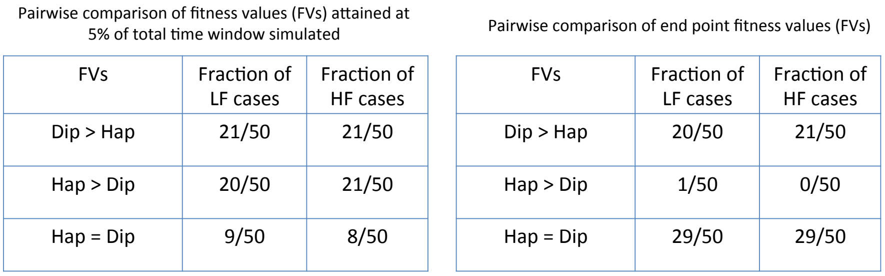
Pairwise comparison of fitness values (FVs) attained at 5% of total time window simulated

FVs	Fraction of LF cases	Fraction of HF cases
Dip > Hap	36/50	17/50
Hap > Dip	3/50	10/50
Hap = Dip	11/50	23/50

Pairwise comparison of end point fitness values (FVs)

FVs	Fraction of LF cases	Fraction of HF cases
Dip > Hap	28/50	47/50
Hap > Dip	0/50	0/50
Hap = Dip	22/50	3/50

Table 5.1: Comparison of fitness values attained at different stages of the adaptation process: Fitness function FF1. *Data shown on the left summarizes the fraction of cases where the different pairwise comparisons for the fitness values attained at* 5% *of the total number of MCMC steps were found to be significant (Dip* $>$ *Hap and Hap* $>$ *Dip) or not (Hap* $=$ *Dip). Data shown on the right summarizes the fraction of cases where the different pairwise comparisons for the end-point fitness values attained were found to be significant (Dip* $>$ *Hap and Hap* $>$ *Dip) or not (Hap* $=$ *Dip). Comparisons were made on the basis of one-sided, Mann-Whitney tests,* $p < 0.05$*, with Bonferroni correction. The comparisons involved fitness values generated using an implementation of the NEA algorithm with the multi-objective fitness FF1, as described in section 4.1.6.2.*



Pairwise comparison of fitness values (FVs) attained at 5% of total time window simulated

FVs	Fraction of LF cases	Fraction of HF cases
Dip > Hap	21/50	21/50
Hap > Dip	20/50	21/50
Hap = Dip	9/50	8/50

Pairwise comparison of end point fitness values (FVs)

FVs	Fraction of LF cases	Fraction of HF cases
Dip > Hap	20/50	21/50
Hap > Dip	1/50	0/50
Hap = Dip	29/50	29/50

Table 5.2: Comparison of fitness values attained at different stages of the adaptation process: Fitness function FF2. *Data shown on the left summarizes the fraction of cases where the different pairwise comparisons for the fitness values attained at* 5% *of the total number of MCMC steps were found to be significant (Dip* $>$ *Hap and Hap* $>$ *Dip) or not (Hap* $=$ *Dip). Data shown on the right summarizes the fraction of cases where the different pairwise comparisons for the end-point fitness values attained were found to be significant (Dip* $>$ *Hap and Hap* $>$ *Dip) or not (Hap* $=$ *Dip). Comparisons were made on the basis of one-sided, Mann-Whitney tests,* $p < 0.05$*, with Bonferroni correction. The comparisons involved fitness values generated using an implementation of the NEA algorithm with the multi-objective fitness FF2, as described in subsection 4.1.6.2.*

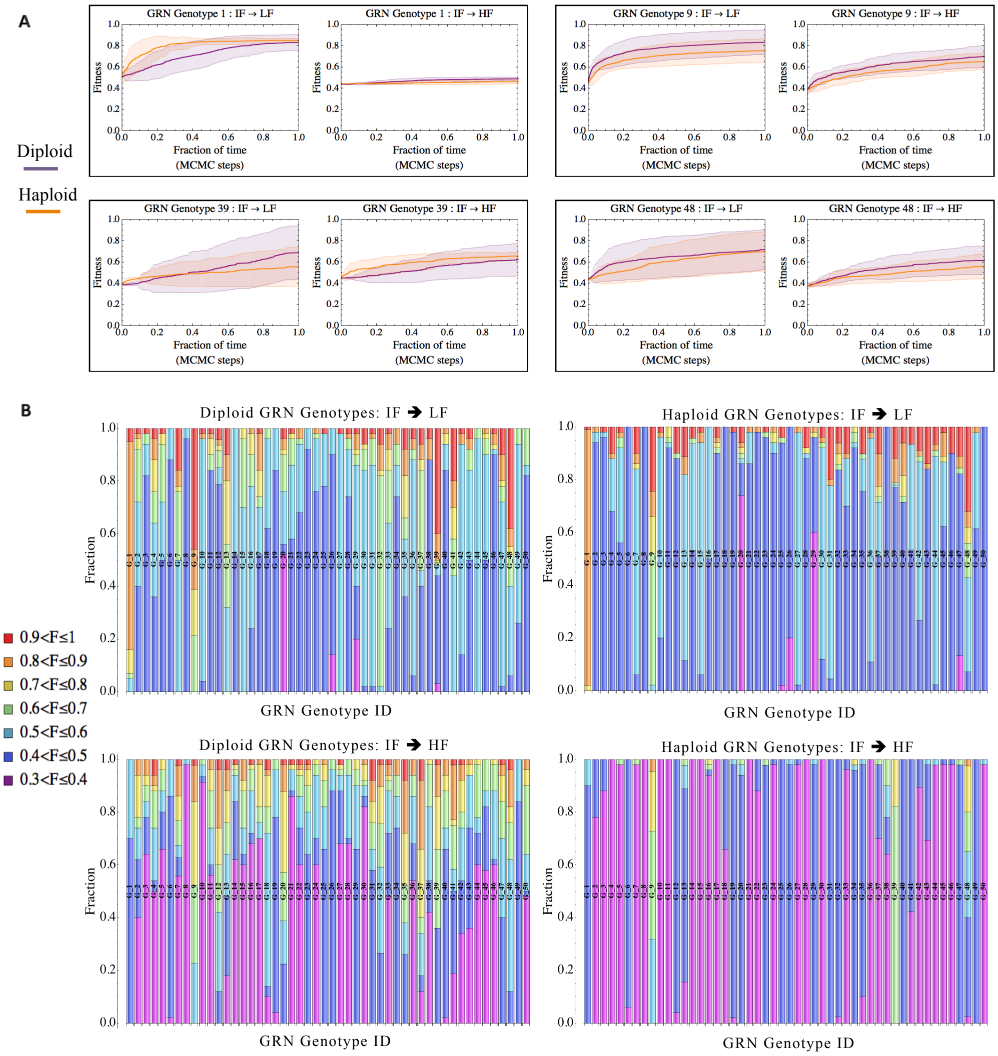


Figure 5.4: Average fitness trajectories: Fitness function FF2. *A, Temporal sequences of fitness values recorded from 50 independent simulation replicates using particular start GRN configurations were averaged out to display the general trend of the adaptation process toward a new phenotypic optimum. Error bars along the trajectories indicate standard deviations. In view of the fact that duplicated (diploid) GRN system configurations present a mutational target twice the size of haploid GRNs, the simulated evolutionary time window for diploids spans twice the number of MCMC steps considered for haploids. Specifically, the length of the simulated mutational pathways was set according to the total number of mutable sites per genome (effective genome size), as follows: haploid GRNs were evolved for 660 (haploid effective genome size) x 5 = 3300 MCMC steps, whereas diploid GRN configurations were evolved for 1320 (diploid effective genome size) x 5 = 6600 MCMC steps. The time (x) axis in the plots shown has been re-scaled to 1. B, distribution of end point fitness values recovered from the 50 evolutionary replicates performed per start GRN configuration. End point fitness values attained were allocated in predefined (color-coded) ranges as indicated by the legend shown on top of the panels. Fitness scores were computed using the multi-objective fitness function (Fitness-F2) described in section 4.1.6.2.*

4.1.6.2). To complement our analyses, we examined the evolutionary accessibility of HF-type and LF-type phenotypic optima from particular start GRN configurations under more restrictive conditions. To do this we simulated adaptive walks conforming to the strong selection-weak mutation (SSWM) regime (see description in subsection 4.1.6.5), which forces evolution toward a new optimum to proceed via the sequential fixation of only beneficial substitutions[367,368]. Overall, the results demonstrate that under such restrictive evolutionary conditions the accessibility of high fitness levels becomes considerably more difficult for both haploid and diploid GRNs (see figure C.13). In fact, we found that under the SSWM regime most of the simulated adaptive walks consistently attained significantly lower end point fitness values (one-sided Mann-Whitney tests, $p << 0.05$) than those attained under the less restrictive NEA algorithm (see figure C.14). Interestingly, we noticed that when evolution toward a new optimum proceeds through the fixation of only beneficial substitutions, haploid GRNs usually attained significantly higher end point fitness values than the duplicated system configurations. This is perhaps due to the fact that it becomes considerably more difficult, without the intervention of neutral evolution, to come across beneficial mutations in the presumably more extensive neutral networks of genotypes associated to the duplicated system configurations. Taken together, our results underscore the crucial role of neutral evolution in the adaptation of biological systems, and suggest that neutral divergence may have been a major determining factor in the establishment and, paradoxically, evolutionary success of polyploids, if this success was adaptive in nature.

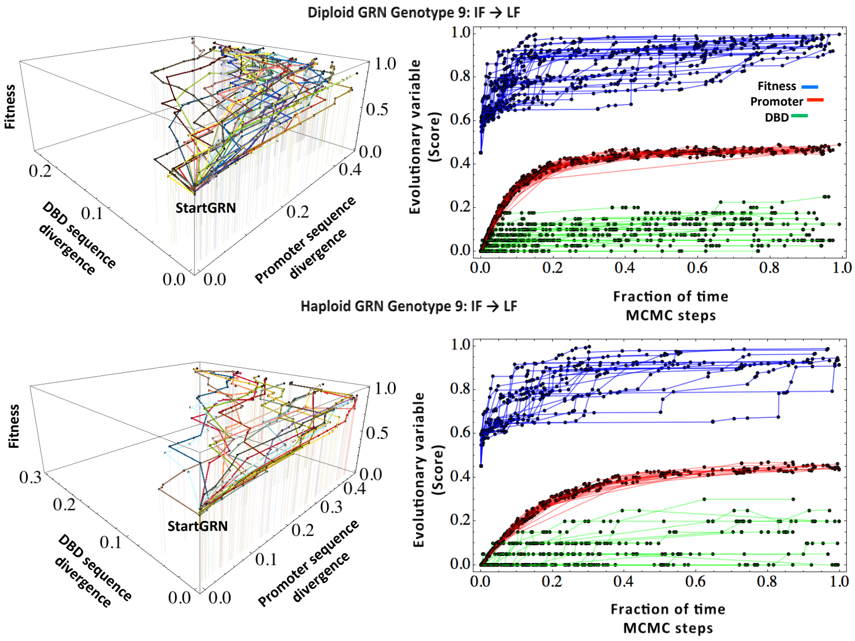
5.2.3 Shedding light on the sequence divergence process driving the evolution of GRNs toward a new phenotypic optimum

Evolutionary adaptation is essentially driven by a sequence divergence process through which increasingly better adapted phenotypes emerge over time. In this section we examined the sequence divergence process that drives the evolution of GRNs toward a new phenotypic optimum. In particular, we asked the following questions: are there general trends in the sequence divergence process? What is the dynamic signature of the sequence divergence process at the *cis*-regulatory regions (gene promoters) and the *trans*-acting elements (DBDs)? How does such dynamic signature compare between evolving haploid GRNs and duplicated system configurations? To address these questions we conducted a detailed analysis of the single-substitution mutational pathways discussed above. Initially, to visually check for general trends in the sequence divergence process, the mutational pathways were projected onto what we term the evolutionary phase space of GRNs. Figures 5.5 and 5.6 (panels shown on the left) portray the so-defined evolutionary phase space where multiple independent mutational pathways are projected for particular start GRN configurations (see also Figures C.15, C.16, C.17, C.18). The resulting trajectories reveal that although the sequence divergence process can take place in many different ways, the projected mutational pathways tend to cluster, forming a cloud resembling a banana-shaped structure, which is heavily biased toward the axis representing the sequence divergence of promoter regions. The emergence of this pattern is linked to the markedly different mutational saturation dynamics of *cis*-regulatory regions and *trans*-acting elements in the GRNs. More specifically, such discrepancies are due to the fact that *cis*-regulatory regions represent mutational targets that likely harbor many more neutral sites than the *trans*-acting elements. To gain more insight into this, we created plots depicting the evolutionary time points (t_i) at which each component of the triplet $(x(t_i), y(t_i), z(t_i))$ projected on the phase space was recorded along a given mutational pathway that attained a high fitness peak (figures 5.5 and 5.6, panels shown on the right). The plots reveal that the divergence process at *cis* and *trans* regulatory sequences possess clearly distinctive dynamic signatures. Evidently, the promoter sequence divergence curves display a temporal behavior with rapid saturation. Also note that as evolving GRNs diverge away from the ancestral configurations the promoter regions consistently tend to reach a saturation point at around 0.45, which is usually achieved within a small fraction (~ 0.25) of the entire evolutionary time window considered. The rapid saturation featured by these sequence divergence curves is again linked to the fact that *cis*-regulatory regions possess a relatively greater potential to accumulate neutral changes compared to the *trans*-acting sequences.

Overall, the distinctive dynamic signatures of the sequence divergence between the promoter regions and the DBD sequences are most likely the result of differential functional constraints. Because nonsynonymous substitutions in DBDs are likely to be more pleiotropic (are less often neutral) than *cis* regulatory changes, the sequence divergence process would be more tightly constrained. To gain further insight into this, we simulated random mutational walks, which allow us to examine the dynamic signature of the sequence divergence process under relaxed selective constraints (see Figure C.21). The results clearly demonstrate that, similar to promoter regions, the divergence of the DBD encoding sequences under relaxed selective constraints tend to rapidly attain a saturation point, demonstrating the important role of the functional constraints imposed by adaptive evolution in the differential dynamics of the sequence divergence at the *cis*-regulatory regions and the *trans*-acting elements. Lastly, in order to gain insight into the extent of sequence divergence under more restrictive conditions we analyzed adaptive walks simulated under the classical SSWM regime[367,368]. Not surprisingly, the results contrast with our previous observations in several respects (see Figure C.20). In particular, we noticed that the trajectories described through the evolutionary phase space tend to be more irregular but usually highly reproducible. In addition, due to the scarcity of strictly beneficial mutations, which makes them more difficult to come across over the course of evolution toward a new optimum, the sequence divergence curves display a radically different temporal behavior. For instance, it is clear that promoter sequences can accumulate only a limited number of substitutions, which gives rise to sequence divergence curves that are far from reaching saturation, whereas the maximum number of substitutions observable for the DBD regions is usually less than 2. Taken together with the fact that SSWM walks generally lead to lower endpoint fitnesses than NEA walks, these results underscore the importance of neutral divergence in the evolvability of GRNs.

5.2.4 Distribution of fitness effects associated to substitutions at *cis* and *trans* sequences

The distribution of fitness effects is a key quantitative signature of any adaptation process[398,399], the quantification of which may reveal salient features of adaptive walks and properties of fitness landscapes. To gain insight into the distribution of fitness effects for the haploid and diploid GRN system configurations evolving toward a new phenotypic optimum, the fitness effect of every single nucleotide substitution tested over the entire course of evolution toward a new optimum was recorded. Figure 5.7(A) displays the distributions of beneficial fitness changes ($\Delta F > 0$) elicited by single nucleotide substitutions at the *cis*-regulatory regions and at the *trans*-acting elements of evolving GRNs. The distributions shown represent the fitness effects of adaptive substitutions recorded from adaptive walks simulated from particular start GRN configurations toward the LF-type phenotypic optimum. In general, our results indicate that a greater proportion of larger beneficial fitness changes tend to be elicited through sequence divergence at



Diploid GRN Genotype 9: IF → LF

Haploid GRN Genotype 9: IF → LF

Figure 5.5: Evolutionary phase space and sequence divergence plots: GRN genotype 9. *Panels on the left depict the phase space of evolving GRNs. Coordinates shown represent the proportion of accumulated changes (with respect to start GRN configurations) in DBD encoding sequences and promoter regions (x, y coordinates), as well as their associated fitness score (z coordinate). Sequence divergence is assessed with respect to the start GRN configurations using a normalized Edit distance, which measures the percentage of dissimilarity (in terms of DNA sequence for promoters and of amino acids for DBDs) between a given ancestral sequence and mutant sequences sampled at a given time point (t_i) over the course of evolution. The concatenation of a sequence of triplets $(x(t_i), y(t_i), z(t_i))$ sampled at different time points over an evolution run describes a trajectory across the phase space. Triplets $(x(t_i), y(t_i), z(t_i))$ projected on the evolutionary phase space were sampled every time there was a fitness increment of $\Delta F \geq 0.001$ over the course of evolution toward a new optimum. Panels shown on the right illustrate the temporal sequence for individual components in the triplets $(x(t_i), y(t_i), z(t_i))$ recorded among different replicates. Time is depicted as the fraction of the total number of MCMC steps simulated. The y axis represents either percentage of dissimilarity between ancestral sequence and mutant sequences (for both promoter regions and DBD encoding sequences, separately) or the fitness value sampled at a given time point (t_i). Evolutionary simulations were performed following the rules specified in the NEA (neutral-evolution-allowed) algorithm, and using fitness function FF1 (see subsection 4.1.6.2).*

the *trans*-acting elements of evolving haploid GRNs compared to duplicated system configurations. Note, for instance, the surprisingly large proportion ($\sim 40\%$) of *trans*-substitutions associated with $\Delta F > 0.1$ recorded over the course of evolution toward the new optimum started from the haploid GRN genotypes

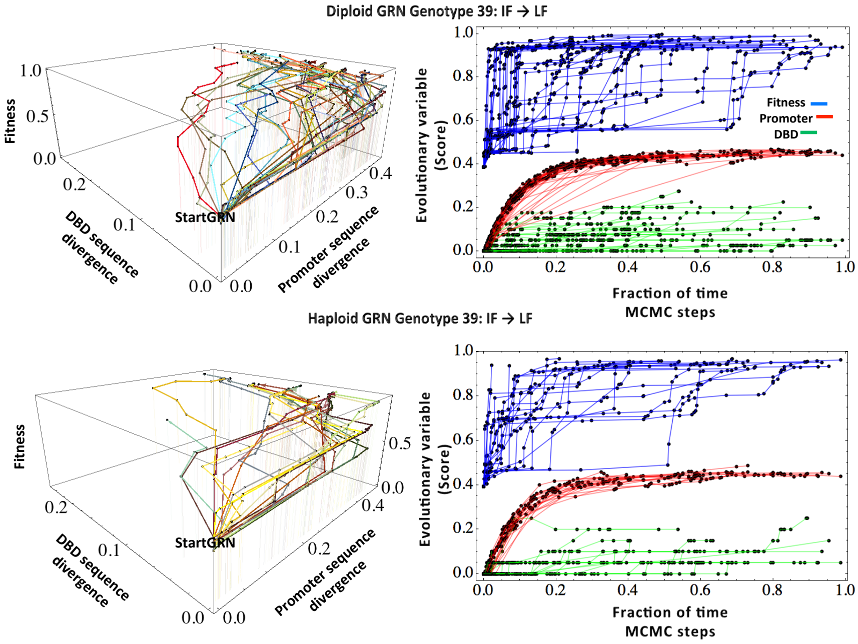


Figure 5.6: Evolutionary phase space and sequence divergence plots: GRN genotype 39. *Panels on the left depict the phase space of evolving GRNs. Coordinates shown represent the proportion of accumulated changes (with respect to start GRN configurations) in DBD encoding sequences and promoter regions (x, y coordinates), as well as their associated fitness score (z coordinate). Sequence divergence is assessed with respect to the start GRN configurations using a normalized Edit distance, which measures the percentage of dissimilarity (in terms of DNA sequence for promoters and of amino acids for DBDs) between a given ancestral sequence and mutant sequences sampled at a given time point (t_i) over the course of evolution. The concatenation of a sequence of triplets $(x(t_i), y(t_i), z(t_i))$ sampled at different time points over an evolution run describes a trajectory across the phase space. Triplets $(x(t_i), y(t_i), z(t_i))$ projected on the evolutionary phase space were sampled every time there was a fitness increment of $\Delta F \geq 0.001$ over the course of evolution toward a new optimum. Panels shown on the right illustrate the temporal sequence for individual components in the triplets $(x(t_i), y(t_i), z(t_i))$ recorded among different replicates. Time is depicted as the fraction of the total number of MCMC steps simulated. The y axis represents either percentage of dissimilarity between ancestral sequence and mutant sequences (for both promoter regions and DBD encoding sequences, separately) or the fitness value sampled at a given time point (t_i). Evolutionary simulations were performed following the rules specified in the NEA (neutral-evolution-allowed) algorithm, and using fitness function FF1 (see subsection 4.1.6.2).*

9 and 39. This difference may be linked again to the presumably increased buffering capacity of diploid GRNs conferred by the presence of initially redundant duplicate gene pairs in the start GRNs. The in-

trinsic buffering capacity of duplicated GRN system configurations may render the large majority of substitutions at *trans*-acting elements only mildly advantageous during a considerable fraction of the entire evolutionary time window considered, before duplicate gene pairs have diverged to some extent. Also worthy of notice is the fact that the shapes of the distributions of fitness effects associated to substitutions at the *trans*-acting elements are quite heterogeneous, which may be linked to the intricacies (*i.e.* genetic background/position in sequence space) of the start GRN configurations. On the other hand, one can notice that the distributions of fitness effects associated to *cis*-regulatory changes tend to be heavily skewed to the left, implying that the large majority of the single nucleotide substitutions bring forth only very slight fitness increments, which seems to be congruent with previous experimental observations[398,399]. Only rarely, we observed that a small proportion ($10\% - 15\%$) of the substitutions at the *cis*-regulatory regions are associated with large fitness changes ($\Delta F > 0.1$), especially in the haploid GRN system configurations.

We also examined the extent to which the rate of substitutions occurring at *cis* and *trans*-regulatory regions was influenced by the mutational target size. To do this, we counted the total number of adaptive substitutions, realized over a given mutational walk, and divided this number by the total extent (given in number of nucleotides) of the *cis*-regulatory regions or the *trans*-acting elements in a haploid/diploid GRN. Figure 5.8(B) illustrates the distributions of these rates for the 50 replicates simulated for different start configurations. We found that in 3 out of the 4 diploid GRN system configurations (genotypes 9, 39 and 48) analyzed the rate at which adaptive substitutions occur at the *trans*-acting elements tend to be higher/equal than the rate of those occurring at the *cis*-regulatory regions (one-sided Mann-Whitney test, $p > 0.1$). This could be due to the fact that diploid GRN system configurations require the duplicate gene pairs to neofunctionalize via sequence divergence at the *trans*-acting elements in order to escape from an initially buffered state that renders the system insensitive to the effects of (beneficial) mutations. In contrast, we found that in just 1 out of the 4 haploid GRN system configurations (genotype 48) the rate at which adaptive substitutions occur at the *trans*-acting elements tend to be higher/equal than the rate of those occurring at the *cis*-regulatory regions (one-sided Mann-Whitney test, $p > 0.06$). Furthermore, we found that the rate of neutral substitutions occurring at the *cis*-regulatory regions tend to be exceedingly higher than those occurring at the *trans*-acting elements (see Figure 5.8).

5.2.5 Quantitative design features of high fitness scoring solutions

Lastly, given that two distinct oscillatory expression phenotypes were set as novel phenotypic optima (LF and HF), it is interesting to ask whether high fitness scoring solutions have been assembled throughout evolution in such a way that they can be allocated into clearly distinguishable classes. To shed light on this, we have examined the design features of diploid GRN system configurations that attained fitness scores $F > 0.9$. As expected for complex network models, we found that many distinct solutions exist for a given functional task. To check for differences in network design features we need a compact description of GRNs in terms of sequence-encoded microscopic features (association constants) for all possible protein-DNA binding events. The rationale behind this is that a sustained oscillatory expression phenotype is elicited through the concerted action of all the time varying regulatory signals within a GRN, which must be in proper balance. Relying on this rationale, for every high fitness scoring solution we computed the ratio of the aggregated DNA binding strength of activating TFs to the aggregated DNA binding strength of repressing TFs (see Figure 5.9). Using these ratios we noted that LF-type and HF-

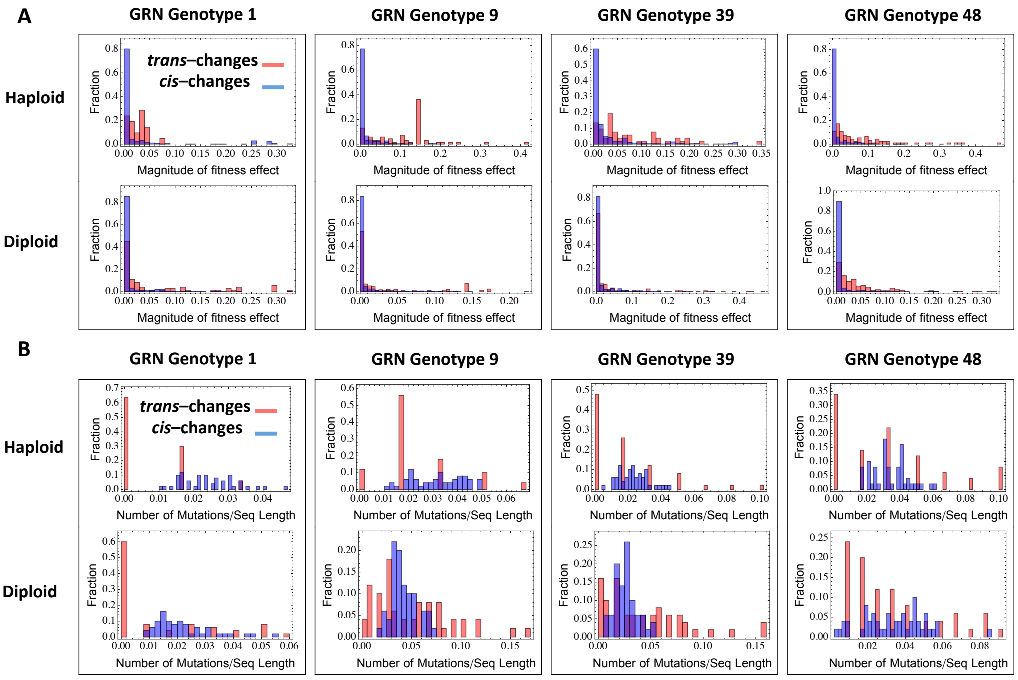


Figure 5.7: Distribution of fitness effects and rate of beneficial substitutions at cis and trans-regulatory sequences. *A illustrates the distribution of beneficial fitness effects ($\Delta F > 0$) associated to adaptive substitutions occurring at the cis-regulatory regions and the trans-acting elements of GRNs evolved toward the LF-type phenotypic optimum, from different start genotypic configurations. B illustrates the distributions describing the rate at which adaptive substitutions occur in cis and trans over the entire course of a mutational walk toward the new optimum (data shown is for the 50 simulation replicates considered per start GRN configuration). The size of the mutation targets considered is as follows: for haploid GRNs the cis-mutation target = 600 nucleotides long and the trans-mutation target = 60 nucleotides long; for diploid GRNs the cis-mutation target = 1200 nucleotides long and the trans-mutation target = 120 nucleotides long.*

type GRN solutions can be clearly distinguished. For instance, we found that in high fitness scoring LF-type GRN configurations the aggregated DNA binding strength of activating TFs was, on average, 2.67 orders of magnitude higher than the aggregated DNA binding strength of repressing TFs, whereas in the HF-type solutions the same relation was, on average, 3.16 orders of magnitude higher (see Figure 5.9). This result demonstrates that evolution of high fitness scoring GRNs with distinctive oscillatory features favors the acquisition of differently parameterized wirings (see Figure 5.9). In other words, evolution toward a newly imposed phenotypic optimum not only promotes changes in the connectivity of GRNs (rewiring) but also extensive quantitative diversification of the strength of regulatory interactions. Not surprisingly, we found that haploid GRNs can assume only a few distinct oscillatory wirings, which exhibit different parameterization of the regulatory linkages (see figures C.22 and C.23).

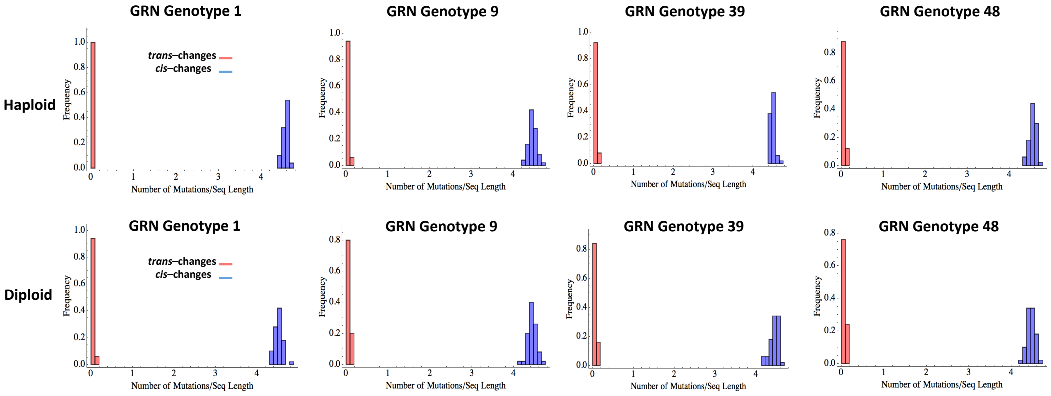


Figure 5.8: Neutral substitutions at *cis* vs. *trans*-regulatory sequences. *The distributions shown describe the rate at which neutral substitutions occur at cis and trans-regulatory sequences over the entire course of a mutational walk toward a new optimum (data shown is for 50 simulation replicates considered per start GRN configuration evolved toward the LF-type optimum). The size of the mutation targets considered is as follows: for haploid GRNs the cis-mutation target = 600 nucleotides long and the trans-mutation target = 60 nucleotides long; for diploid GRNs the cis-mutation target = 1200 nucleotides long and the trans-mutation target = 120 nucleotides long.*

5.3 Discussion

In this study we developed a fine-grained mechanistic GPM modeling approach to study the evolution of GRNs. Unlike most conventional GPM models, the point of departure in our modeling approach is an explicitly defined genotypic encoding, which provides the basis for simulating the evolution of individual GRNs across sequence space. By conducting extensive simulation experiments that mimic possible evolutionary trajectories toward a newly imposed phenotypic optimum, we were able to derive quantitative estimates on the impact of WGD on the evolutionary accessibility of high fitness scoring oscillatory expression phenotypes. Our results provide numerical evidence that, we believe, sheds new light on the navigability of the fitness landscape and the evolvability of GRNs subsequent to WGD events. Moreover, our study offers fresh insights into the distribution of fitness effects associated to single nucleotide substitutions occurring at the *cis* and the *trans* regulatory components of GRNs.

Contrary to popular belief, analyses of single substitution mutational pathways describing the evolution of GRNs toward a new optimum demonstrate that duplication of a system configuration does not necessarily speed up the adaptation process. A possible explanation for this observation may be linked to the existence of buffering mechanisms in diploid GRNs which, under certain conditions, could prevent these systems from adapting faster than haploid GRNs. Several arguments could be invoked to explain these observations. For instance, it could be argued that duplicated components may render an evolving biological system more robust to changes, and thus less prone to experience significant improvements in functional performance[272]. In fact, it is widely believed that a molecular interacting system with functionally redundant duplicate genes would tend to be resilient to several sources of perturbations, including

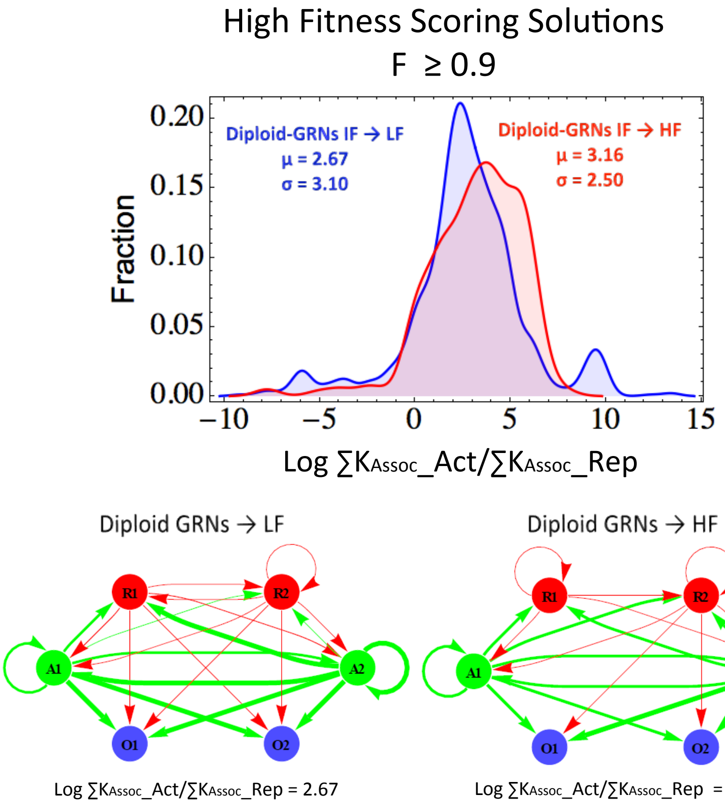


Figure 5.9: Quantitative design principles of high fitness scoring diploid GRN system configurations. *Distributions depicted on top represent the ratio (in logarithmic scale) of aggregated DNA binding strength of activating TFs to the aggregated DNA binding strength of repressing TFs, among high fitness scoring diploid GRN system configurations evolved toward both the LF-type and the HF-type phenotypic optima. Statistics shown were computed using Gaussian kernel density estimators of the empirical distributions. Wirings shown at the bottom represent instances of high fitness scoring solutions. The thickness of the edges in the regulatory wirings displayed is proportional to the aggregated DNA binding strength of a given TF over all possible binding sites on the promoter regions of the target genes. Repressor and activator transcriptional regulators are shown in red and green, respectively; output genes are shown in blue.*

potentially beneficial and deleterious mutations[90,92,400]. Another factor that may explain our results relies on the idea that raising the dimensionality of the fitness landscape, for instance via a WGD event, is likely to result in the proliferation of local fitness maxima (fitness barriers), which may render the accessibility of increasingly higher fitness levels a rather improbable event[230,401]. On the other hand, it has been argued that an increase in the dimensionality of genotype space is expected to result in a concomitant increase in the number of accessible mutational pathways toward a new optimum[402–404], by facilitating the transition between local optima and the global optimum in rugged fitness landscapes[257]. Our results do not conclusively support either of these ideas. Instead, we found that the rate at which pre- and post-WGD GRNs adapt toward a newly imposed optimum can be highly context-dependent. Specifically, we found a complex interplay between initial evolutionary conditions determined by genetic and non-genetic factors, such as the underlying structure of a start GRN genotype (genetic background), the nominal values of (partly) environmentally determined network control parameters, as well as quantitative aspects of the newly imposed phenotypic optimum, which can severely constrain the rate of adaptation of GRNs. A more consistent pattern seems to emerge regarding the adaptation of GRNs over long evolutionary time scales. In this case, we found that the evolutionary accessibility of newly imposed phenotypic optima after a WGD is frequently, but not always, improved. An interesting observation was that this improved “long-term evolvability” in post-WGD GRNs is more prevalent when GRNs are required to evolve toward LF-type phenotypic optimum. Although our study only considers two phenotypic optima as evolutionary targets, these results indicate the existence of a bias in the evolvability of certain expression phenotypes. In summary, our study reveals an unanticipated complexity underlying the evolutionary potential of GRNs, and suggests that the evolvability of biological systems possesses an intricate multifactorial basis that can be difficult to dissect through coarse-grained mathematical models of the GPM.

Beyond question, *cis* and *trans* regulatory changes have played a pivotal role in the evolution of physiological[405,406] and morphological[72,75] features. Nevertheless, the existing controversy regarding the relative contribution of *cis* vs. *trans* regulatory changes to adaptive evolution, an in particular following a WGD event, remains far from being resolved[81,82]. Examination of the distribution of fitness effects associated to substitutions at the *cis* and the *trans*-acting sequences of GRNs evolved toward newly imposed phenotypic optima shed some light into this issue. An important insight gained from our analysis is that for particular start haploid GRNs significantly larger fitness gains ($\Delta F > 0.1$) are more frequently achieved through sequence divergence of the *trans*-acting elements compared to the corresponding duplicated GRN system configurations. Due to the presumably increased buffering capacity of diploid GRN system configurations, by virtue of carrying duplicate gene pairs with initially identical regulatory roles, the acquisition of *trans* regulatory changes with relatively large fitness effects may prove more difficult than in haploid GRNs. In particular, it is likely that the acquisition of substitutions with larger fitness effects in the diploid GRN system configurations analyzed might require longer evolutionary time scales, given that they are expected to spend longer periods of time drifting across extended neutral networks of genotypes.

Furthermore, we found that single nucleotide substitutions in the *cis*-regulatory regions of both haploid and diploid GRN system configurations have predominantly mild fitness effects (distributions heavily skewed to the left), which seems to be congruent with previous experimental observations[398,399]. Nevertheless, we observed that a small proportion ($\approx 15\%$) of the substitutions at the *cis*-regulatory regions can bring forth relatively large fitness gains ($\Delta F > 0.1$). Taken together, these observations seem to suggest that the rewiring of GRNs following a WGD event may be achieved through concerted changes in both *cis* and *trans*-regulatory regions that drive the evolution of the system toward a new optimum through gradual fitness increments. To gain more insight into this, we have examined the extent to which the rate of substitutions occurring at *cis* and *trans*-regulatory regions was influenced by their mutational target size (*i.e.* the total sequence length of the *cis* and *trans*-regulatory regions in a GRN). Our analysis indicated that the rate at which beneficial substitutions occur at *trans*-acting sequences tend to be relatively higher/equal than the rate of those occurring at *cis*-regulatory regions in the diploid system configurations.On the other hand, we found that *trans*-divergence is more widespread in diploids (although with smaller fitness effects, on average) compared to haploid GRNs. Together, these observations suggest that, at least in the start GRNs analyzed, the adaptive rewiring of diploid system configurations is likely to take place via neofunctionalizing changes involving sequence divergence of both *cis* and *trans*-regulatory regions.

Although our study provides fresh insights into the evolutionary potential of GRNs, it is fair to say that our mechanistic GPM modeling approach offers only a first look at what is in reality a more complex multidimensional space of mutable network control parameters. For instance, our study has concentrated only on the mutationally accessible parameter space that determines the wiring of GRNs. Therefore, in order to gain further insight into the evolutionary potential of GRNs, the quantitative features of other regulatory layers must be accounted for and adequately incorporated as sequence-encoded parameters in GRN models, in order to assess their evolutionary impact on, for instance, the evolvability of the system. In particular, it would be interesting to examine whether our predictions hold if network evolution is allowed to proceed via mutations capable of modulating cooperative protein-protein interactions involved in DNA binding recognition[359], or in the non-linear degradation of multimeric proteins[362]. Finally, as discussed above, we have focused on studying the impact of gen(om)e duplications on the traversability of fitness landscapes by simulating mutational trajectories describing the evolution of individual GRN system configurations (adaptive walkers). Nevertheless, a more comprehensive picture requires population-based simulations to dissect the potential contribution of important population genetic parameters, such as the effective population size and recombination, as well as variable mutation rates, on the evolvability of GRNs. It is possible that if we consider evolution of populations of GRNs undergoing different sorts of genetic modifications beyond point mutations, such as recombination and large-scale genetic changes, some discrepancies with respect to our current results emerge.

“Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things.”

Isaac Newton

6

Proximate and ultimate consequences of dosage balance alterations in GRNs

Abstract

The functioning of molecular interacting networks can be quite sensitive to changes in the balanced concentration levels among network components. This is particularly true for transcriptional regulatory systems where alterations in gene dosage have been shown to represent a primary source of dosage-dependent expression phenotypes. This important variational aspect of the genotype-phenotype map (GPM) is the central topic of the gene dosage balance hypothesis (GDBH), which proposes a series of principles to explain the mechanistic underpinnings of dosage balance effects in the context of molecular networks. Over the last few years, several ideas elaborated in the GDBH have been tested through quantitative network modeling approaches, yielding novel insights on the dosage dependent functioning of molecular networks. However, we still lack a clear quantitative understanding of the role of dosage balance alterations in the modulation of network dynamics, mainly due to the fact that most current network models fail to capture essential mechanistic details of, for instance, transcriptional regulation. Here we use the mechanistic GPM modeling framework described in chapter 4 to investigate the proximate and ultimate consequences of dosage balance alterations in oscillatory gene regulatory networks (GRNs). We first assessed the immediate fitness impact of single gene duplication and deletion in the ensemble of start GRNs previously used to conduct evolutionary simulations (see chapter 5). Next, we simulated the evolution of GRNs carrying an extra copy of one of the genes (imbalanced GRNs) toward new phenotypic optima, and compared their evolvability with that of haploid and diploid GRN system configurations. Lastly, we examined the impact of single gene duplication, deletion and amplification of gene copies on the expression dynamics of high fitness-scoring GRNs previously evolved. Our results reveal that: 1) under a fixed GRN topology, single gene duplications can give rise to a broad range of phenotypic responses as a function of quantitative differences in the strength of regulatory linkages; 2) due to pervasive detrimental changes in fitness, single duplications of regulatory genes generally have an adverse impact on the evolvability of GRNs, although a few exceptions exist where the imbalanced GRNs outperform their haploid and diploid counterparts (balanced GRNs); 3) in evolved GRNs, gradual modulation of oscillatory expression dynamics can be effectively achieved in response to amplification of the activator-encoding gene, whereas only qualitative changes in expression dynamics are achievable through amplification of the repressor-encoding gene; and 4) in GRNs with duplicate gene copies the fitness impact of dosage balance alterations largely depends on the extent of functional divergence between paralogous regulators. Our study demonstrate that detailed and biologically realistic GPM models are necessary tools to gain mechanistic insight on the proximate and ultimate consequences of dosage balance effects in GRNs.

Important note: this is a work in progress; theres still too many holes to be filled to present this as a finalized piece of work. Additional simulation experiments and analyses are undergoing. The new simulation results will be contrasted and interpreted in the light of those presented in the study described below.

Author contribution

All content within this chapter was written by myself. It contains the results of a research paper designed by me and professor Steven Maere.

6.1 Introduction

Gene copy-number variation, an important source of genetic variation within and among populations[407], may have variable impacts on phenotypes and thus fitness, with a possible bias toward detrimental effects[408]. For instance, increased disease susceptibility has usually been linked to gene duplications[409,410] and gene deletions[411], with deletions being presumably more deleterious than duplications[407,408]. Alternatively, several cases exist where changes in gene copy number have been associated to traits with potential adaptive benefits[412–414]. Moreover, gene copy number variation has been found to cause significant alterations in the concentration of transcripts in humans[415], it has been associated with growth rate changes in bacteria[416,417], and has recently been implicated in the alteration of dynamic aspects of genome-wide expression patterns throughout different developmental stages[418].

A common unifying theme among the aforementioned cases relates to the widespread idea that the functioning of molecular networks behind complex GPMs is typically dosage balance sensitive. This recurring variational property of molecular networks has been extensively elaborated in the gene dosage balance hypothesis (GDBH)[155–157], which aims to explain the mechanistic underpinnings of dosage balance effects[156,157,160]. Essentially, the GDBH posits that abnormal phenotypes result from dosage balance alterations that compromise the functioning of molecular interacting networks, for instance, by altering the binding kinetics and mode of assembly of the components of macromolecular complexes[158,159]. In other words, the GDBH predicts that altered stoichiometric relationships among the components of macromolecular complexes induce drastic reductions of the assembled complex, thus producing unassembled intermediates and free subunits[158], which may have detrimental effects[155,160,161]. From a GRN perspective, the GDBH posits that abnormal expression phenotypes would result from dosage balance alterations that impact on the DNA occupancy profiles of transcriptional regulators at the promoter region of target genes[158,159]. In fact, this observation seems to hold under a wide range of scenarios. Take for instance the case of developmental regulatory cascades where the expression of downstream target genes is controlled by upstream regulators[163,164,419]. Here, alterations in the dosage of any one regulator on top of the hierarchy typically triggers effects that propagate across the entire system, ultimately modulating the expression of a battery of key developmental genes[163,164,419]. Moreover, due to its predictive power, the GDBH has also served as an explanatory framework for evolutionary phenomena, such as the usual preferential retention of certain regulatory and interacting gene classes observed in several lineages that have undergone one or several rounds of genome duplications along their evolutionary history[110,157,161,179].

Because of the complex non-linear nature of the GPM of molecular networks, systems biology-inspired network models have become popular as quantitative tools to interrogate the impact of perturbations (*i.e.* by tweaking the parameters of network models to mimic, somehow, the effect of genetic variation) on the dynamical behavior of a wide range of molecular networks[209,357,358,420,421]. In particular, systems biology-inspired modeling approaches have proven instrumental in the elucidation of network-based mechanisms underlying the dosage dependent functioning of a great variety of cellular information processing systems[158,172,422–428], as well as in the design and optimization of synthetic circuits[172,429–431]. It should be noticed, however, that most present-day network models usually rely on standard coarse-grained approximations to account for transcriptional regulatory processes, wherein essential mechanistic details are assumed to be encapsulated in aggregated parameters, thus limiting a model's scope and predictive power. Perhaps most critical, the use of coarse-grained mathematical representations of molecular networks necessarily implies an arbitrary treatment of genetic perturbations. For instance, it is standard procedure to simulate gene copy number variation by adjusting certain parameters of a fixed model structure in a way reflecting proportional changes in gene dosage (*e.g.* 2-fold changes in the value of a parameter describing the maximal transcription rate of a given gene to mimic a duplication event)[423,426]. Based on this approximation, evolutionary hypotheses have been derived regarding the potential impact of different aspects related to dosage balance effects in molecular networks, such as the link between neutral processes and the organization of signal transduction pathways into different classes of dosage sensitive signaling proteins[432], as well as the origin of functional innovation in regulatory network motifs[423]. Nevertheless, our knowledge on the potential role of dosage balance effects in, for instance, the evolvability of molecular networks remains largely fragmentary (but see[382,433] for insightful discussions).

Here we used the mechanistic GPM modeling framework described in chapter 4 to investigate the proximate and ultimate consequences of dosage balance alterations in oscillatory GRNs. As discussed in chapter 4, in the context of this modeling framework GRNs possess an explicitly defined genome representation (see chapter 4). This allows us to adequately assess the impact of gene copy number variation on GRN dynamics by altering its genomic structure (genotype), and correspondingly its regulatory wiring, which is assembled from individual protein-DNA interactions. Most importantly, in view of the fact that this GPM model is built upon fine-grained aspects of transcriptional regulation, one can then unambiguously assess the impact of, for instance, competitive DNA binding between non-divergent transcription factor duplicates on the expression dynamics of GRNs. Although, the role of this competitive DNA binding mode in the onset of complex GRN dynamics, such as oscillatory expression phenotypes, has been questioned on the basis of coarse-grained modeling (*i.e.* Hill functions) of transcriptional regulation (see[434]), a recent modeling study that relies on a more principled approach (*i.e.* thermodynamic model) demonstrates that such a DNA binding mode is the key mechanism underlying the oscillatory behavior of a duplicated auto-regulatory motif[435]. Intriguingly, a link has been previously noted in the literature between the presence of paralogous genes and oscillatory processes[436,437], and accumulating evidence seems to suggest an important link between the maintenance of oscillations in regulatory systems and the presence of duplicate genes[424,438].

Based on our fine-grained GPM modeling framework we assessed the immediate fitness impact of single gene duplication and deletion in the ensemble of start GRNs previously used to conduct evolutionary simulations (see chapter 5). Furthermore, we conducted evolutionary simulations toward new phenotypic optima with GRNs carrying an extra copy for either an activator, a repressor or an output gene (imbalanced GRNs), and compared their capacity to attain high fitness levels (our operational definition of evolvability) with that of haploid and diploid GRN system configurations. Lastly, we examined the impact of single gene duplication, deletion and amplification of gene copies on the oscillatory expression phenotype of high fitness-scoring (haploid and diploid) GRNs previously evolved.

6.2 Results

6.2.1 Sensitivity to dosage balance alteration can be modulated through quantitative changes of a fixed network topology

All the (ancestral) GRN configurations previously considered as starting points for evolutionary simulations (see chapter 5) were interrogated for sensitivity to dosage balance alterations, by duplicating (in the haploid GRNs) and deleting (in the diploid GRN system configurations) the transcriptional activator and repressor encoding genes. The results revealed a wide range of heterogeneous phenotypic responses to dosage balance alterations (see Figure 6.1(A)), which is surprising given that the wiring of the GRNs considered is topologically indistinguishable (Smolen-like topologies, see chapter 5, figure 5.1). Overall, one can observe that dosage balance alterations in the model GRNs can have variable impacts on the amplitude, frequency as well as the phase of the oscillatory expression phenotype. This observation demonstrates that in our artificial transcriptional regulatory systems dosage sensitivity can be finely modulated through quantitative differences given a fixed regulatory wiring. Essentially, the distinct phenotypic responses observed to dosage balance alterations achieved through single duplication of activator/repressor genes arise from a complex interplay between sequence encoded quantitative features (protein-DNA binding affinities), which parameterize the strength of the regulatory linkages within a GRN, and additional network control parameters, such as basal kinetic rates (*i.e.* mRNA and protein half-lives). Among the cases analyzed we found that in $5/50$ and $11/50$ of the start haploid GRN configurations the stable oscillatory expression phenotype was not disrupted upon duplication of the repressor or the activator gene, respectively (see Figures C.24 and C.25), whereas the phenotypic impact of removal of one of the copies of the repressor encoding gene in the start diploid GRNs was consistently buffered in $39/50$ of the cases, while deletion of the activator encoding gene was found to disrupt the sustained oscillatory phenotype in $49/50$ of the cases.

It should be noticed that in all of the cases where a haploid GRN configuration is able to oscillate upon duplication of the activator gene, the corresponding diploid system configuration is also able to oscillate upon deletion of a repressor gene copy, with only minor differences existing between the two oscillatory expression profiles in terms of the period see Figures C.24 and C.25). Note that such relationship does not imply that in our model GRNs the phenotypic impact of dosage balance changes are always equal for duplication of the repressor gene in a haploid GRN and deletion of the activator gene in the corresponding diploid system configuration, due to the presence of a concentration dependent factor intended to account for differences in cell volume between haploid and diploid cases (see detailed description of this models feature in chapter 4, subsection 4.1.4). In fact, we observed many cases (28/50) where deletion of the repressor gene in a diploid GRN system configuration does not disrupt the stable oscillatory expression phenotype, while duplication of the activator in the corresponding haploid GRN elicited an abrupt

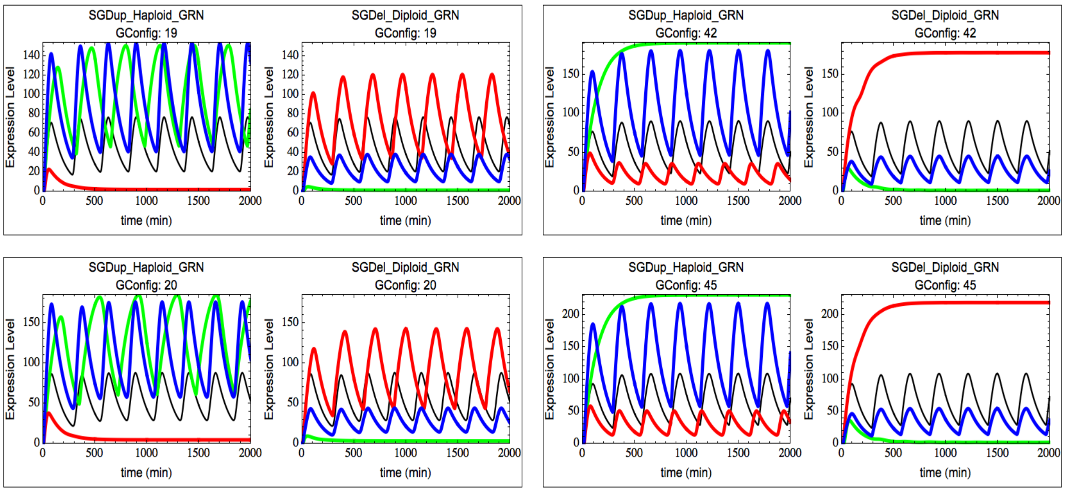


Figure 6.1: Immediate impact of dosage balance alteration on expression dynamics. *A, impact of duplications/deletions of activator (green-color coded profile), repressor (red color-coded profile) and downstream output (blue color-coded profile) genes on the expression dynamics of haploid and diploid GRN system configurations, with respect to the expression phenotype of the unperturbed system (black color-coded profile). The phenotypic readout of a GRN is taken as the time varying concentration of the protein encoded by the downstream output gene.*

phenotypic transition between the stable oscillatory behavior an a steady state expression profile. Furthermore, it should be noticed that duplication/deletion of the downstream output gene consistently alters the amplitude of the oscillatory expression pattern in haploid (over-expressed amplitude) and diploid (under-expressed amplitude) GRN system configurations. We also interrogated another ensemble of oscillatory GRNs for dosage balance effects (see randomly generated configurations in figures C.26 and C.27). Interestingly, we found in this additional ensemble of model GRNs that in $24/50$ and $43/50$ of the start haploid GRN configurations the stable oscillatory expression phenotype was not disrupted upon duplication of the repressor and the activator gene, respectively. Further, we observed that the phenotypic impact of removal of one of the copies of the repressor-encoding gene in the diploid GRN system configurations was consistently buffered in 43/50 of the cases, while deletion of the activator-encoding gene was found to disrupt the sustained oscillatory phenotype in 48/50 of the cases. Intriguingly, 17/50 of the haploid GRN configurations were found to oscillate upon duplication of either the activator or the repressor-encoding gene.

To gain insight into the quantitative design principles underlying the dosage sensitive nature of the model GRNs, we performed a clustering analysis of the 100 model GRNs discussed above (50 start GRN configurations used to simulate evolution + 50 additional configurations). The clustergram (see Figure 6.2, A) was generated based on vectors where the entries are given by the $K^{Assoc}_{j,x(i;n,m)}$ values associated to every regulatory linkage in a GRN (*i.e.* TF P_j

binding to a permissible DNA motif on the promoter region of a target gene i), as well as the activator-mRNA half-life, the activator-protein half-life, the repressor-mRNA half-life and the repressor -protein half-life parameter values. Our analysis shows that the different classes of wirings (*i.e.* robust both to activator and repressor SGD, robust only to activator SGD, robust only to repressor SGD, and non-robust ones) tend to be scrambled all over the clustergram, with only few exceptions where some of the configurations corresponding to the same class of wiring were allocated in a similar cluster. In other words, our analysis indicates that it is generally quite difficult to clearly pinpoint the quantitative basis of the dosage sensitive nature of the model GRNs. Nevertheless, we found an interesting association between the half-life of the activator-encoding mRNA and the different classes of wirings considered (see Figure 6.2, B). Specifically, we observed that the half-life of the activator-encoding mRNAs for some of the oscillatory configurations found to be robust both to duplication of the activator and repressor genes is significantly longer than some of the configurations found to be robust only to duplication of either the activator or the repressor, separately. Together, these observations indicate that the dosage sensitive nature of the topologically indistinguishable GRNs analyzed possess an intricate mechanistic basis, which can be modulated through relatively small quantitative changes in regulatory linkages among genes and in basal kinetic parameters. In the context of our artificial transcriptional regulatory systems, this finding implies that the dosage sensitive nature of GRNs can be readily evolved or engineered through *cis/trans* regulatory changes.

6.2.2 Disruption of dosage balance can reveal novel expression phenotypes with adaptive potential

Under certain circumstances, disruption of the dosage balance in cellular information processing systems may prove advantageous, for instance, by uncovering new phenotypic variants with potential adaptive benefits at the cellular level[439 440]. As shown above, dosage balance effects are pervasive among our model GRNs, which raises the question of: what would be the immediate fitness benefits of such dosage balance changes? To answer this question, we examined the adaptive benefit, using the multi-objective fitness function FF1 (see detailed explanation in subsection 4.1.6.2) obtained upon duplication /deletion of the activator or the repressor encoding gene with respect to the LF-type and HF-type phenotypic optima, as well as the ancestral phenotype (IF-type), in the start GRN configurations ((see chapter 5, figure 5.1)). The analysis shown in figure 6.1(B) reveals that duplication of the activator or the repressor gene in the haploid GRN configurations is mostly deleterious, being less than 30% the fitness of the unperturbed GRN in the majority of the cases analyzed ($41/50$ in the LF-type, $46/50$ in the IF-type and $49/50$ in the HF-type for SGD of the activator gene; and $49/50$ in all the target phenotypes for duplication of the repressor gene). In particular, note that the cases interrogated for dosage balance effects under the IF-type expression pattern (the original phenotype in the start GRNs) demonstrate that our model GRNs are generally quite sensitive to SGD events. Analogously, we found that deletion of a copy of the repressor gene in diploid GRN system configurations was found to be considerably less advantageous than the system bearing the whole set of copies intact, whereas deletion of a

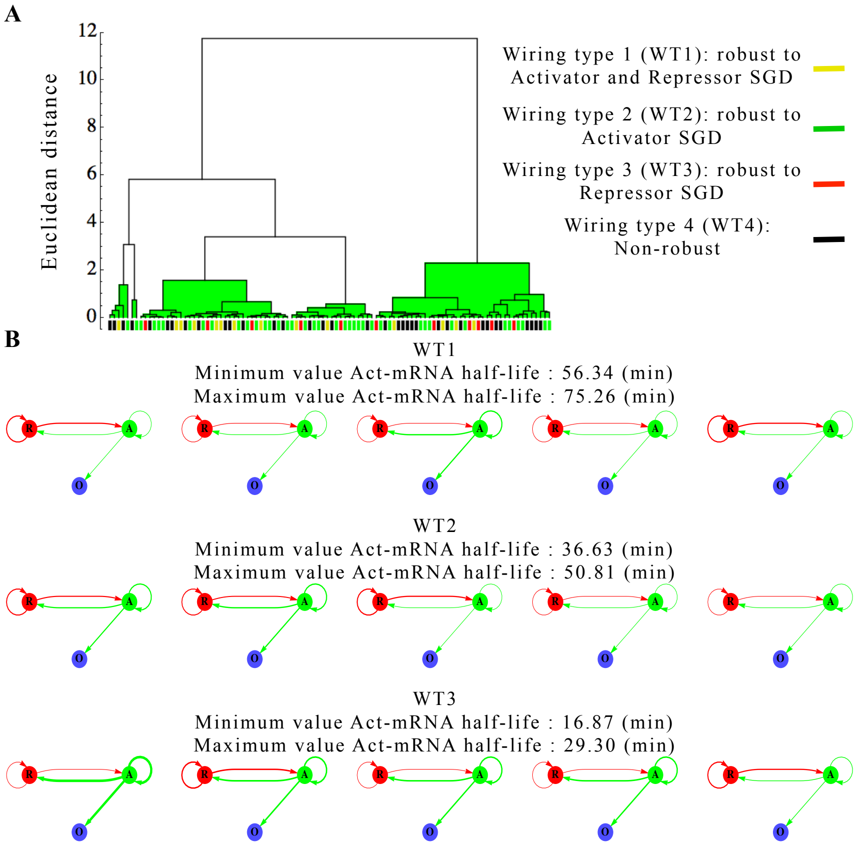


Figure 6.2: Association between quantitative features of GRNs and dosage balance effects. *A, Clustergram of the 100 GRNs considered (50 GRNs used as starting points for evolutionary simulations + 50 additional GRNs). The clustergram is generated based on vectors containing the $K^{Assoc}_{j,x(i;n,m)}$ parameter value associated to every regulatory linkage in a GRN, as well as the Activator-mRNA half-life, the Activator-protein half-life, the Repressor-mRNA half-life and the Repressor-protein half-life. Vectors were normalized by the maximal value of each entry found across the ensemble of 100 configurations. B, Wirings that map to specific sub-clusters in the clustergram shown in A. The most distinctive feature between the different classes of wirings shown (i.e. wirings which are partially dosage compensated for Rep-SGD (bottom row) and Act-SGD (middle row), as well as compensated for Act and Rep SGD events (top row)) is the Activator-mRNA half-life, which is a kinetic parameter critically involved in the specification of oscillatory expression dynamics in regulatory circuits. The thickness of the edges in the regulatory wirings displayed is proportional to the binding strength ($K^{Assoc}_{j,x(i;n,m)}$) of a given TF for a given DNA motif on the promoter region of a target gene.*

copy of the activator encoding gene proved only disadvantageous. Taken together, our results suggests that dosage balance alterations in transcriptional regulatory systems are generally deleterious, although a small chance exist that slightly advantageous novel phenotypes can be triggered in response to duplication or deletion events, depending on the quantitative features of the regulatory system (*i.e.* the strength of regulatory linkages among genes).

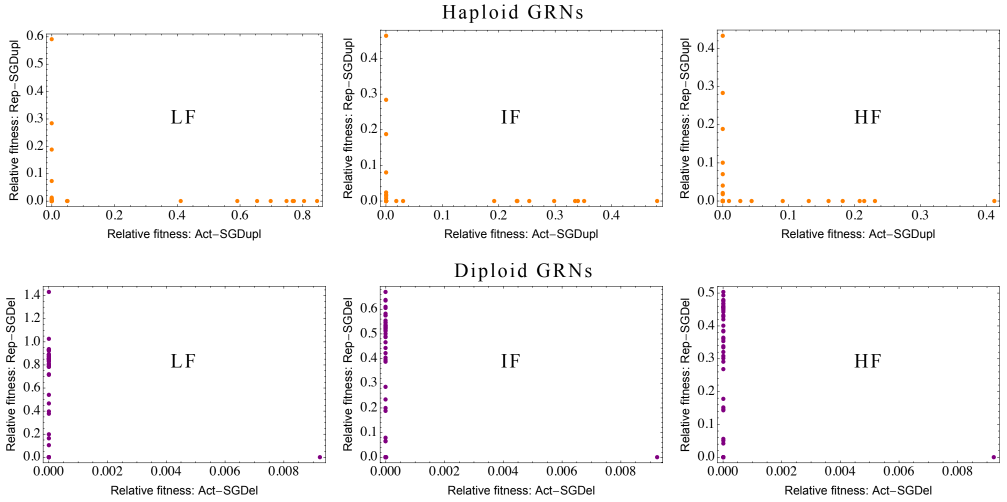


Figure 6.3: Relative fitness impact of dosage balance alteration. *The figure shows the relative fitness benefit conferred by duplication and deletion of the activator and the repressor encoding genes. The fitness of the unperturbed and the perturbed (with duplicated/deleted genes) GRN system configuration is assessed with respect to LF-type and HF-type phenotypic optima, as well as the ancestral phenotype (IF-type), and the relative fitness benefit is taken as the ratio of fitness scores for the perturbed to the unperturbed GRN system configuration. Fitness scores were computed using fitness function FF2, as described in section 4.1.6.2.*

6.2.3 Impact of dosage balance effects on the evolvability of GRNs

Upon a sudden environmental change, the fitness associated to a given genotype may be a major determinant of its evolutionary fate. In the context of an adaptive walk, in particular, where evolution toward a newly imposed optimum on the fitness landscape typically proceeds through the gradual accumulation of beneficial mutations (see [368,441]), the initial fitness value of a founder genotype may usually play an important role in determining the course and the outcome of the adaptation process [368,441]. In the case of a GRN carrying an extra copy of a regulatory gene, this must first get established in a population, which can be achieved by several means [134], for evolution to act upon, and it will do so only if it is not immediately outcompeted by a balanced system configuration (*i.e.* a haploid or diploid GRN). Given that single gene duplication events are often associated with undermined fertility and/or survival, such genotypes will tend to be removed (*i.e.* through strong purifying selection) from the gene pool of populations. Therefore, in this section we focused on the examination of the course and the outcome of the adaptation process of particular imbalanced GRNs with initial fitness ≥ 0.1, and compared their evolvability with that of balanced (haploid/diploid) GRN system configurations. To do this, we conducted evolutionary simulations mimicking the adaptation of GRNs toward the LF-type and HF-type phenotypic optima, using as starting points the ensemble of (ancestral) GRN configurations considered in chapter 5. In this case, imbalanced GRN system configurations carrying an extra copy for one of the genes (activator, repressor, output) were evolved toward the two phenotypic optima considered using the same evolutionary simulation protocol implementing the multi-objective fitness function FF1 (see subsection 4.1.6.2 and Figure 5.2), as used in chapter 5. Figures C.28 - C.32 depict the total ensemble of average fitness trajectories simulated. The panels A-D shown in Figure 6.4 depict the immediate impact of single gene duplication events on the expression dynamics of certain start GRNs, as well as their average fitness trajectories together with those trajectories corresponding to the respective balanced GRN system configurations. Our results demonstrate that under certain conditions a SGD event can prove advantageous in the long term compared to the balanced GRNs. Notably, the configurations analyzed indicate that the adaptation process toward the LF and HF-type phenotypic optima in the GRNs carrying an extra copy of the activator-encoding gene (Figure 6.4 A,B) is significantly speeded up compared with the rate of adaptation in the balanced GRN system configurations. For instance, we found that the fitness values attained at 5% of the entire evolutionary time window simulated tend to be significantly higher for the imbalanced GRNs (one-sided Mann-Whitney test, $p > 0.5$). Further, for these particular configurations we found that the magnitude of the end point fitness values attained by the imbalanced GRNs are fairly similar to those fitness values attained by the diploid GRNs (one-sided Mann-Whitney test, $p > 0.7$), but are significantly higher than those attained by the haploid GRNs (one-sided Mann-Whitney test, $p > 0.2$). On the other hand, we observed that both the course and the outcome of the adaptation process in the GRNs carrying an extra copy of the repressor-encoding gene shown in Figure 6.4 (C,D) are quite distinct: while both the fitness values attained at 5% and 100% (end point fitness) are significantly higher in the imbalanced GRN configuration evolved toward the LF-type optimum (panel C) compared with the balanced GRNs (one-sided Mann-Whitney test, $p > 0.5$), for the GRN configuration evolved toward the HF-type optimum (panel D) we found that statistically significant higher end point fitness values were attained by the imbalanced GRN compared only with the haploid system configuration (one-sided Mann-Whitney test, $p > 0.9$). Together, these results reveal that under particular conditions a single duplication of a regulatory gene may have a positive impact on the evolvability of a GRN.

6.2.4 Modulation of expression dynamics through amplification of regulatory gene copies

Gene copy number changes, such as gene amplification, have long been known as an important source of genetic variation within and among populations[407], and are major drivers of phenotypic variation at the expression level typically associated with detrimental changes[407,408], and less often with adaptation at the cellular and organismal levels[412–414].

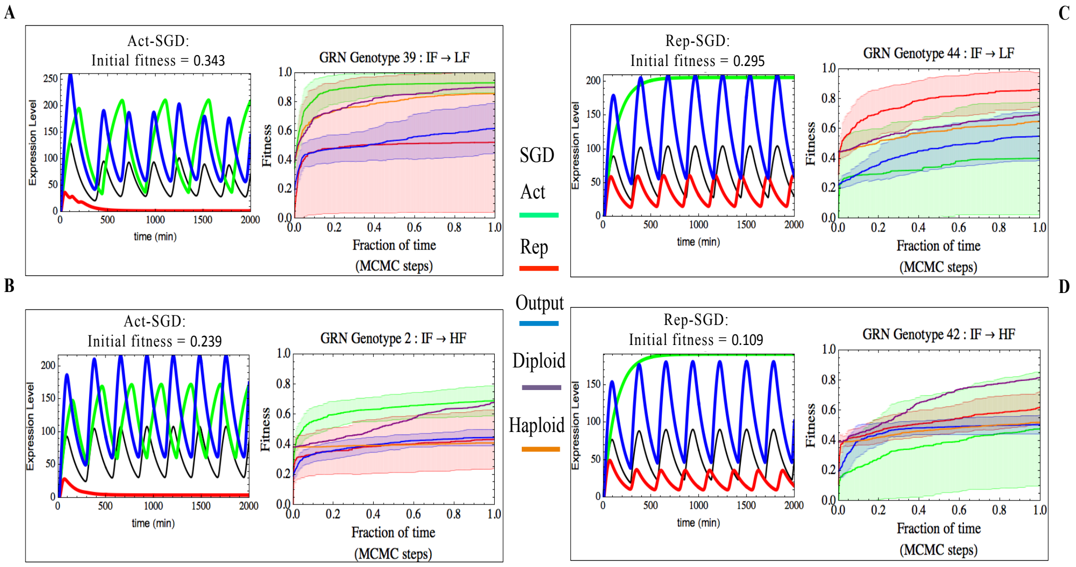


Figure 6.4: Average fitness trajectories of particular GRN configurations evolving under imbalanced vs. balanced conditions. *the expression phenotype of start haploid GRNs upon duplication of the activator (green-color coded expression profile), repressor (red color-coded expression profile) and downstream output (blue color-coded expression profile) gene, as well as the expression phenotype of the unperturbed system (black color-coded expression profile). The phenotypic readouts are taken as the time varying concentration of the protein encoded by the downstream output gene. On top of figures A and B are shown the (initial) fitness score (computed with respect to the LF-type phenotypic optimum) associated to the haploid GRN carrying an extra copy of the activator (A) and repressor (B) gene. C-D, Temporal sequences of fitness values recorded from 50 independent simulation replicates, using the start GRN configurations corresponding to A and B, were averaged out to display the general trend of the adaptation process toward a new phenotypic optimum. The trajectories shown correspond to evolving GRN configurations carrying an extra copy of the activator (green color-coded trajectory), repressor (red color-coded trajectory), and output genes (blue color-coded trajectory), separately. For comparison purposes, average fitness trajectories are also displayed for the corresponding haploid (orange color-coded trajectory) and diploid (purple color-coded trajectory) GRNs. Error bars along the trajectories indicate standard deviations. Fitness scores were computed using the multi-objective fitness function (Fitness-F2) described in section 4.1.6.2.*

In this section we examined the impact of amplification of regulatory gene copies on expression dynamics, and associated fitness, in a large ensemble of model (haploid) GRNs. Specifically, we focused on high fitness-scoring haploid GRNs (856 LF-type solutions in total with fitness scores > 0.9) that have been previously evolved *in silico*, using an implementation of the already described evolutionary simulation protocol with the multi-objective fitness function FF1 (see subsection 4.1.6.2 and Figure 5.2). Figure 6.5 (top panels) illustrates the distribution of relative fitness at different gene copy number variants. These results clearly demonstrate that as the number of gene copy number variants increases the proportion of GRNs with relatively high fitness scores is significantly greater when the gene being amplified is the one encoding for the activator TF, compared to the gene encoding for the repressor TF. The plots depicting the average trend of the relative fitness as a function of gene copy number variants (see Figure 6.5, bottom panels) demonstrate that the relative fitness score tends, on average, to decay exponentially as a function of increasing number of activator-encoding gene copies, suggesting that copy number variation in the activator-encoding gene can induce a broad range of phenotypic variation in oscillatory expression dynamics. In contrast, copy number variation in the repressor-encoding gene causes an abrupt decline in the relative fitness across the ensemble of evolved GRNs being analyzed. This finding seems to suggest that in cascade-like GRNs driven by negative feedback mechanisms, gradual modulation of oscillatory expression phenotypes can be better achieved under copy number variation in the activator-encoding gene, whereas amplification of the repressor-encoding gene would typically result in phenotypic transitions involving qualitative changes in expression dynamics (e.g. transition between oscillations and steady state expression profiles).

Furthermore, we have found particular cases where the relative fitness changes in a non-monotonic fashion as a function of copy number variation in the activator-encoding gene. For instance, figure 6.6 displays two different cases where fitness abruptly declines and then recovers as the number of activator-encoding gene copies increases. Examination of the phenotypic impact of copy number variation confirms that the original stable oscillatory expression pattern is considerably altered and then restored at different number of activator-encoding gene copies. In contrast, stable oscillations are systematically disrupted for copy number variants > 1 in the repressor-encoding gene. These observations demonstrate that in transcriptional regulatory feedback circuits subject to substantial changes in the dosage of one of the components, non-monotonic phenotypic responses can arise as a consequence of complex non-linear regulatory interactions. Based on these results it is tempting to speculate on the idea that even for negative feedback-driven molecular networks operating through a combination of transcriptional and post-transcriptional mechanisms, such as the mammalian circadian clock[424] and the yeast cell cycle network[426], which are subject to strong stabilizing selection for oscillatory dynamics, variation in the number of gene copies might be restricted to molecular components with activating regulatory roles.

6.2.5 Dosage balance effects in duplicated GRN system configurations with divergent gene copies

In addition to gene duplication, gene deletion represents an important source of dosage balance effects, the consequences of which are usually linked to detrimental effects[160,408,422]. Using our GPM modeling framework we examined the impact of gene duplication/deletion on the expression phenotype of GRNs carrying duplicate gene pairs. One of our main interests is to gain mechanistic insight into the resolution of dosage balance effects in dupli-

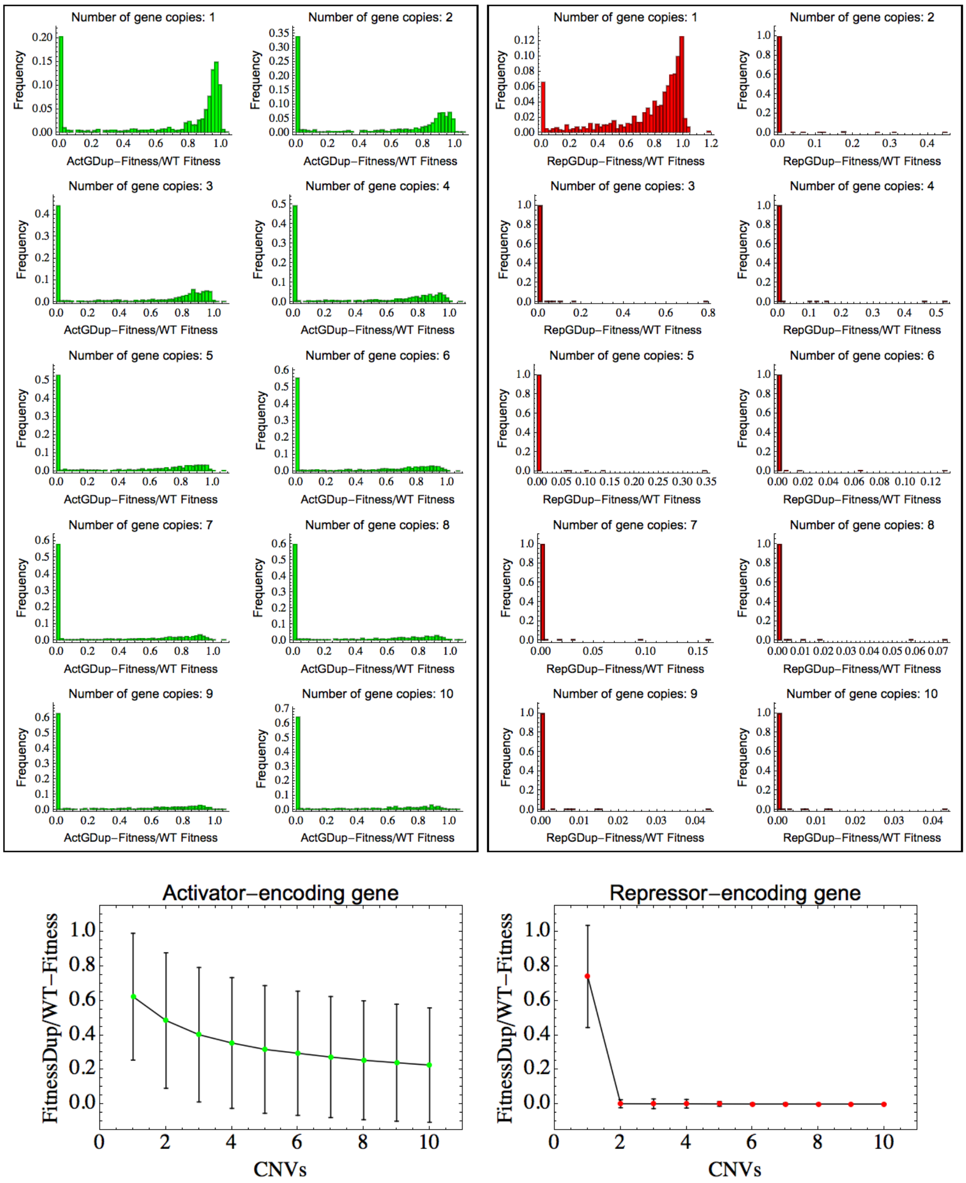


Figure 6.5: Fitness impact of copy number variation. *Distributions of fitness effects associated to activator and repressor gene copy number variants. The top panels illustrate the distribution of relative fitness scores calculated at different copy number variants for activator (green-color coded distributions) and repressor (red-color coded distributions) encoding genes, in a large ensemble (856) of high fitness-scoring LF-type GRNs previously evolved in silico. The bottom panels depict how the average trend of the relative fitness decays at different gene copy number variants (CNVs) for the activator and repressor TFs*

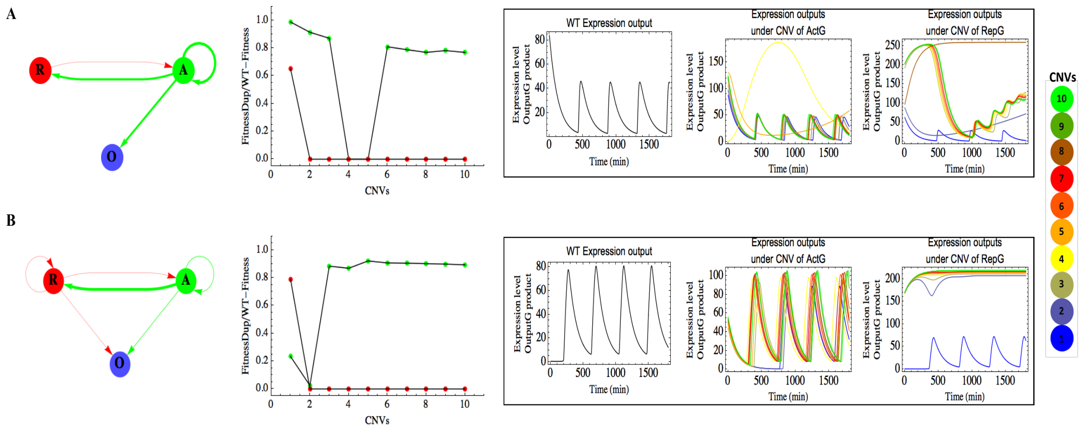


Figure 6.6: Non monotonic changes in fitness in response to copy number variants (CNVs). *The cases illustrated represent GRNs where changes in fitness in response to CNVs in activator and repressor-encoding genes exhibit a non-monotonic trend.*

cated GRN system configurations over the course of evolution (*e.g.* during the genome fractionation process typically observed subsequent to a WGD event). Similar to our previous analyses, we interrogated for dosage balance effects a large ensemble (900 LF-type solutions in total with fitness scores > 0.9) of diploid GRNs evolved *in silico*, using an implementation of the previously described evolutionary simulation protocol with the multi-objective fitness function FF1 (see subsection 4.1.6.2 and Figure 5.2). Importantly, analysis of the impact of gene duplication/deletion on the fitness of duplicated GRN system configurations requires careful consideration of several details. Critically, this type of evolved GRNs are usually characterized by divergent duplicate gene pairs, in the sense that they have acquired distinct functional roles within the network (*e.g.* in terms of either connectivity, interaction strength, or both) over the course of evolution (see Figure 6.7). In fact, we have noticed that all GRNs exhibit duplicate gene pairs that have experienced different sorts of functional alterations involving *cis*/*trans* regulatory changes. Moreover, since the connectivity of GRNs carrying duplicate gene pairs is substantially more complex than that of GRNs carrying single copy genes (see Figure 6.7), a vast number of distinct regulatory wirings are usually recovered through *in silico* evolution, several of which have lost one of the gene copies codifying for either the activator or the repressor, or less frequently for both regulators. These observations demonstrate that in a small fraction of the GRN configurations some of the dosage balance constraints have been resolved over the course of evolution. These observations underscore the critical role of network rewiring, achieved through sequence divergence of *cis* and *trans*-regulatory regions of duplicate gene pairs, in the resolution of dosage balance constraints.

In order to adequately assess the impact on fitness of dosage balance effects in the duplicated GRNs evolved, we excluded system configurations whose wirings contained nodes that were entirely disconnected from the rest of the system, or effector nodes (*e.g.* transcriptional activators and repressors) that did not feedback onto the system (*e.g.* nodes that were regulated by other nodes but did not regulate other nodes in the GRN, approximately 5% of the solutions evolved). Keeping in mind these considerations, we examined the fitness impact of duplication and deletion of every activator and repressor-encoding gene, separately, in a set of GRNs carrying divergent duplicate gene pairs (Figure 6.8). Our results show the existence of several interesting patterns. For instance, we found that the oscillatory expression dynamics in a great proportion of the GRNs analyzed can be buffered (relative fitness scores ≥ 0.9) under the presence of an additional copy for the activator ($\sim 42\%$ of the cases), the repressor ($\sim 27\%$ of the cases), or both genes ($\sim 11\%$ of the cases), separately. As can be noticed, the fitness impact tends to be less severe when one of the activator gene copies is duplicated, compared to duplication of the repressor-encoding gene copies. Given that the activator and repressor gene copies in the GRNs analyzed have typically acquired distinct functional roles within the GRN (more often in terms of both connectivity and strength of regulatory linkages with other genes), it is generally quite difficult to conclude whether the buffering capacity of the GRNs is attributable to an active dosage compensation mechanism (*e.g.* interlocked feedback loops or feedforward sub-circuits, see[424]), or it is simply the result of dosage balance constraints being resolved through network rewiring over the course of evolution, which may render a GRN insensitive to an increased dosage in one of the gene duplicates. By contrast, it is noticeable that deletion of regulatory gene copies is typically associated with adverse changes in fitness ($\sim 70\%$ and $\sim 82\%$ of the cases involving deletion of gene copies of the activator and repressor, respectively), which is generally indicative of disruption of oscillatory expression dynamics (see Figure 6.9). Interestingly, these observations are congruent with several reported cases showing that gene deletions tend to be more deleterious than duplications[407,408,411]. Regarding the fitness impact of deletion of regulatory genes, our results seem to suggest that dosage balance constraints in GRN system configurations carrying duplicate gene pairs may require long evolutionary time periods to be resolved.

Furthermore, our results are particularly interesting in the light of classical dosage balance-related phenomena such as haploinsufficiency[160] and gene essentiality[442]. Our results, nevertheless, must be interpreted with causation owing to the presence of divergent regulatory features between the members of duplicate gene pairs. This is a critical point that deserves special attention given that haploinsufficiency studies assume that both alleles perform equivalent biochemical functions[160,422]. Under such condition, any signature of haploinsufficiency can thus be unambiguously attributed to the sensitivity of the system to an effective reduction in gene dosage. By contrast, the divergent regulatory features existing between the members of duplicate gene pairs in the GRNs analyzed here introduce an important source of variation that can be difficult to interpret. For instance, in a network context sufficient reasons exist to believe that divergent duplicate gene pairs do not make equivalent contributions to the overall functional performance (fitness) of a GRN. Indeed, our analysis demonstrates that differences of several orders of magnitude can exist when comparing the fitness impact of duplication/deletion of the members of a duplicate gene pair (see Figure 6.10). Taken together, our results seem to suggest that functional differences between the members of duplicate gene pairs acquired over the course of evolution represent important sources of non-linearity which may render GRNs more (or less) sensitive to gene losses.

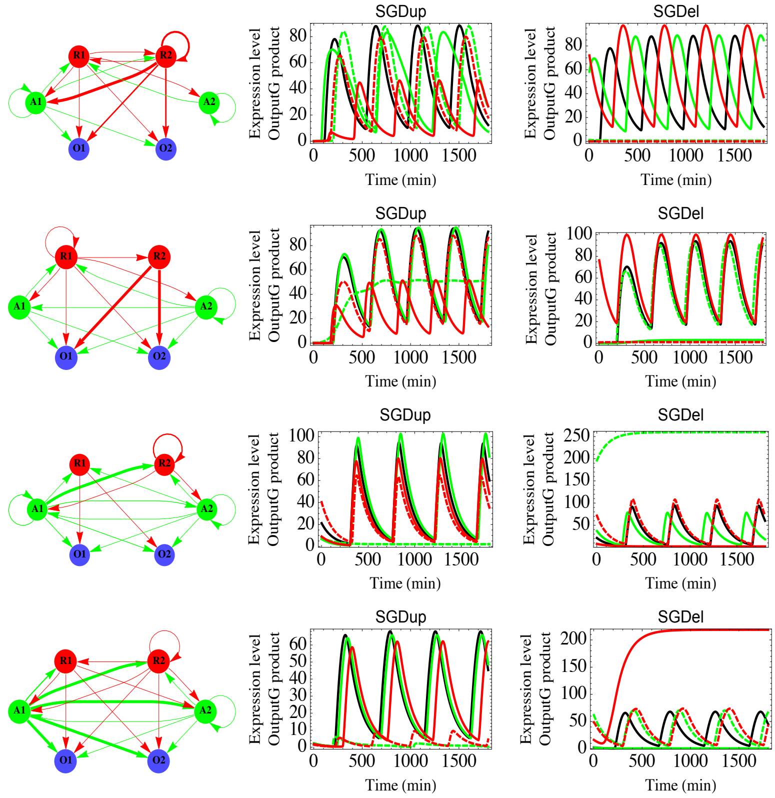


Figure 6.7: GRNs with divergent duplicate gene pairs. *The GRNs shown represent typical cases designed through in silico evolution involving thousands of cycles of mutation and selection for LF-type oscillatory expression phenotypes. Note that these GRNs possess highly divergent (e.g. in terms of connectivity and regulatory strength) duplicate gene pairs that have acquired distinct functional roles within the GRNs (the thickness of the edges in the regulatory wirings displayed is proportional to the aggregated DNA binding strength of a given TF over all possible binding sites on the promoter regions of the target genes). Duplicate gene pairs can diverge over the course of evolution through the gradual accumulation of nucleotide changes in cis and/or trans-acting sequences, by which duplicates subfunctionalize or neofunctionalize. Owing to such functional divergence, notable differences in expression dynamics between duplicates are often observed in response to duplication/deletion of each copy individually. The phenotypic responses are taken as the sum of the expression profiles of the downstream output genes. Red color-coded time courses represent the phenotypic responses of the GRNs under duplication/deletion of repressor-encoding gene copies (dashed line is for repressor-encoding gene copy 2, R2). Green color-coded time courses represent the phenotypic responses of the GRNs under duplication/deletion of activator-encoding gene copies (dashed line is for activator-encoding gene copy 2, A2). Black color-coded time courses represent the expression profile of the unperturbed GRN system configuration.*

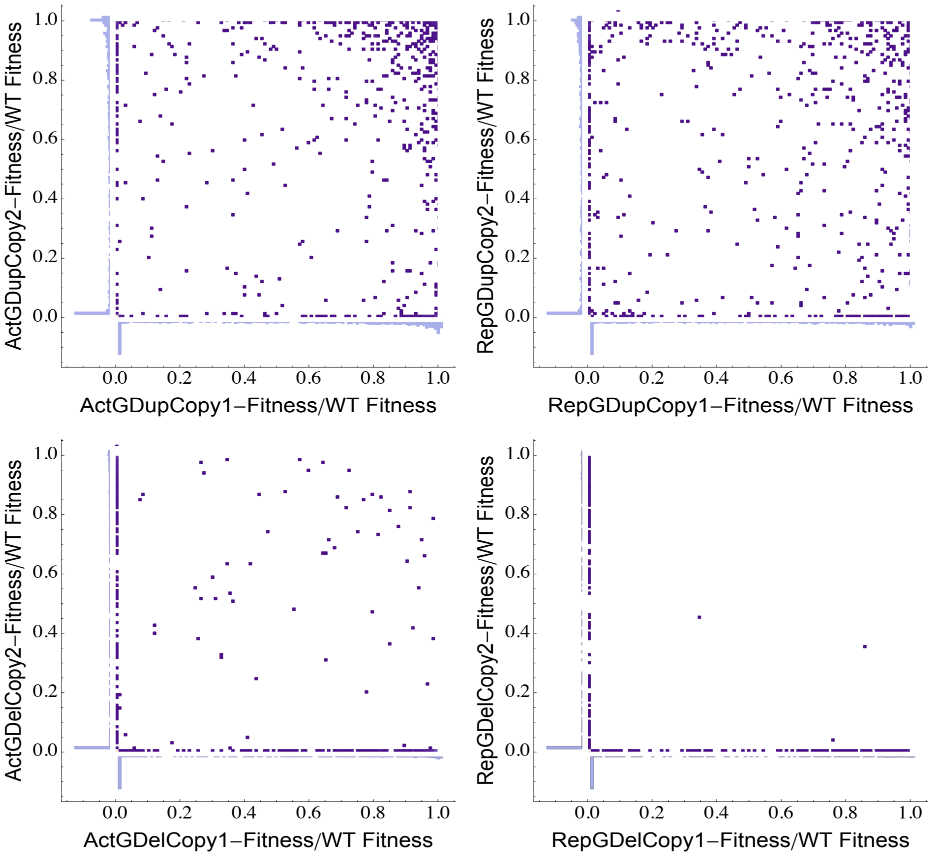


Figure 6.8: Relative fitness impact of dosage balance alterations in GRNs with divergent duplicate regulatory gene pairs. *Density histogram for the bivariate distribution of relative fitness scores associated to every GRN interrogated for dosage balance effects involving single duplication (top panels) and deletion (bottom panels) of each gene copy encoding for the activator or the repressor encoding gene.*

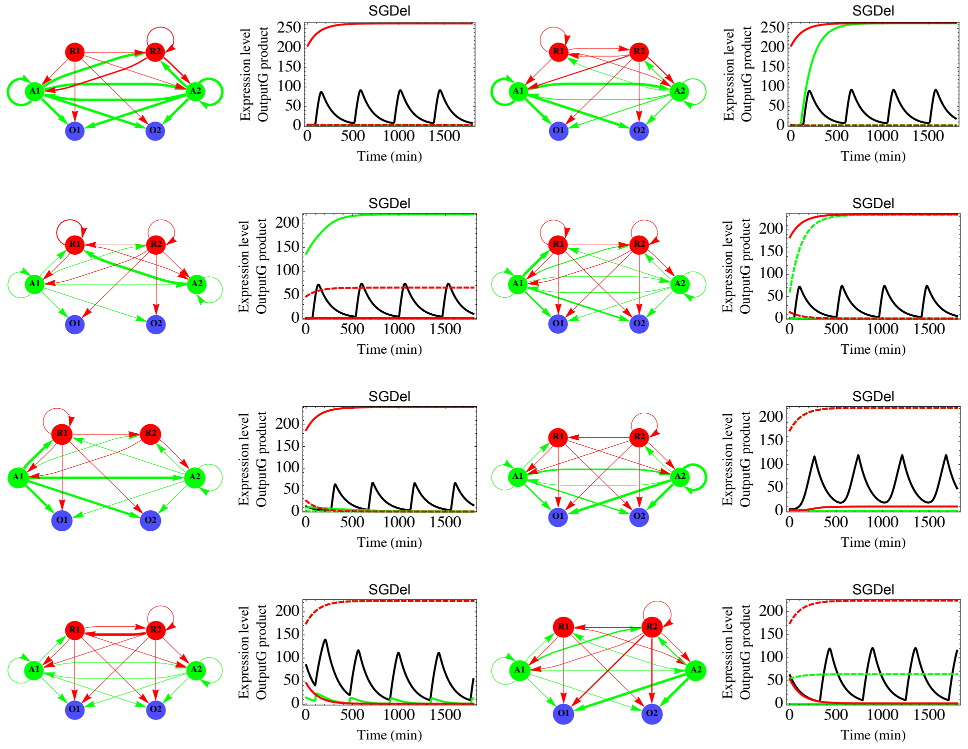


Figure 6.9: Impact of gene deletion on the oscillatory expression dynamics of GRNs with divergent duplicate gene pairs. *The panels illustrate different examples of evolved GRNs with divergent duplicate gene pairs whose oscillatory behavior is severely compromised (i.e. the amplitude) or fully disrupted upon deletion of any activator or any repressor-encoding gene copy. Red color-coded time courses represent the phenotypic responses of the GRNs under duplication/deletion of repressor-encoding gene copies (dashed line is for repressor-encoding gene copy 2, R2). Green color-coded time courses represent the phenotypic responses of the GRNs under duplication/deletion of activator-encoding gene copies (dashed line is for activator-encoding gene copy 2, A2). Black color-coded time courses represent the expression profile of the unperturbed GRN system configuration.*

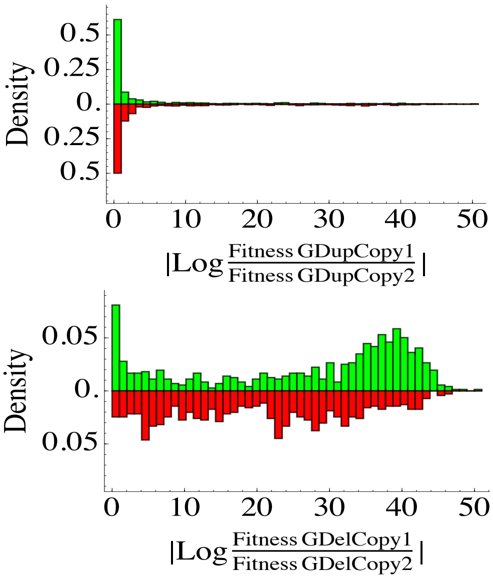


Figure 6.10: Asymmetry in the fitness contribution between duplicate regulatory gene pairs. *Differences in fitness between members of a duplicate gene pair when tested for duplication (top panel) or deletion (bottom panel) of the activator (green color-coded distributions) or the repressor (red color-coded distributions) encoding genes.*

6.3 Discussion

Disruption of dosage balance in transcriptional regulatory systems, cell signaling pathways and macromolecular complexes has often been unequivocally associated with adverse fitness effects [156,157,207,443]. This observation provides support to one of the core explanatory principles of the GDBH, which attributes a key role to dosage balance effects in the systematic pattern of differential gene retention (loss) that has been documented in several species whose genomes exhibit signatures of ancient polyploidization events (paleopolyploids) [137,179,444]. Here, based on a fine-grained, mechanistic GPM modeling framework, we showed that dosage balance alteration, achieved through gene copy number variation including single gene duplication/deletion and gene amplification, represent an important source of phenotypic variation at the level of dynamic expression phenotypes in cascade-like GRN models. Although dosage balance effects are predominantly detrimental in the context of our artificial systems, we found that under particular conditions they prove advantageous by bringing forth novel expression phenotypes with adaptive potential, which may be particularly important in terms of a system's evolvability.

Intriguingly, we found that under a fixed start network topology the degree of sensitivity to dosage balance alterations can be readily modulated by changing the strength of regulatory linkages among the genes (i.e. protein-DNA binding affinities), as well as through variation in network control parameters such as basal kinetic rates (i.e. mRNA half-lives). In particular, we found that dosage balance alterations achieved through duplication/deletion of the activator or repressor genes can have variable impacts on the amplitude, frequency as well as the phase of the oscillatory expression phenotype of start haploid and diploid GRN configurations. However, we observed that in most of the cases interrogated for dosage balance effects a single duplication/deletion event triggers a drastic phenotypic transition from a stable oscillatory behavior toward a steady state pattern. Overall, our results underscore the importance of using detailed mechanistic GPM models in order to adequately capture the concentration-dependent nature of transcriptional regulatory systems. In this respect, it is important to note that efforts have been made to develop increasingly more realistic models of transcriptional regulation to shed light on the quantitative nature of dosage balance effects in regulatory systems[160,445]. However, to the best of our knowledge, the predictions made so far are only applicable to a single-gene system operating under equilibrium conditions[160,445], thus avoiding generalizations to more complex regulatory systems involving feedback control and time varying concentration levels of the molecular components. Our results are worth considering in the light of one of the core principle of the gene dosage balance hypothesis (GDBH), which considers the connectivity of a gene within a network as a major predictor of the phenotypic impact of a gene's under- or over-expression[158]. Our study does not disprove this view but rather extends it by demonstrating that subtle quantitative differences in the strength of the regulatory linkages can play a critical role in determining the impact of dosage balance alterations on the expression dynamics of GRNs. In the light of these results, it is tempting to speculate that the preferential retention and loss of gene duplicates in molecular networks may not be entirely determined by the network topology, as commonly inferred through purely topological models[395], but it may also be severely constrained by quantitative differences in the regulatory linkages within GRNs operating under a fixed topology.

Beyond question, dosage balance effects-related phenomena are particularly critical factors underlying the evolution of molecular networks subsequent to gen(om)e duplication events. Surprisingly, these issues have remained largely understudied so far (but see[382,433] for insightful discussions). In order to shed light into the evolutionary consequences of dosage balance effects, we simulated the evolution of imbalanced GRN system configurations carrying an extra gene copy, and found that the immediate impact of dosage balance alterations on the oscillatory expression phenotype, and associated fitness, may be a major determinant of the course and the outcome of the adaptation process toward a new phenotypic optimum. Interestingly, we observed that for particular start GRNs able to prevent stable oscillatory expression dynamics from being disrupted upon single duplication of the activator or repressor gene, significantly higher fitness values are usually attained at different time points (i.e. fitness values sampled at 5% and 100% the total evolutionary time window considered) over the course of the adaptation process, compared to haploid and diploid counterparts (balanced configurations). These results show that under specific conditions single gene duplication events may improve the evolvability of GRNs compared to balanced system configurations.

Another critical aspect of dosage balance-related phenomena concerns the possible impact of amplification of gene copies, on the dynamical behavior of GRNs. Our analyses demonstrate that in negative feedback-driven GRNs gradual modulation of oscillatory expression dynamics can be effectively achieved through amplification of the activator-encoding gene, whereas amplification of the repressor-encoding gene seems to be unequivocally associated with phenotypic transitions involving qualitative changes in expression dynamics (*e.g.* transition from oscillatory to steady state dynamics). An intriguing finding is that under certain conditions the expression dynamics of GRNs can alternate between oscillatory and steady state regimes at increasing number of activator-encoding gene copies, which is the result of the non-linear nature of transcriptional regulation. It is interesting to note that some of our predictions seem to be in line with previous observations in synthetic biology studies. In particular, our results seem to recapitulate the dosage dependent functioning of a minimal regulatory circuit composed of coupled activator and repressor modules implemented in a bacterial system shown to exhibit a range of dynamical behaviors, including oscillatory expression dynamics[429]. An important observation in this study was that increases in the activator module copy number caused the system's expression dynamics to shift from a region in phenotype space characterized by damped oscillations to another region defined by prolonged oscillations[429]. Likewise, a recent study demonstrated that the oscillatory expression dynamics of a tunable synthetic mammalian oscillator could be reproducible as long as specific relative levels of the molecular species involved were kept in balance[430]. The relative levels of the molecular components of the synthetic circuit are largely determined by the relative amounts of expression vectors used to transfect mammalian cells. Interestingly, a computational model of the circuit predicted that, even when the ratios of the molecular species were held constant, absolute plasmid concentrations could be used to modify the period and amplitude of the oscillations[430]. Overall, these observations seem to be in accordance with our results showing that increasing changes in the dosage of the molecular components of feedback-driven transcriptional regulatory cascades can be an effective way to modulate expression dynamics over a broad range.

In addition, our results seem to be consistent, to some extent, with previous simulation studies showing that copy number variation in distinct regulatory network motifs, including oscillatory circuits, can lead to multiple orders of magnitude change in gene expression as well as qualitative changes in circuit dynamics[423,446]. We find, nevertheless, important differences both in the granularity of the models used and in the way copy number variation is simulated, which are worth discussing. Firstly, it should be noted that we simulate copy number variation by increasing the number of regulatory gene copies contained in the genome representation of GRNs, whereas Mileyko *et.al.*[423,446] simulated copy number variation by increasing the number of network motifs as a whole. Moreover, these simulation studies rely on standard coarse-grained approximations to transcriptional regulation in which crucial mechanistic details are presumably encapsulated in aggregated parameters. In addition, the use of coarse-grained mathematical representations of GRNs necessarily implies an arbitrary treatment of genetic perturbations. For instance, it is standard procedure to simulate gene copy number variation by adjusting certain parameters of a fixed model structure in a way reflecting proportional changes in gene dosage (*e.g.* 2-fold changes for a single gene duplication). In contrast, in the context of our GPM modeling approach crucial fine-grained details of gene regulation can be explicitly accounted for, including for instance, the case of a transcriptional regulator binding multiple DNA motifs with distinct affinities, transcriptional activators and repressors engaged in competitive binding to partially or fully overlapped DNA motifs, as well as competitive DNA binding occurring between non-divergent duplicate transcription factors. Importantly, in our modeling framework all these fine-grained details of transcriptional regulation are captured in sequence-encoded rules contained in the minimal genome representation of GRNs. Using such genome representation as the basis for conducting genetic modifications, one can then simulate a gene duplication (deletion) event by adding to (deleting from) a target gene, which causes model structure to change accordingly (*e.g.* by changing the number of parameters and equations accounting for the expression dynamics of each molecular species individually). Obviously, the highlighted differences between modeling approaches are expected to result in fundamentally different predictions regarding the impact of dosage balance effects on network dynamics.

We also examined the dosage dependent nature of GRNs carrying duplicate gene pairs. One of the emerging patterns observed across our simulations is that the dosage dependent nature of a GRNs can be determined by the degree of functional divergence between the members of a duplicate gene pair. In particular, we noticed that the fitness impact of a deletion/duplication could vary by several orders of magnitude between duplicate gene pairs. These observations were consistent across the ensemble of GRNs analyzed. A possible explanation for this pattern relates to the fact that duplicate genes may have acquired over the course of evolution substantially different roles within the GRN. Duplicate genes can subfunctionalize or neofunctionalize through the gradual accumulation of changes in *cis*-regulatory regions and/or *trans*-acting elements, which can derive in duplicate genes acquiring distinct roles in terms of the number, type and strength of regulatory linkages within a GRN[86,383,447]. Overall, our observations show that if duplicate gene pairs are differentially wired within a GRN, the fitness impact of gene deletion/duplication can be expected to vary, even over several orders of magnitude, between the members of a duplicate gene pair. It should be noticed that the GRNs analyzed have been the product of an evolutionary optimization process implementing a directional selection regime (*i.e.* selection for a new phenotypic optimum). Therefore, in order to gain a more general idea on the resolution of dosage balance constraints in duplicated GRN system configurations, it will be necessary to conduct evolutionary simulations under a stabilizing selection regime, and assess whether the resolution of dosage balance constraints is achieved differently in GRNs evolved toward a new phenotypic optimum (*i.e.* adaptive rewiring) compared to those evolved across neutral domains in genotype space (*i.e.* neutral rewiring) where the fitness is kept invariant over time.

In this study we have focused on oscillatory expression dynamics as reference quantitative phenotypes for exploring the dosage sensitive nature of cascade-like GRNs. This has been motivated by previous observations showing the importance of regulatory imbalances in cellular processes characterized by periodic biochemical activities, the disruption of which can cause severe detrimental effects, such as cell cycle arrest and abnormal morphology[448], as well as altered circadian rhythms[449,450]. In this respect, it is difficult to generalize our findings given that the dynamical realization of other biologically relevant phenotypes, such as bistable or pulse-like expression patterns, may require distinct molecular rules beyond those implemented by purely transcriptionally based regulatory systems. However, one might expect some common patterns to emerge at a more coarse-grained description of molecular networks, such as the network topology level[216]. Nevertheless, our work favors the view that increasingly more detailed mechanistic models are necessary to adequately study the functional and evolutionary properties of molecular networks[392]. This may have far reaching implications in several contexts, such as in the rational design of robust synthetic circuits, in the investigation of the etiology of complex diseases (e.g. carcinogenesis), as well as in the study of the evolution of emergent system properties (*e.g.* evolvability).

"Satisfaction of one's curiosity is one of the greatest sources of happiness in life"

Linus Pauling

7

Discussion and Future Perspectives

Author contribution

All content within this chapter was written by myself and revised by professor Steven Maere.

7.1 The importance of mechanistic modeling to study the evolutionary potential of biological systems

Understanding the relationship between the genotype and the phenotype lies at the center of almost any biological problem conceivable. Although many different approaches exist to address genotype-phenotype mapping problems, the ultimate goal is, nevertheless, to try to reveal regularities regarding the way phenotypes vary as a function of changes in the genotype. The era of systems biology brought into the mainstream of molecular and evolutionary biology a great variety of engineering-inspired quantitative frameworks that allow the study of GPM problems at different levels of biological organization from a more mechanistic perspective. In particular, systems biology-inspired network modeling approaches have gained increasing attention as quantitative tools to study the evolutionary potential (evolvability) and origin of emergent system properties such as robustness and modularity. Nevertheless, the level of granularity afforded by these modeling approaches is not sufficient to adequately study all the intricacies of evolving biological systems, the most pressing factor being the understanding of how molecular networks gradually acquire novel phenotypes and properties as they navigate the fitness landscape via discrete changes in the genotype.

As demonstrated in this work, the use of a mechanistic model of transcriptional regulation, in combination with a suitable modeling framework to explicitly account for the genotypic encoding of GRNs, offers the possibility to simulate their functional and evolutionary properties at a level of resolution that is well beyond of those models operating at a more abstract level. In the context of our GPM modeling approach, crucial fine-grained details of gene regulation can be explicitly accounted for, including the case of a transcriptional regulator binding multiple DNA motifs with distinct affinities, transcriptional activators and repressors engaged in competitive binding to partially or fully overlapped DNA motifs, as well as competitive DNA binding occurring between non-divergent duplicate transcription factors. Importantly, in our modeling framework all these fine-grained details are captured in sequence-encoded rules contained in a minimal genome representation of a GRN. Using a GRN's genome representation as the basis for conducting genetic modifications, one can then simulate a gene duplication (deletion) event by adding (deleting) a target gene, which causes model structure to change accordingly (*e.g.* by changing the number of parameters and equations accounting for the expression dynamics of each molecular species individually).

A particularly intriguing evolutionary aspect of GRNs that is within the scope of what can be investigated with the GPM modeling framework presented here is the contribution of neutral evolution in the adaptation of GRNs. In this respect, our simulation results (see chapter 5) demonstrated that when evolving GRNs were allowed to traverse the fitness landscape via neutral substitutions, the accessibility of high fitness scoring solutions was typically substantially improved. This finding was indicative that GRNs were able to reach a newly imposed phenotypic optimum by drifting through neutral domains (neutral network of genotypes) in sequence space associated to different sub-optimal fitness levels. Given the level of resolution at which the adaptation of the model GRNs was simulated (*i.e.* via a mutational walk across sequence space involving single nucleotide substitutions at each step of the algorithm), several interesting questions could be addressed regarding the role of neutrality in adaptation, such as: what is the minimal number of neutral substitutions required to move from one sub-optimal peak to a higher one (shortest mutational path) on the fitness landscape of GRNs. This type of questions are simply outside the scope of what most conventional systems biology-inspired network models can offer, because those coarse-grained models rely on a continuous parameter space to simulate network evolution, where key aspects of fitness landscapes such as mutational neighborhood, connectivity and accessibility cannot be easily interpreted. Similarly, if one considers evolutionary scenarios where distinct classes of dynamic expression phenotypes (*e.g.* oscillatory, bistable, and pulse-like expression patterns) exist, which are particularly well suited (adapted) to specific environments (*e.g.* subject to stabilizing selection), using a model like this one could adequately assess the portion of sequence space (*i.e.* the extent of the neutral network) associated to each optimal expression phenotype. By mapping such functional domains on the fitness landscape of GRNs one could eventually create a portray of sequence space suitable to investigate efficient ways to navigate the fitness landscape and to study, for instance, phenotypic innovation.

Several practical implications can be envisioned for the type of evolutionary scenarios outlined above, which can be adequately explored with the GPM model developed in this work. For instance, in the context of synthetic implementations of regulatory circuits one of the most challenging tasks is the exploration of (largely unknown) extensive regions of the sequence space of the circuits[215,216,451]. Understanding the organization of the fitness landscape (*e.g.* the extent of connectivity between functionally distinct regions) of relatively large synthetic circuits can be crucial for the implementation of novel circuit functionalities (expression patterns) achieved through a minimal number of mutations (see[452] for a case where a non-functional circuit is transformed into a functional one by means of a directed evolution approach), and also to understand how far in sequence space could a circuit travel before its functionality is disrupted (network robustness)[216]. Another interesting network evolution aspect that falls within the scope of what can be explored with the GPM modeling framework discussed, and that may have important implications as well in the design of synthetic circuits, is the relative contribution of neutral vs. adaptive sequence divergence between duplicate gene pairs in the acquisition of novel regulatory roles within the GRN context. By monitoring the gradual accumulation of changes at the *cis*-regulatory regions (gene promoters) and *trans*-acting elements (DNA binding domains) that drive the functional diversification of paralogous transcriptional regulators over the course of evolution, one could elucidate potential network rewiring mechanisms by which the architecture of extant GRNs could have been shaped[86]. Due to rapid advances in our ability to synthesize DNA, which have enabled researchers to construct, for instance, a whole bacterial genome[453] and a partial eukaryotic chromosome[454], it won't take longer before the insights gained from *in silico* evolution experiments, and the hypotheses derived thereof, can be proved and further refined by means of synthetic evolutionary biology[451], directed evolution of regulatory circuits[455], and evolutionary genome engineering[456], which together with microbial experimental evolution hold the promise to radically alter our view on the evolutionary potential of complex cellular information processing networks.

In the present work, we have concentrated on oscillatory expression dynamics as reference quantitative phenotypes for exploring proximate and ultimate consequences of gen(om)e duplications in the model GRNs. Moreover, we have focused on purely transcriptional regulatory systems, while other types of molecular systems, such as signaling and metabolic networks, rely on the implementation of other type of biochemical mechanisms, such as phosphorylation-mediated post-transcriptional regulation and enzyme-catalyzed reactions, to achieve their functional tasks. In this sense, this work can be considered quite narrow in scope, a shortcoming that emerges in virtually any modeling project that involves time-consuming simulation experiments. In evolutionary studies, in particular, one is always restricted to exploring just a small fraction of the constellation of existing biological systems, mainly because one needs to collect meaningful statistics, by performing many simulation replicates under a given set of conditions, in order to draw conclusions in an unbiased manner. In reality, many regulatory networks and their parts are usually multifunctional (*e.g.* TFs with dual regulatory roles). By necessity, analysis of their evolvability implies a focus toward one or a few particular network functions, and the degree of evolvability of such functions, most likely, does not necessarily equate with the evolutionary potential of other functions. Because of these biological and computational limitations, one is restricted to exploring only a limited number of alternative wiring configurations and only a limited region of sequence space; one has to restrict oneself to some criteria of network behavior (autonomous, stable oscillations of an appropriate period) at the expense of others (*e.g.* oscillator entrainment by light, bi-stable behavior, pulse-like expression dynamics, etc.); and one can explore only a very limited number of alternative mathematical representations of the biological system under study. Obviously, these limitations may preclude precise estimates and comprehensive understanding of the evolutionary potential of biological systems. Despite all these limitations, one might expect some common patterns to emerge, for instance, at a more coarse-grained description of molecular systems, such as the network topology level[216].

Although our study provides fresh insights into the evolutionary potential of GRNs, it is fair to say that our mechanistic GPM modeling approach offers only a first insight into what is in reality a more complex multidimensional space of mutable network control parameters. For instance, our study has concentrated only on the mutationally accessible parameter space that specifies the wiring of GRNs. Therefore, in order to gain a more comprehensive idea on the real evolutionary potential of GRNs, the quantitative features of other regulatory layers must be accounted for and adequately incorporated as sequence-encoded parameters in the network models, in order to assess their evolutionary impact on, for instance, the evolvability of the system. In particular, it would be interesting to examine whether our predictions hold if network evolution is allowed to proceed via mutations modulating cooperative protein-protein interactions involved in DNA binding recognition[359] or in the non-linear degradation of multimeric proteins[362]. Finally, we have focused on studying the impact of gen(om)e duplications on the traversability of fitness landscapes by simulating mutational trajectories describing the evolution of individual GRN system configurations.

Nevertheless, a more comprehensive picture would require ensemble-based simulations to dissect the potential contribution of important population genetic parameters, such as population size and recombination, on the evolvability of GRNs. It is possible that if we consider evolution of populations of GRNs undergoing different sorts of genetic modifications, including point mutations, recombination and large scale genetic changes, highly likely additional insights will emerge that are beyond the scope of the current computational framework.

7.1.1 The impact of allopolyploidization on GRN expression dynamics and innovation

An important aspect a mechanistic GPM modeling framework can shed light on regards the short-term impact of allopolyploidization, following a genome merging event, on the expression dynamics of GRNs. Given that allopolyploidization is often accompanied by massive reorganization of the transcriptome[457–459], this phenomenon is of great interest in *e.g.* plant biology research[460], because it has the potential to improve biotechnologically relevant products. In plants, altered expression outputs in allopolyploids have been linked to a variety of factors, such as activation of transposable elements[461], the gain and loss of repeated sequences[462], and mostly (increased) heterozygosity, which is introduced by merging two distinct genomes. Importantly, the inability of regulatory networks from two divergent genomes to successfully coordinate their actions within a hybrid polyploid nucleus might be largely responsible for substantial gene expression changes detected in allopolyploids[460,463]. In fact, evidence seems to suggest that altered gene expression outputs in polyploids can be achieved without signs of genome reorganization and epigenetic changes, supporting the idea that cross interactions between divergent regulatory hierarchies could be largely responsible for this phenomenon (see[463] and references therein). Therefore, by creating genome-wide regulatory variation through hybridization and interaction between diverged regulatory hierarchies[460,463], allopolyploidy might promote speciation events[464], it can improve plant vigor by heterotic effects[126], and it may also prove advantageous in accessing new ecological niches or surviving ecological crises[459,465].

Recently, it has been shown that allopolyploidization can have profound effects in the modulation of regulatory pathways underlying important physiological and metabolic traits intimately linked to biomass production and heterosis traits[449,465,467]. However, several questions at the mechanistic level remain to be addressed as to how allopolyploidization modulates the dynamic behavior of molecular networks. Using the mechanistic GPM modeling framework developed in this work, we started to investigate the impact of allopolyploidization on the expression dynamics of GRNs. To do this, we have gathered large ensembles of allopolyploid GRNs by merging pairs of GRNs (parental lines) that have been previously evolved for oscillatory expression dynamics (using the same *in silico* evolution approach as explained in chapter 5). We have initially focused on GRNs evolved from particular start oscillatory configurations (ancestral genotypes) toward lower and higher-frequency oscillatory expression dynamics. Allopolyploid GRN system configurations have been assembled from haploid and diploid GRNs evolved from a

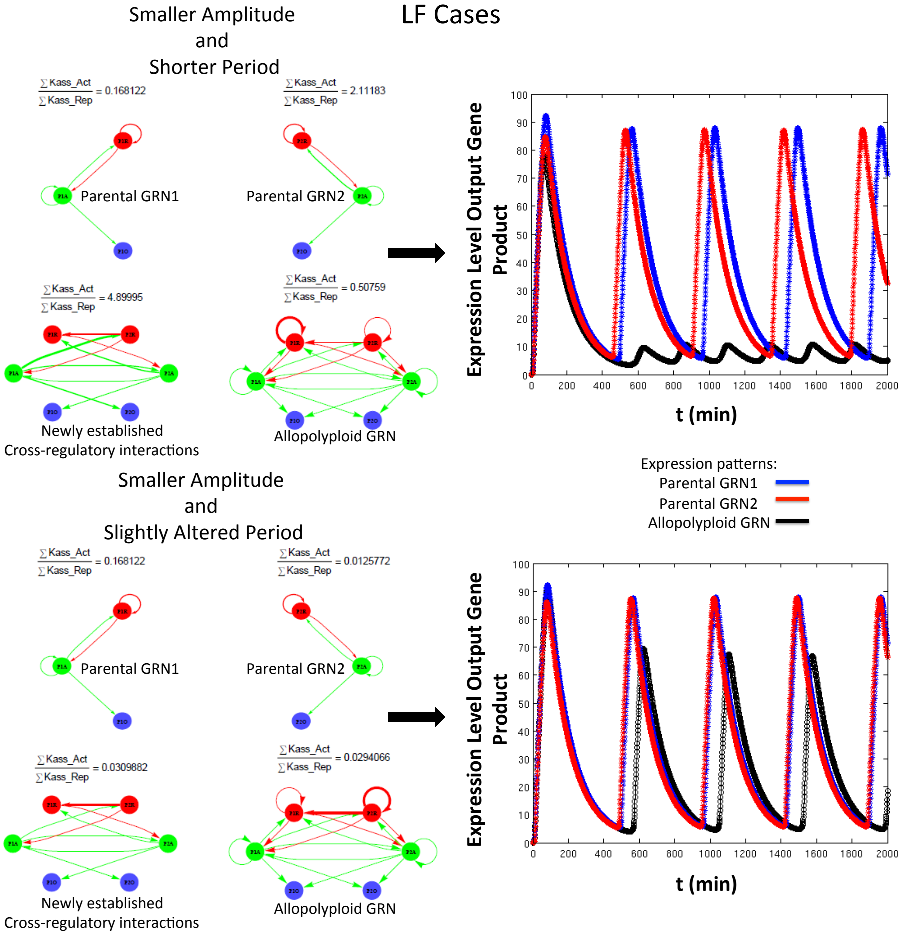


Figure 7.1: In silico genome merging events to investigate the impact of allopolyploidization in the expression dynamics of GRNs. *GRN genotypes previously evolved from the same ancestor(top panels) or distantly-related (bottom panels) GRN genotypes are joined together to create a hybrid polyploid. Newly established cross-regulatory interactions are then assessed, and the impact of these on the expression dynamics is evaluated and compared with the phenotypes of the parental lines. Illustrated are two types of crosses where the parental GRNs exhibit low-frequency (LF) oscillatory expression phenotypes. The thickness of the edges in the regulatory wirings displayed is proportional to the aggregated DNA binding strength of a given TF over all possible binding sites on the promoter regions of the target genes. Repressor and activator transcriptional regulators are shown in red and green, respectively; output genes are shown in blue. The quantity shown on top of each wiring represent the ratio (in logarithmic scale) of aggregated DNA binding strength of activating TFs to the aggregated DNA binding strength of repressing TFs, over the entire wiring.*

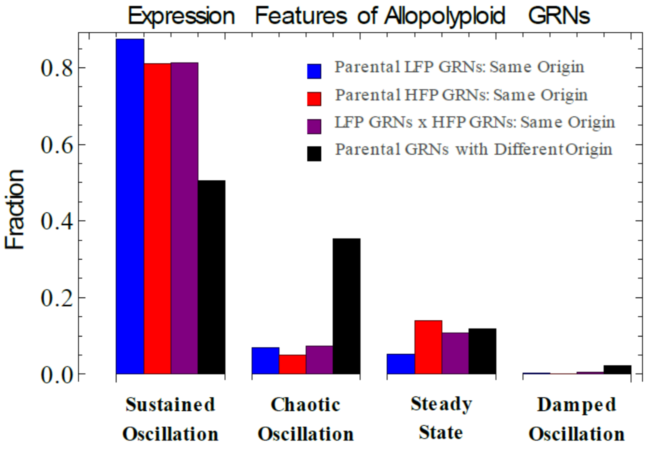
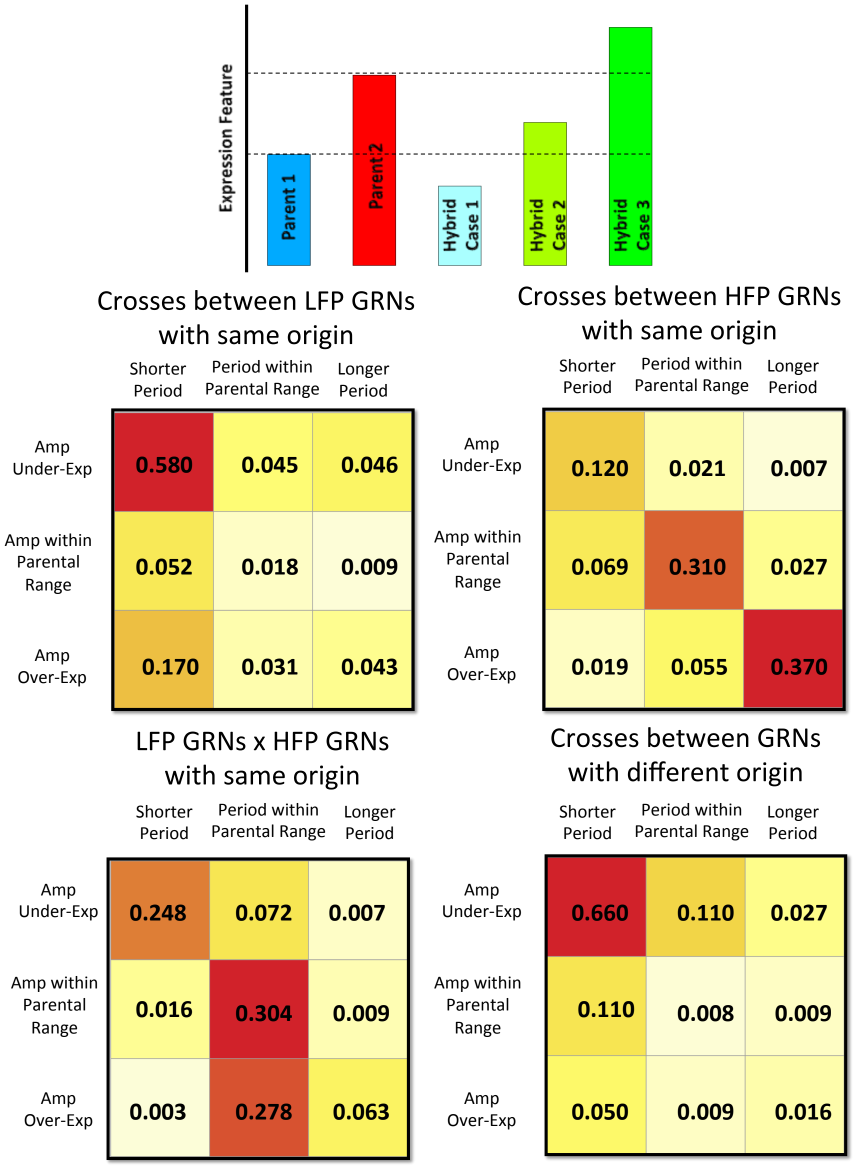


Figure 7.2: Assessing the fraction of different expression phenotype categories observed across the different ensembles of allopolyploid GRNs created. *Genome merging experiments were conducted using parental GRNs that have been evolved for low-frequency or high-frequency oscillatory expression phenotypes from the same ancestral, or distantly-related, GRN genotypes.*

given ancestral genotype. Importantly, int the context of these artificial systems, cross regulatory linkages between pairs of divergent genomes being merged together emerge as a regulatory system property in the newly formed allopolyploid GRN system configurations, which are established via individual protein-DNA interaction events, the quantitative properties of which are encoded in the genotype. Our preliminary analyses demonstrate that allopolyploidization represents an important source of regulatory innovation by creating a wide range of cross regulatory linkages between pairs of homeologous GRNs, which variably impact on expression dynamics. For instance, figure 7.1 illustrates two distinct genome merging events where one can appreciate the type of regulatory linkages formed in the newly established hybrid polyploid GRNs. Note that not only the type of cross regulatory linkages formed between pairs of homeologous GRNs can differ, but also the strength of the interactions (represented by the thickness of the regulatory edges in the wirings shown in figure 7.1), which reflects the extent of regulatory divergence between homeologous GRNs acquired over the course of evolution. Also note that in the two cases illustrated, genome merging can effectively modulate both the amplitude and the period of the oscillatory expression output in the allopolyploid system configurations compared to their parental GRNs.

Furthermore, our preliminary analyses also demonstrate that allopolyploidization in GRNs not only creates a wide spectrum of phenotypic variation in oscillatory expression dynamics, but that it can also induce drastic phenotypic transitions. For instance, we have observed that merging pairs of GRN genotypes that have been evolved from the same ancestral configuration, or from distantly related configurations, or that have been evolved toward different phenotypic optima (*e.g.* high-frequency or low frequency expression dynamics), can result in allopoly-



Crosses between LFP GRNs with same origin

	Shorter Period	Period within Parental Range	Longer Period
Amp Under-Exp	0.580	0.045	0.046
Amp within Parental Range	0.052	0.018	0.009
Amp Over-Exp	0.170	0.031	0.043

Crosses between HFP GRNs with same origin

	Shorter Period	Period within Parental Range	Longer Period
Amp Under-Exp	0.120	0.021	0.007
Amp within Parental Range	0.069	0.310	0.027
Amp Over-Exp	0.019	0.055	0.370

LFP GRNs x HFP GRNs with same origin

	Shorter Period	Period within Parental Range	Longer Period
Amp Under-Exp	0.248	0.072	0.007
Amp within Parental Range	0.016	0.304	0.009
Amp Over-Exp	0.003	0.278	0.063

Crosses between GRNs with different origin

	Shorter Period	Period within Parental Range	Longer Period
Amp Under-Exp	0.660	0.110	0.027
Amp within Parental Range	0.110	0.008	0.009
Amp Over-Exp	0.050	0.009	0.016

Figure 7.3: Assessing patterns of transgressively expressed features in oscillatory expression phenotypes. *Relying on the mid-parent value and the best-parent value notions*[126,466] *(see top panel) we have assessed the impact of genome merging on the amplitude and frequency of oscillatory expression phenotypes in allopolyploid GRNs with respect to their corresponding parental GRNs.*

ploid GRNs with different expression dynamics, such as chaotic oscillations, damped oscillations or steady state expression patterns (see figure 7.2). We have also assessed the prevalence of transgressive phenotypic features by comparing the amplitude and period of the oscillatory expression patterns in the allopolyploid GRNs with respect to their corresponding parental GRNs (see Figure 7.3), using, for instance the mid-parent value and the best-parent value notions [126,466]. Our preliminary results show that transgressive quantitative features in the oscillatory expression phenotypes of allopolyploid GRN system configurations are quite common, and that the pattern of transgressively expressed features observed largely depends on whether the parental lines share the same (recent) ancestral origin, or derive from distantly related ancestral configurations, as well as on the type of expression phenotype they have been evolved to. To shed further light on the impact of genome merging and hybridization on the expression dynamics of GRNs, several additional analyses must be performed. For instance, the relationship between the extent of divergence between pairs of parental GRNs and the impact of genome merging on expression dynamics is a key aspect of allopolyploidization. To investigate this, one could perform further evolutionary simulations from particular start genotypes, by applying both stabilizing and directional selection on expression dynamics, and assemble allopolyploid GRN system configurations from parental GRNs sampled at different time points over the course of evolutionary runs. The expression phenotypes of allopolyploids and corresponding parental GRNs could be monitored at different time points over the course of evolution in order to assess the way in which the extent of divergence between parental GRNs relates to changes in the expression dynamics of allopolyploid system configurations, compared to the phenotypes of the parental GRNs.

7.1.2 Evolutionary systems biology approaches to study cancer cellular networks

Unraveling the consequences of large-scale genetic perturbations (*e.g.* gene and genome duplications) is not only central to understanding the origin of intriguing evolutionary processes involving long time periods, such as species diversification and biological complexification, but is also crucial to deciphering important cell biological phenomena such as carcinogenesis, which is itself the result of an evolutionary process driven by the principles of mutation and selection by which cells acquire malignant phenotypes during the lifespan of an organism (somatic evolution). From a genomic point of view, large-scale modifications such as chromosomal amplifications, deletions, inversions, and translocations are key signatures of malignant cells[468,469]. Interestingly, most tumors have been found to have genomes in the triploid to tetraploid range[470], with tetraploid intermediates being typically found in both murine and human cancers[471]. By rewiring molecular networks, such large-scale genetic modifications drastically increase the chances of a cell to acquire a set of cancer hallmark traits[472–474] at once, and then transform a slow-growing cancer clone into a fast-growing one, therefore speeding up tumor formation[473,474]. Based on this idea, it has been suggested that genome duplication could be the rate-limiting step for tumor development, and therefore could be an early-warning signal for fast-growing clone formation[473]. Intriguingly, under this somatic evolutionary scenario, large-scale genetic modifications may not be intrinsically linked to either beneficial or detrimental changes in fitness[439]; rather, the impact of a “macromutation” would depend on the shape of the fitness landscape (defined by the nature of the selection pressure imposed by a tissue micro-environment) and the nature of the macromutation under consideration, which could bring about specific phenotypic changes, *i.e.* at the gene expression level, with potential selective advantages for malignant cells[439].

Given that cancer cellular phenotypes result from the deregulation of complex cellular information processing networks (*i.e.* transcriptional, signaling and metabolic), evolutionary systems biology approaches hold promise for future exciting avenues for research on the somatic evolutionary process underlying carcinogenesis. Firstly, network modeling can help us identify the mechanistic basis of robustness and fragility of cancer hallmark networks[473,474]. For instance, a recent study focused o a large-scale characterization of network motifs in cancer signaling pathways suggests that positive signaling regulatory network motifs tend to be preferentially used by cancer driving mutating oncogenes. In contrast, negative signaling regulatory network motifs were found to be preferentially used by methylated genes and tumor suppressors in cancer cells[475]. These observations can serve as a point of departure for the development of mechanistic network models to attempt a classification of cancer network motifs in terms of dynamics and functionality[476], which could help us rationalize the wealth of (topological) information from high-throughput experiments in terms of, for instance, a network's information processing capabilities (*e.g.* the ability to convert an array of environmental stimuli into particular gene expression patterns)[476]. A dynamical picture of cancer hallmark networks[473,474] would eventually aid in future research on personalized and predictive cancer therapies. Secondly, from an evolutionary perspective, the most interesting aspect of carcinogesis regards the impact of all sorts of macromutations (*e.g.* single gene duplication/deletion events, chromosomal amplifications, and polyploidization) on the information processing capabilities of cancer network motifs, which can bring about big phenotypic leaps resembling punctuated equilibria dynamics[477] during the somatic evolution of malignant cells[478]. In this context, the arsenal of mechanistic GPM modeling tools and *in silico* evolution approaches offered by evolutionary systems biology holds great promise for advancing our understanding of the somatic evolutionary process driving the transformation of normal cells into malignant ones. In particular, an evolutionary systems biology approach to carcinogenesis could shed new light on how macromutations rewire cancer hallmark networks[473,474], which is key to understanding the fitness landscape of malignant cells[472].

A

Summary

The relationship between the genotype and the physical or biochemical characteristics of biological systems, referred to as the genotype-phenotype map (GPM), lies at the center of most research fields in the life sciences. Most GPMs are intrinsically linked to the functioning of complex molecular networks (*e.g.* the regulatory networks controlling the expression of cellular phenotypes), which operate in a highly non-linear manner. This makes the study of GPMs by mere intuitive reasoning alone quite challenging. The inter-disciplinary field of systems biology offers an ample range of quantitative tools to study not only the inner workings of molecular networks but also to shed light on the evolutionary potential (evolvability) and origin of emergent systems properties such as robustness and modularity. Although some of the mechanistic underpinnings of molecular networks can be reasonably captured in most systems biology-inspired network models, they generally fail to account for the genetic encoding of the systems. In this sense, these models are of limited use to adequately study the evolutionary potential of molecular networks. In this work, I concentrated on designing an adequate mechanistic network modeling framework to study the impact of gene and genome duplications on the evolvability of gene regulatory networks (GRNs). I demonstrate that the use of a fine-grained model of the GPM allows the study of the evolutionary potential of GRNs at an unprecedented detail. In particular, by simulating the evolution of GRNs across an explicitly defined genotype/sequence space, rather than in a continuous parameter space, I was able to examine how GRNs acquire increasingly better adapted dynamic expression phenotypes (our operational definition of evolvability) as they navigate the fitness landscape via discrete changes in the genotype. This allowed me to quantify the impact of duplication of a GRN system configuration on the evolvability of oscillatory expression phenotypes, as well as to assess the relative contribution of changes in *cis*-regulatory regions and *trans*-acting elements in the adaptation of GRNs. Furthermore, my simulation results provide fresh insight into the proximate and ultimate consequences of dosage balance effects. In particular, I found that the model GRNs exhibit a broad range of phenotypic responses to single gene duplication and deletion, as well as to amplification, of activator, repressor and output genes. In addition, *in silico* evolution of GRNs under dosage balance constraints demonstrated that due to pervasive detrimental changes in fitness, single duplications of regulatory genes generally have an adverse impact on the evolvability of GRNs, although a few exceptions exist where the imbalanced GRNs outperform their haploid and diploid counterparts (balanced GRNs). Overall, the work presented here has revealed an intricate multifactorial basis of the evolvability of GRNs, which can be difficult to dissect through coarse-grained mathematical representations of the GPM.

“Writers should not fear jargon”

Trevor Quirk

B

Glossary

Boolean: A data type with only two possible values (e.g., true or false, zero or one, present or absent, or on or off).

Emergent property: A feature that is not a property of any individual part of the system, but only emerges in the context of the entire system from how the individual parts interact.

Epistasis: epistatic effects between alleles are deviations from the expected additive phenotype or fitness effects of allelic changes, due to interaction between the alleles involved. Several kinds of epistasis can be distinguished (see Chapter 3 – Figure 3.3).

Evolvability: Can be defined as 1) the extent to which a population can produce new selectable allelic variation; or 2) as the internal disposition of biological systems to vary in the face of genetic perturbations, which determines their potential for future evolutionary change (adaptations).

Fitness landscape: A metaphorical representation of the relation between the genotype and fitness of an organism. Fitness landscapes are often depicted as 3D landscapes, but are high-dimensional functions in reality.

Group selection: Group selection for a particular trait entails that the trait is selected for on the level of (sub)populations rather than on the level of individuals.

Modularity: The extent to which a system or network can be subdivided in modules, or independently functioning parts.

Neutral network: A set of genotypes that have equivalent fitness and that are linked through neutral mutation pathways.

Robustness: The extent to which a system is phenotypically insensitive to perturbation. Two general kinds of robustness are usually distinguished for molecular systems: genetic robustness (i.e., invariance to mutation or recombination) and environmental robustness (i.e., invariance to micro- or macroenvironmental perturbations).

Statistical thermodynamics: A branch of statistical physics that studies the average behavior of a thermodynamical system (e.g., the interaction of macromolecules) using probability theory.

Transfer function: a mathematical representation of the inputoutput relation of a linear time-invariant system.

"A picture is worth a thousand words"

proverb

C

Supplementary Figures

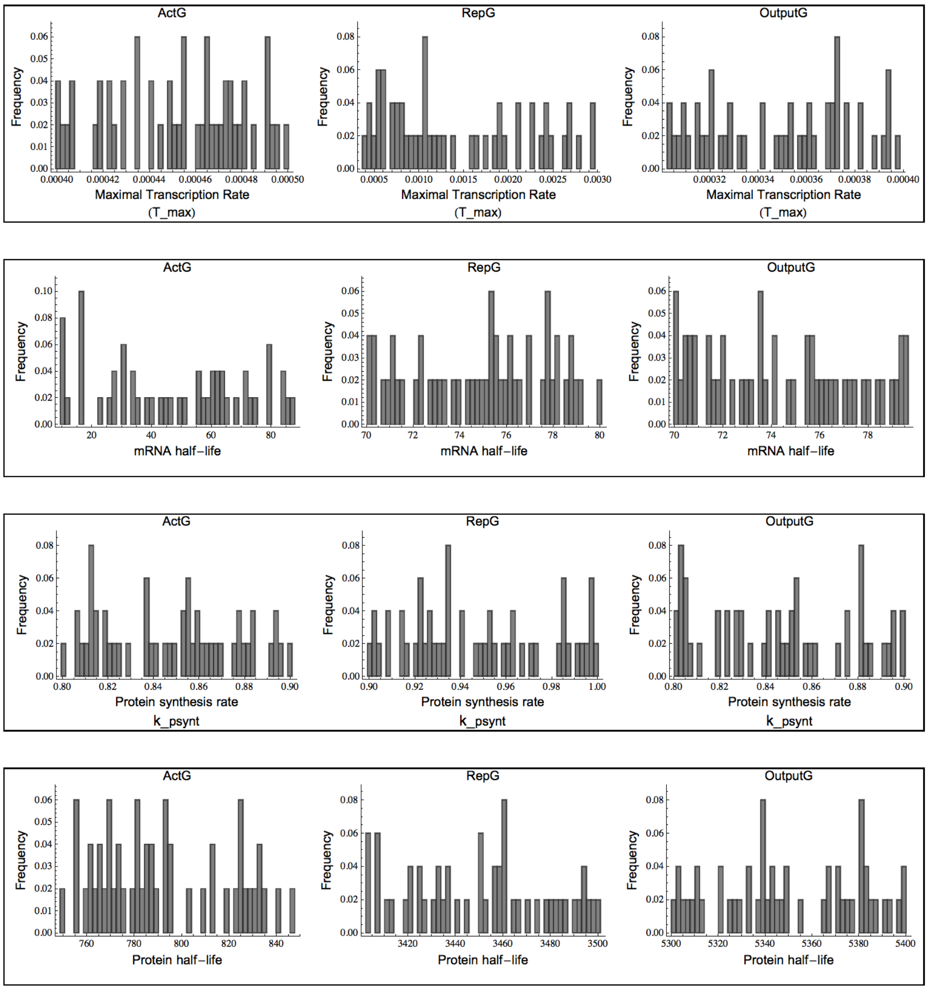


Figure C.1: Distribution of basal kinetic rates for the GRN configurations used as starting points of evolution toward new phenotypic optima.

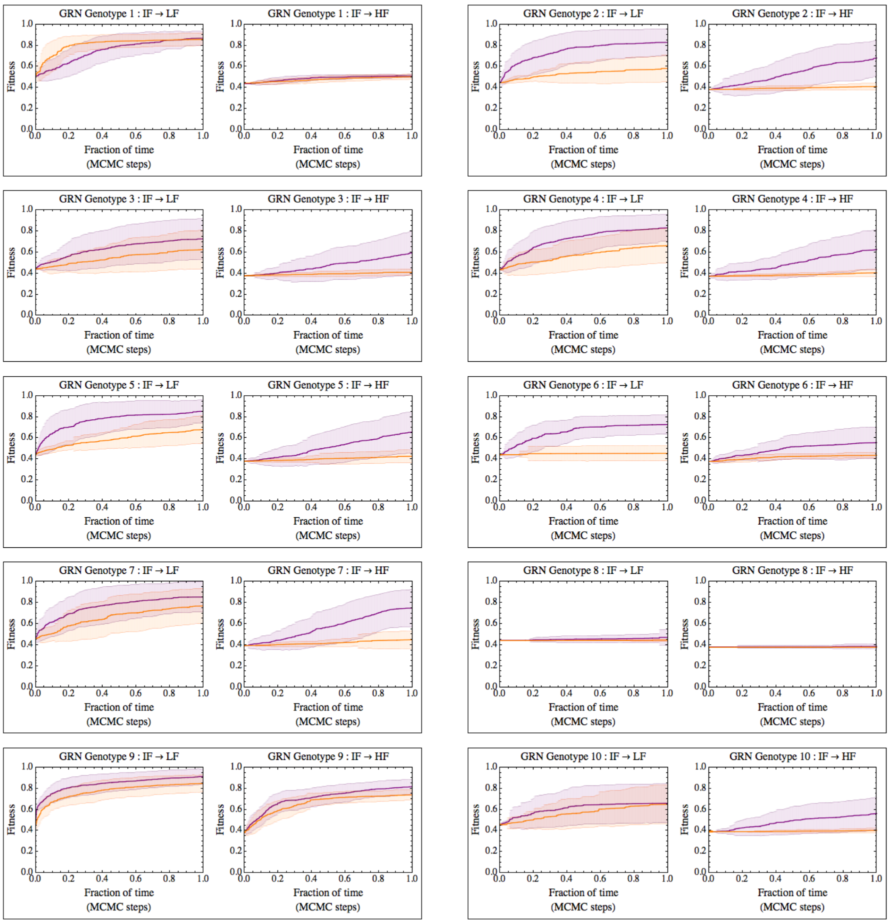


Figure C.2: Diploid vs. haploid average fitness trajectories: Fitness function FF1, set I. *Temporal sequences of fitness values recorded from 50 independent simulation replicates, using 10 different start GRN configurations, were averaged out to display the general trend of the adaptation process toward a new phenotypic optimum. Evolutionary simulations were performed using an implementation of the NEA (neutral-evolution-allowed) algorithm with the multi-objective fitness function FF1 (see subsection 4.1.6.2). Error bars along the trajectories indicate standard deviations. In view of the fact that duplicated (diploid) GRN system configurations present a mutational target twice the size of haploid GRNs, the simulated evolutionary time window for diploids (purple-color coded trajectories) spans twice the number of MCMC steps considered for haploids (orange-color coded trajectories). Specifically, the length of the simulated mutational pathways was set according to the total number of mutable sites per genome (effective genome size), as follows: diploid GRNs were evolved for 1320 (diploid effective genome size) x 5 = 6600 MCMC steps, whereas haploid GRNs were evolved for 660 (haploid effective genome size) x 5 = 3300 MCMC steps, and the the time (x) axis in the plots was re-scaled to 1.*

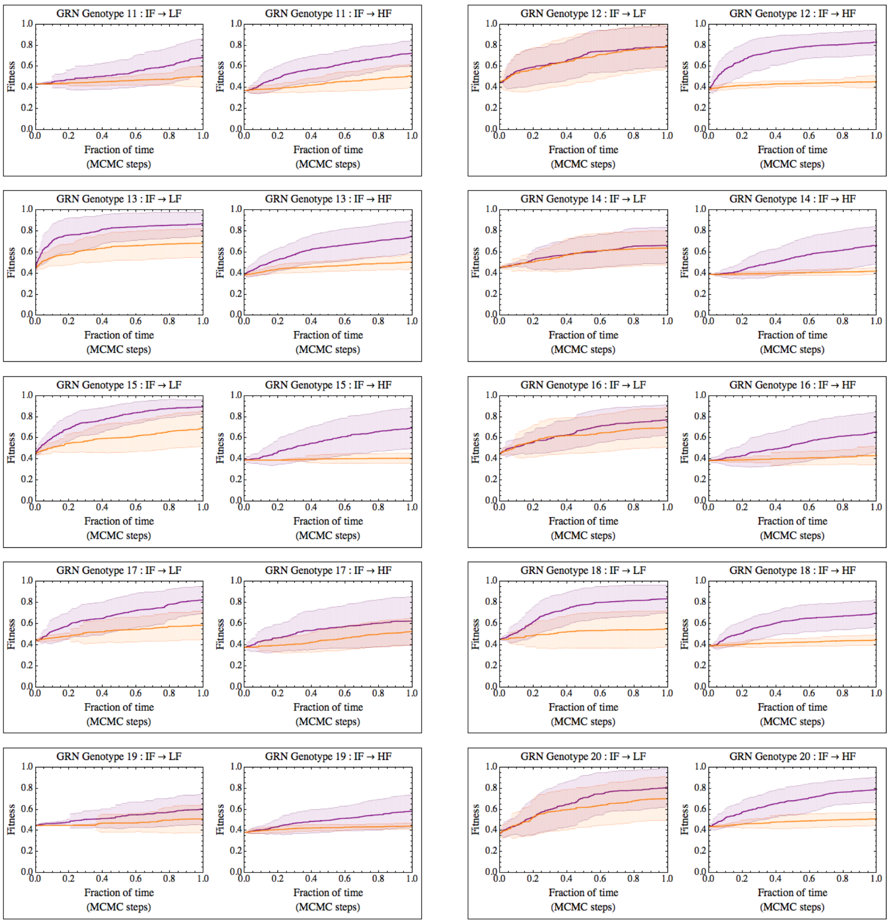


Figure C.3: Diploid vs. haploid average fitness trajectories: Fitness function FF1, set II. *Description is the same as in Figure C.2.*

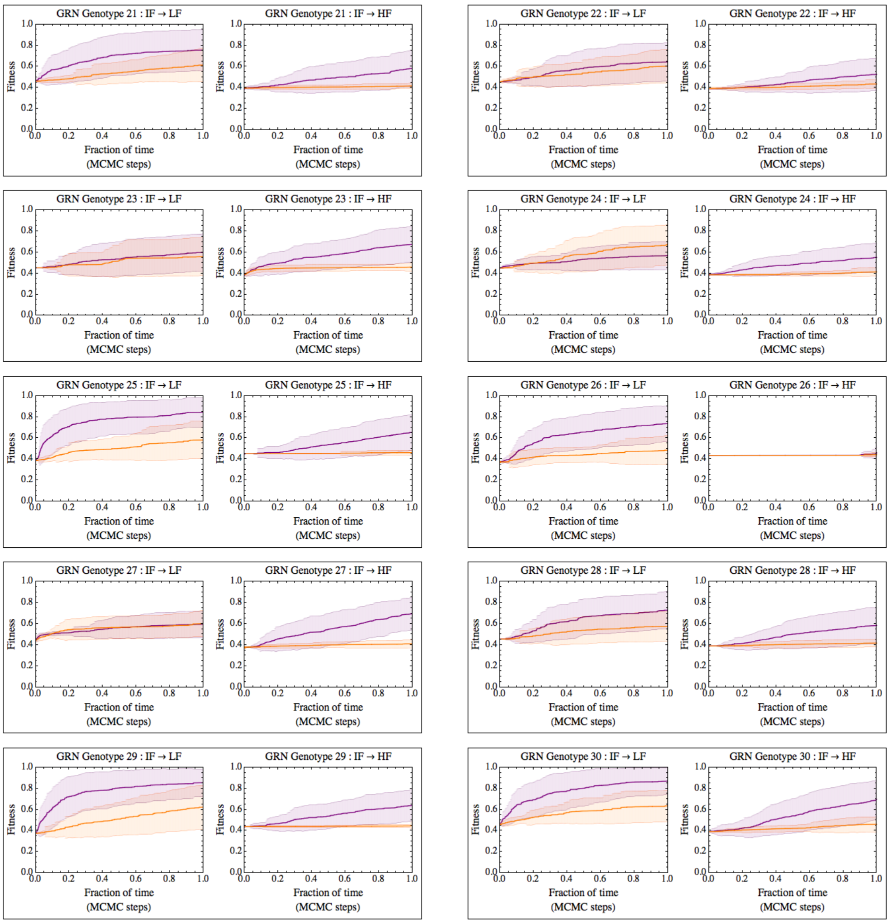


Figure C.4: Diploid vs. haploid average fitness trajectories: Fitness function FF1, set III. *Description is the same as in Figure C.2*

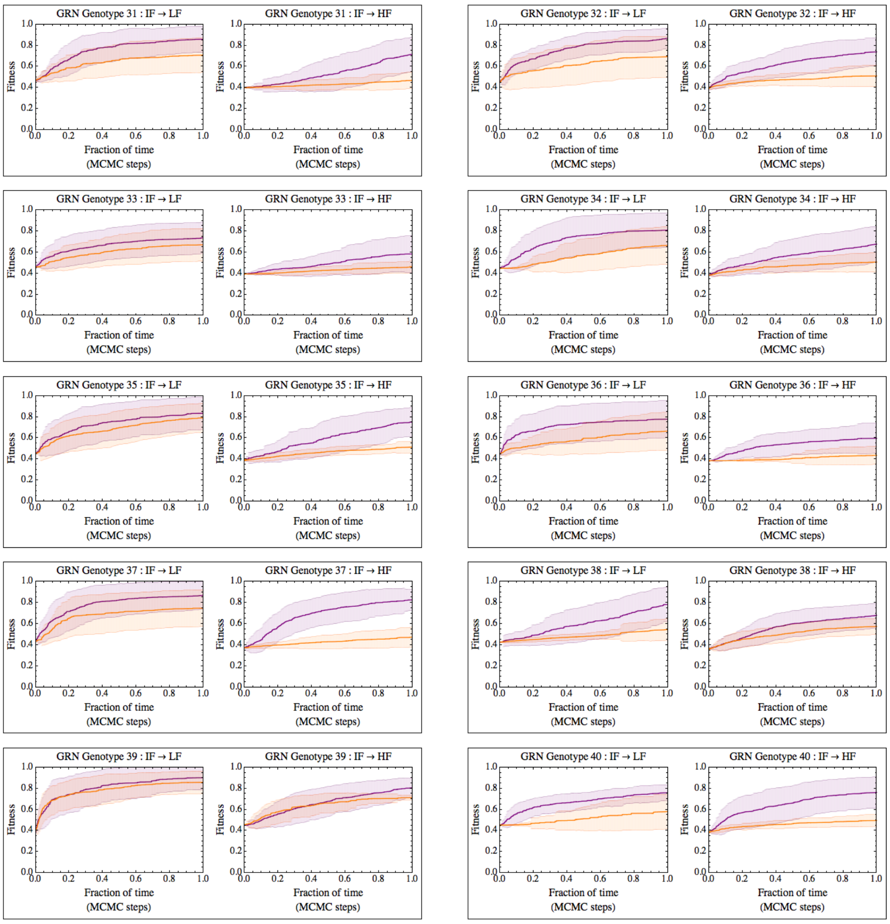


Figure C.5: Diploid vs. haploid average fitness trajectories: Fitness function FF1, set IV. *Description is the same as in Figure C.2.*

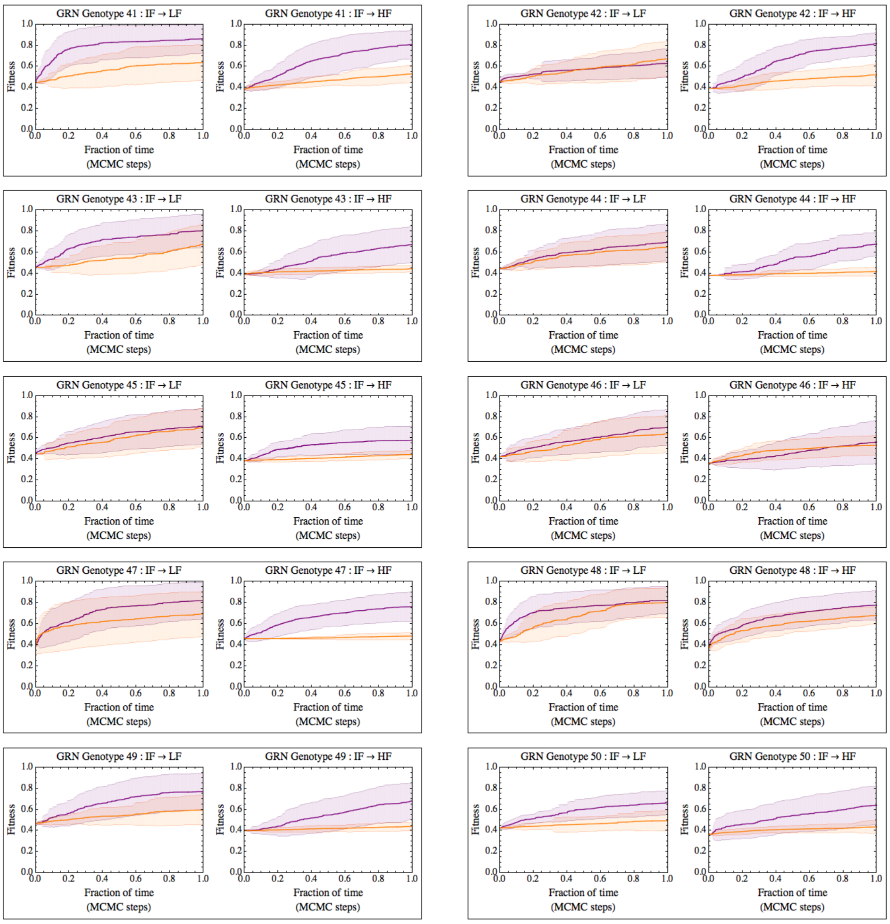


Figure C.6: Diploid vs. haploid average fitness trajectories: Fitness function FF1, set V. *Description is the same as in Figure C.2.*

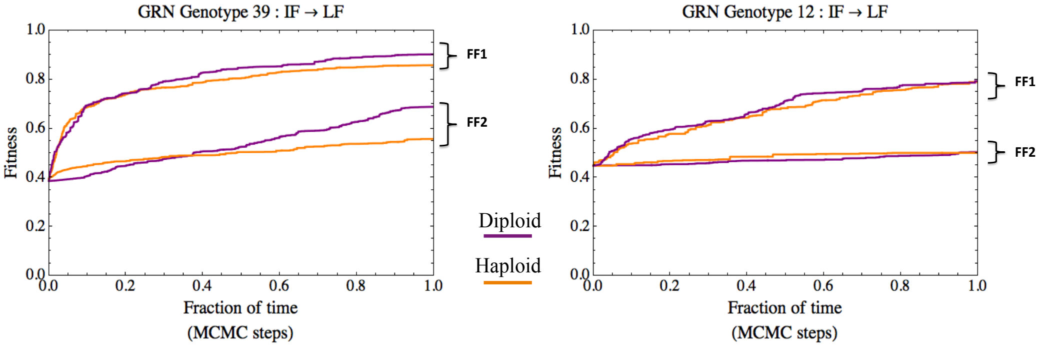


Figure C.7: Comparison of evolutionary trajectories simulated under FF1 vs. FF2. *Temporal sequences of fitness values recorded from 50 independent simulation replicates, using start GRN configurations 12 and 39, were averaged out to display the general trend of the adaptation process toward a new phenotypic optimum. Evolutionary simulations were performed using an implementation of the NEA (neutral-evolution-allowed) algorithm with the multi-objective fitness functions FF1 and FF2 (see subsection 4.1.6.2). Error bars along the trajectories indicate standard deviations. In view of the fact that duplicated (diploid) GRN system configurations present a mutational target twice the size of haploid GRNs, the simulated evolutionary time window for diploids (purple-color coded trajectories) spans twice the number of MCMC steps considered for haploids (orange-color coded trajectories). Specifically, the length of the simulated mutational pathways was set according to the total number of mutable sites per genome (effective genome size), as follows: diploid GRNs were evolved for 1320 (diploid effective genome size) x 5 = 6600 MCMC steps, whereas haploid GRNs were evolved for 660 (haploid effective genome size) x 5 = 3300 MCMC steps, and the the time (x) axis in the plots was re-scaled to 1.*

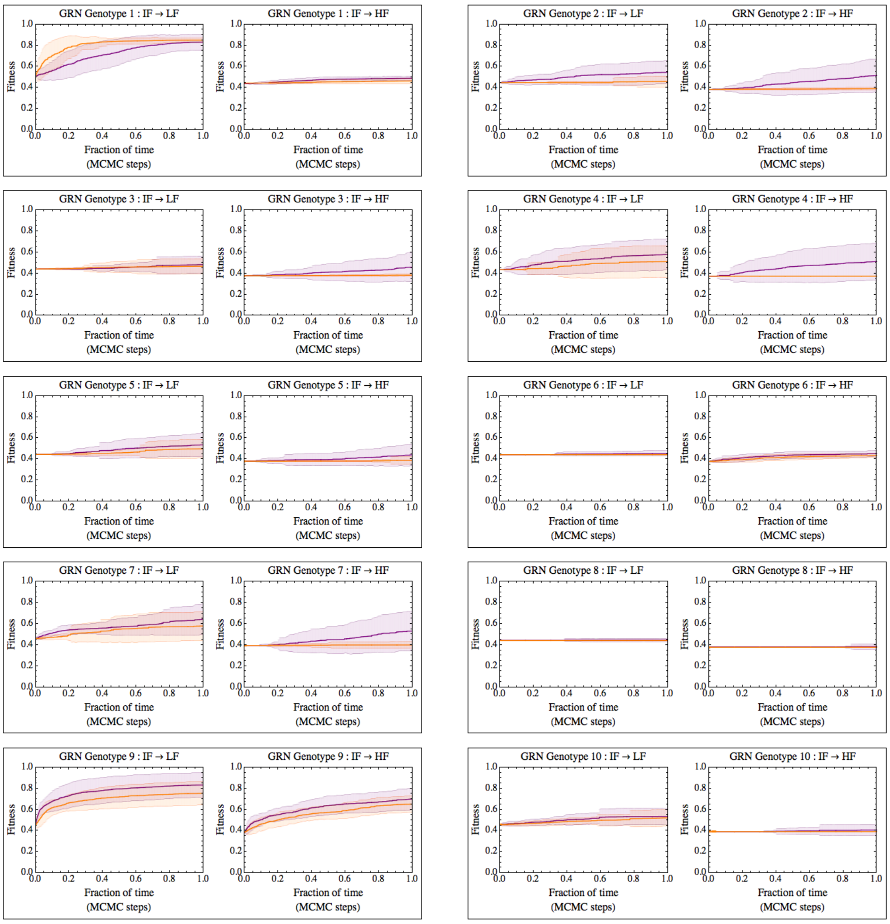


Figure C.8: Diploid vs. haploid average fitness trajectories: Fitness function FF2, set I. *Temporal sequences of fitness values recorded from 50 independent simulation replicates, using 9 different start GRN configurations, were averaged out to display the general trend of the adaptation process toward a new phenotypic optimum. Evolutionary simulations were performed using an implementation of the NEA (neutral-evolution-allowed) algorithm with the multi-objective fitness function FF2 (see subsection 4.1.6.2). Error bars along the trajectories indicate standard deviations. In view of the fact that duplicated (diploid) GRN system configurations present a mutational target twice the size of haploid GRNs, the simulated evolutionary time window for diploids (purple-color coded trajectories) spans twice the number of MCMC steps considered for haploids (orange-color coded trajectories). Specifically, the length of the simulated mutational pathways was set according to the total number of mutable sites per genome (effective genome size), as follows: diploid GRNs were evolved for 1320 (diploid effective genome size) x 5 = 6600 MCMC steps, whereas haploid GRNs were evolved for 660 (haploid effective genome size) x 5 = 3300 MCMC steps, and the the time (x) axis in the plots was re-scaled to 1.*

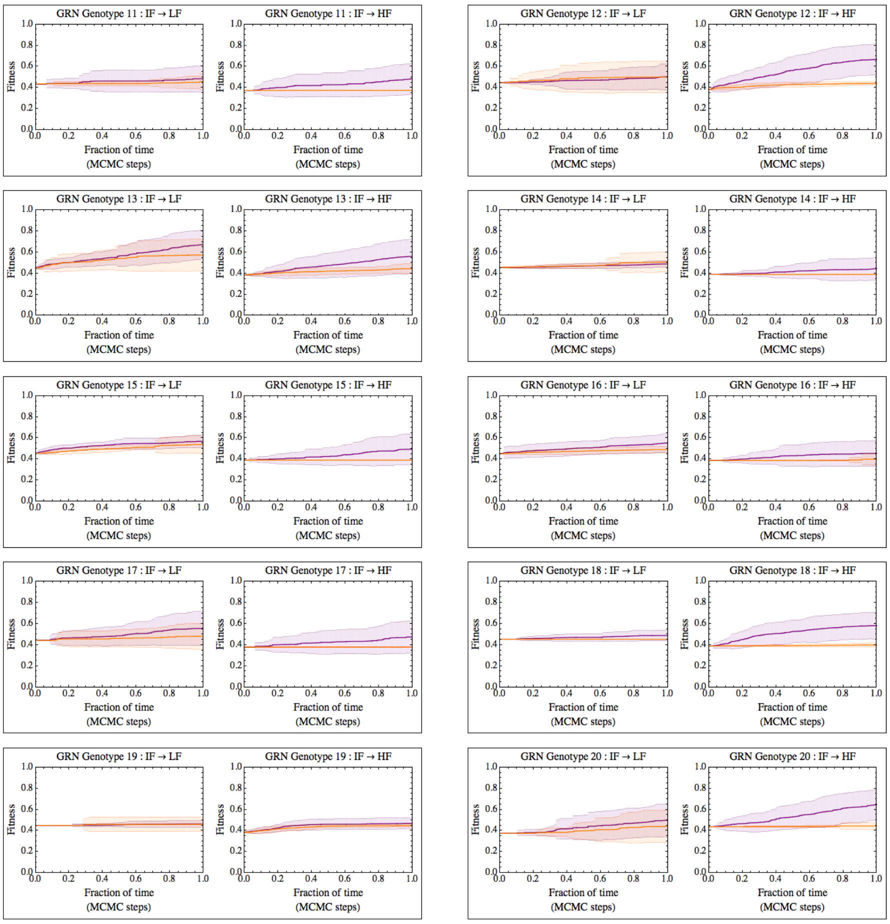


Figure C.9: Diploid vs. haploid average fitness trajectories: Fitness function FF2, set II. *Description is the same as in Figure C.8.*

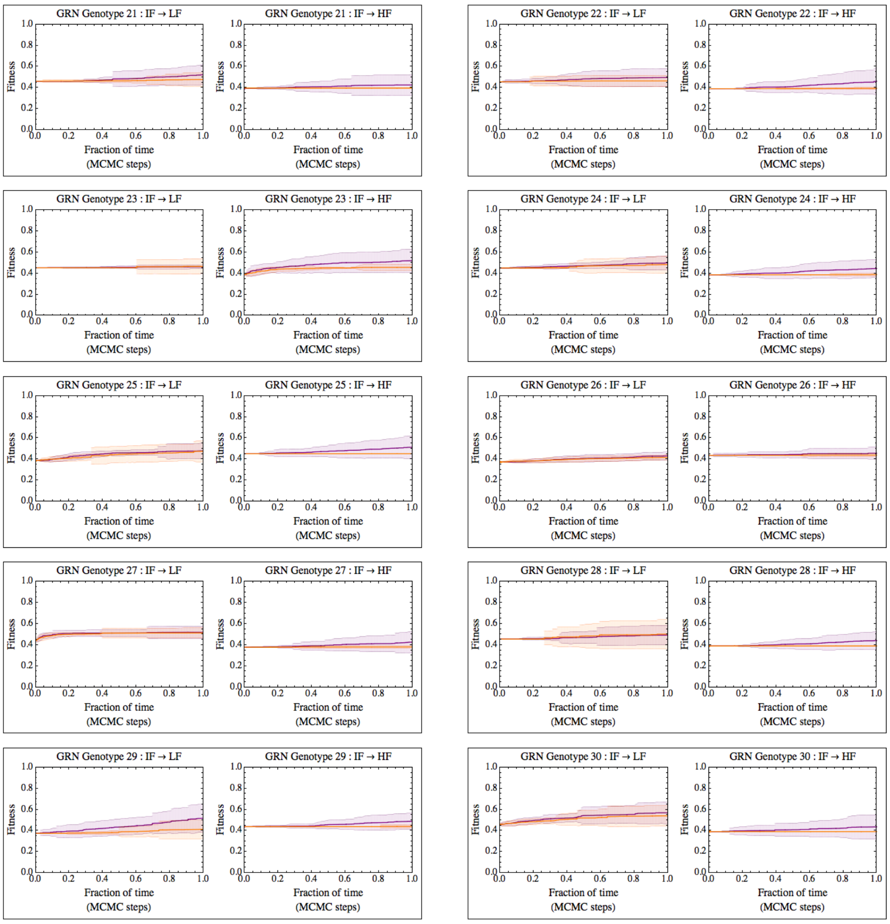


Figure C.10: Diploid vs. haploid average fitness trajectories: Fitness function FF2, set III. *Description is the same as in Figure C.8.*

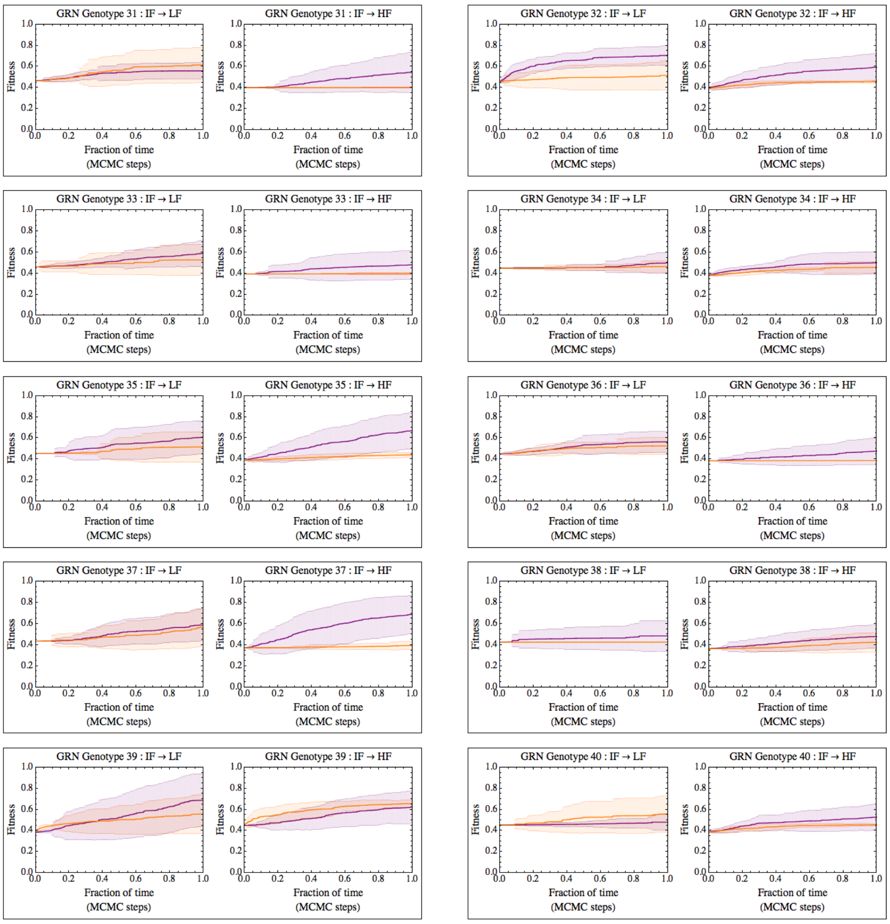


Figure C.11: Diploid vs. haploid average fitness trajectories: Fitness function FF2, set IV. *Description is the same as in Figure C.8.*

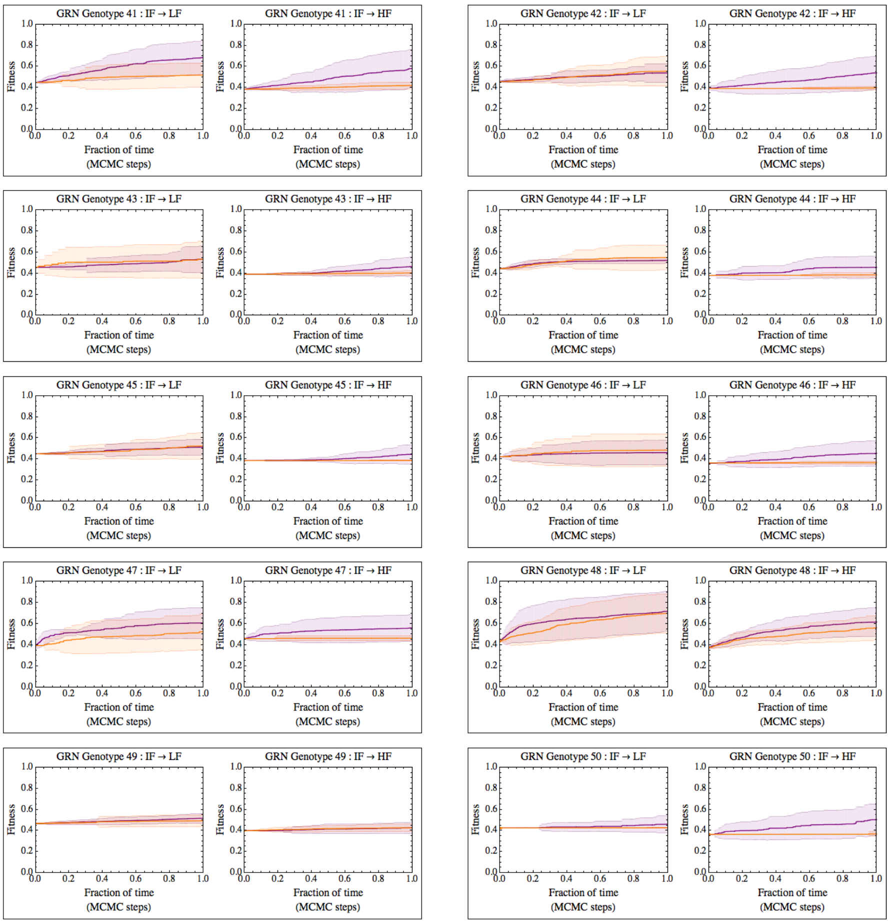


Figure C.12: Diploid vs. haploid average fitness trajectories: Fitness function FF2, set V. *Description is the same as in Figure C.8.*

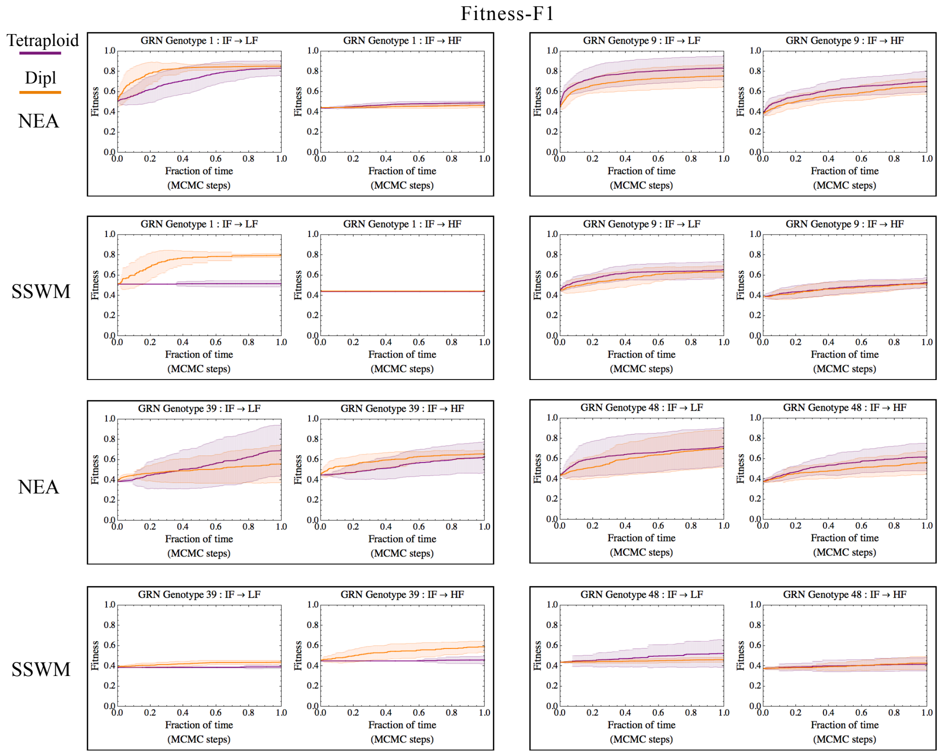


Figure C.13: Evolutionary performance of GRNs under NEA vs. SSWM. *The plots display the average fitness trajectories for diploid and haploid GRNs recorded from particular start GRN configurations. Evolutionary simulations were performed using implementations of the NEA (neutral-evolution-allowed) algorithm and the adaptive walk algorithm (under the strong selection-weak mutation (SSWM) regime) with the multi-objective fitness function FF1 (see subsection 4.1.6.2). In view of the fact that duplicated (diploid) GRN system configurations present a mutational target twice the size of haploid GRNs, the simulated evolutionary time window for diploids (purple-color coded trajectories) spans twice the number of MCMC steps considered for haploids (orange-color coded trajectories). Specifically, the length of the simulated mutational pathways was set according to the total number of mutable sites per genome (effective genome size), as follows: diploid GRNs were evolved for 1320 (diploid effective genome size) x 5 = 6600 MCMC steps, whereas haploid GRNs were evolved for 660 (haploid effective genome size) x 5 = 3300 MCMC steps, and the the time (x) axis in the plots was re-scaled to 1.*

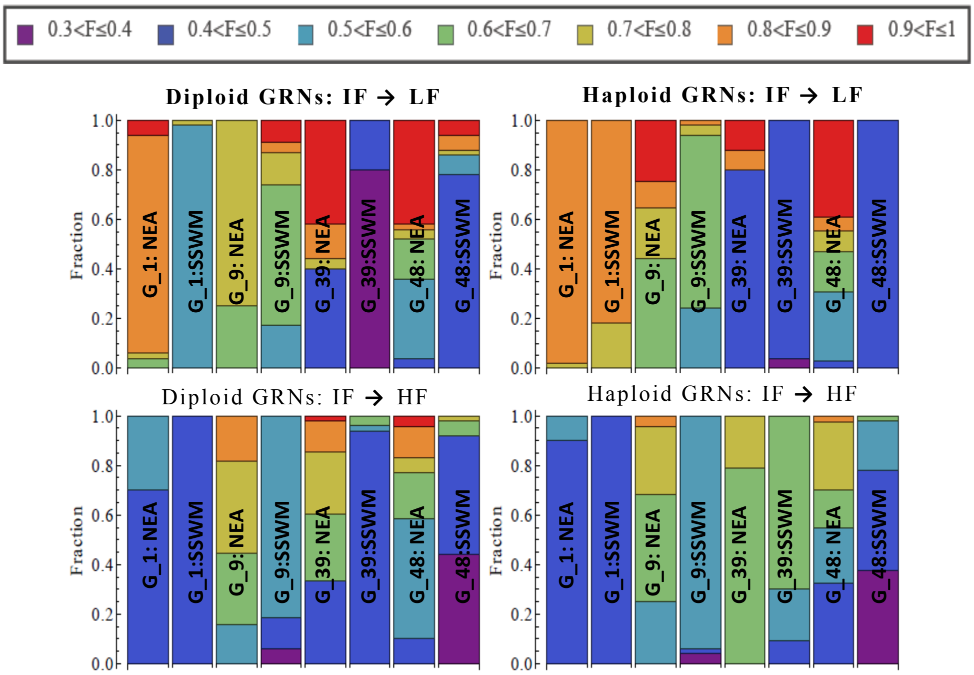
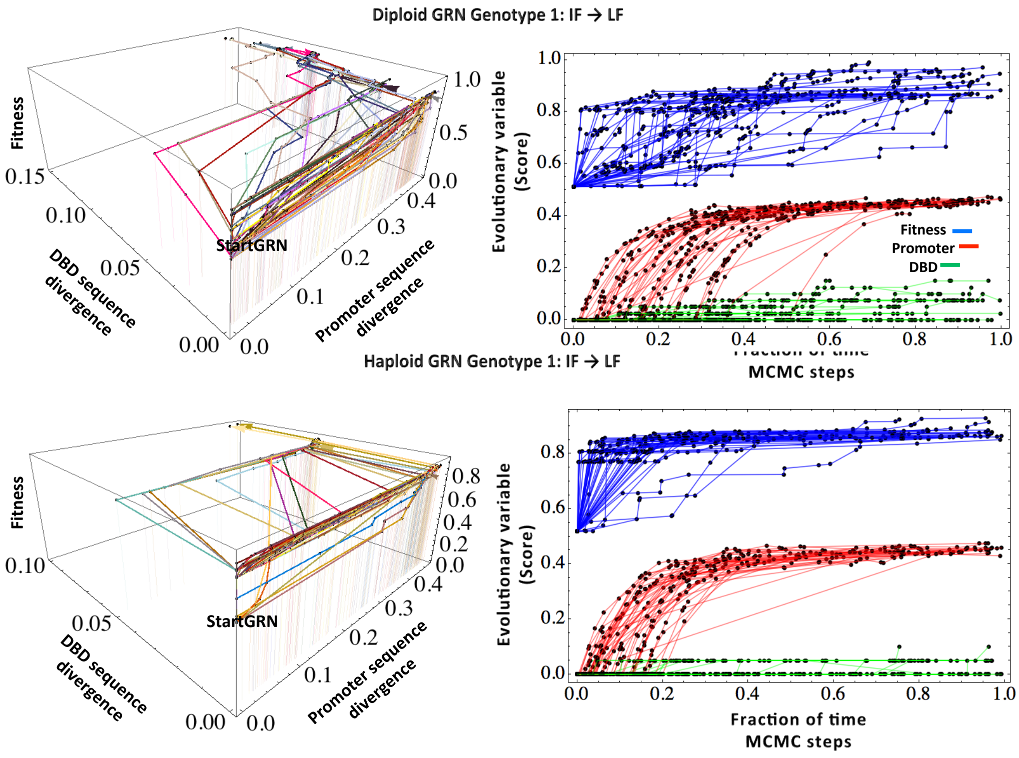


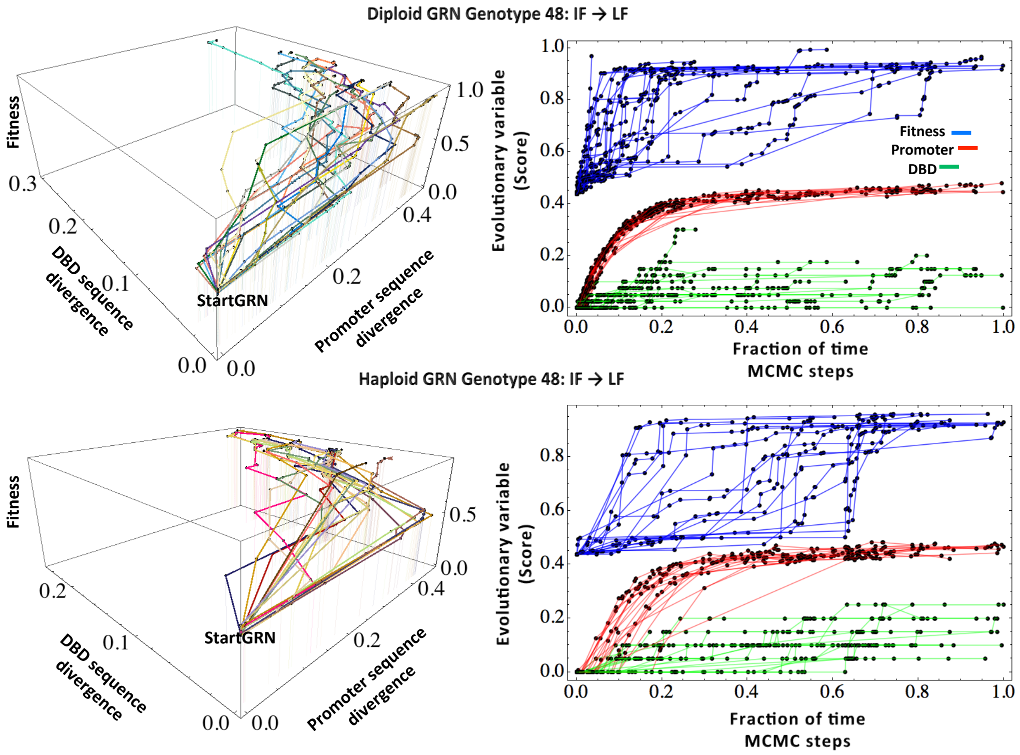
Figure C.14: Comparison of NEA vs. SSWM accessible end point fitness values. *Distribution of end point fitness values recovered from 50 simulation replicates using particular start GRN configurations (labeled as G_x). End point fitness values attained were allocated in predefined (color-coded) ranges as indicated by the legend shown on top of the panels. Data generated using Fitness-F1 (see subsection 4.1.6.2)*



Diploid GRN Genotype 1: IF → LF

Haploid GRN Genotype 1: IF → LF

Figure C.15: Evolutionary phase space and sequence divergence plots: Fitness function FF1, GRN genotype 1 *Panels on the left depict the phase space of evolving GRNs. Coordinates shown represent the proportion of accumulated changes (with respect to start GRN configurations) in DBD encoding sequences and promoter regions (x, y coordinates), as well as their associated fitness score (z coordinate). Sequence divergence is assessed with respect to the start GRN configurations using a normalized Edit distance, which measures the percentage of dissimilarity (in terms of DNA sequence for promoters and of amino acids for DBDs) between a given ancestral sequence and mutant sequences sampled at a given time point (t_i) over the course of evolution. The concatenation of a sequence of triplets $(x(t_i), y(t_i), z(t_i))$ sampled at different time points over an evolution run describes a trajectory across the phase space. Triplets $(x(t_i), y(t_i), z(t_i))$ projected on the evolutionary phase space were sampled every time there was a fitness increment of $\Delta F \geq 0.001$ over the course of evolution toward a new optimum. Panels shown on the right illustrate the temporal sequence for individual components in the triplets $(x(t_i), y(t_i), z(t_i))$ recorded among different replicates. Time is depicted as the fraction of the total number of MCMC steps simulated. The y axis represents either percentage of dissimilarity between ancestral sequence and mutant sequences (for both promoter regions and DBD encoding sequences, separately) or the fitness value sampled at a given time point (t_i). Evolutionary simulations were performed using an implementation of the NEA (neutral-evolution-allowed) algorithm with FF1 (see subsection 4.1.6.2).*

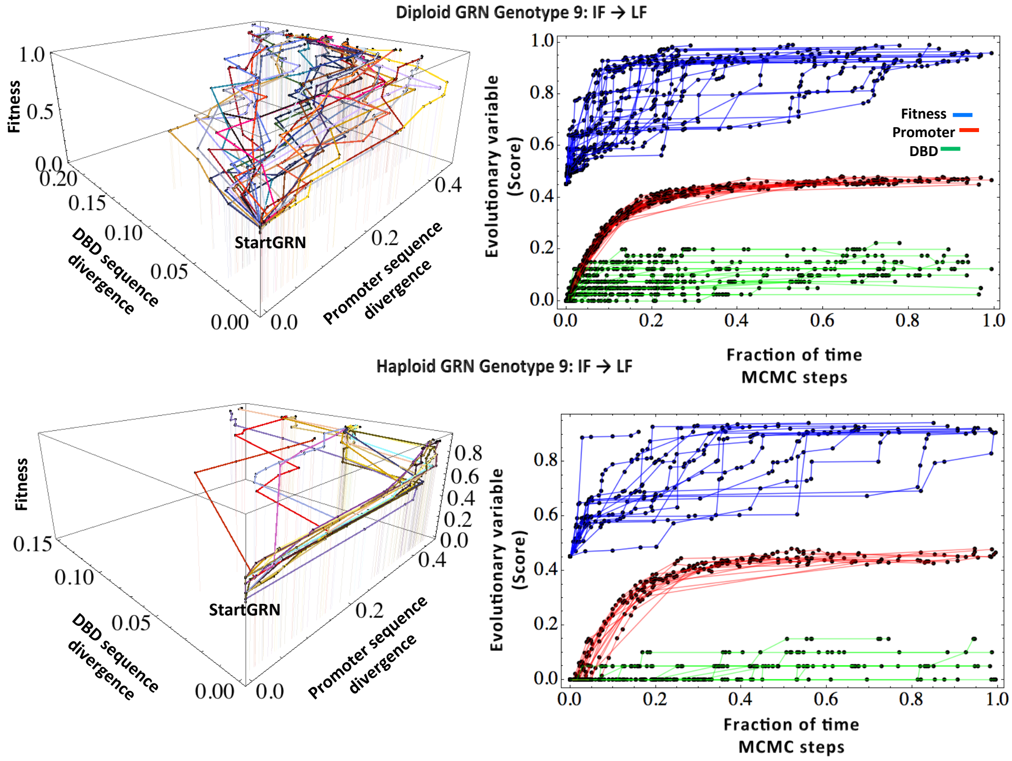


Diploid GRN Genotype 48: IF → LF

Haploid GRN Genotype 48: IF → LF

Figure C.16: Evolutionary phase space and sequence divergence plots: Fitness function FF1, GRN genotype 48 *Panels on the left depict the phase space of evolving GRNs. Coordinates shown represent the proportion of accumulated changes (with respect to start GRN configurations) in DBD encoding sequences and promoter regions (x, y coordinates), as well as their associated fitness score (z coordinate). Sequence divergence is assessed with respect to the start GRN configurations using a normalized Edit distance, which measures the percentage of dissimilarity (in terms of DNA sequence for promoters and of amino acids for DBDs) between a given ancestral sequence and mutant sequences sampled at a given time point (t_i) over the course of evolution. The concatenation of a sequence of triplets $(x(t_i), y(t_i), z(t_i))$ sampled at different time points over an evolution run describes a trajectory across the phase space. Triplets $(x(t_i), y(t_i), z(t_i))$ projected on the evolutionary phase space were sampled every time there was a fitness increment of $\Delta F \geq 0.001$ over the course of evolution toward a new optimum. Panels shown on the right illustrate the temporal sequence for individual components in the triplets $(x(t_i), y(t_i), z(t_i))$ recorded among different replicates. Time is depicted as the fraction of the total number of MCMC steps simulated. The y axis represents either percentage of dissimilarity between ancestral sequence and mutant sequences (for both promoter regions and DBD encoding sequences, separately) or the fitness value sampled at a given time point (t_i). Evolutionary simulations were performed using an implementation of the NEA (neutral-evolution-allowed) algorithm with FF1 (see subsection 4.1.6.2).*

Fitness-F1


Diploid GRN Genotype 9: IF → LF

Haploid GRN Genotype 9: IF → LF

Figure C.17: Evolutionary phase space and sequence divergence plots: Fitness function FF2, GRN genotype 9 *Panels on the left depict the phase space of evolving GRNs. Coordinates shown represent the proportion of accumulated changes (with respect to start GRN configurations) in DBD encoding sequences and promoter regions (x, y coordinates), as well as their associated fitness score (z coordinate). Sequence divergence is assessed with respect to the start GRN configurations using a normalized Edit distance, which measures the percentage of dissimilarity (in terms of DNA sequence for promoters and of amino acids for DBDs) between a given ancestral sequence and mutant sequences sampled at a given time point (t_i) over the course of evolution. The concatenation of a sequence of triplets $(x(t_i), y(t_i), z(t_i))$ sampled at different time points over an evolution run describes a trajectory across the phase space. Triplets $(x(t_i), y(t_i), z(t_i))$ projected on the evolutionary phase space were sampled every time there was a fitness increment of $\Delta F \geq 0.001$ over the course of evolution toward a new optimum. Panels shown on the right illustrate the temporal sequence for individual components in the triplets $(x(t_i), y(t_i), z(t_i))$ recorded among different replicates. Time is depicted as the fraction of the total number of MCMC steps simulated. The y axis represents either percentage of dissimilarity between ancestral sequence and mutant sequences (for both promoter regions and DBD encoding sequences, separately) or the fitness value sampled at a given time point (t_i). Evolutionary simulations were performed using an implementation of the NEA (neutral-evolution-allowed) algorithm with FF2 (see subsection 4.1.6.2).*

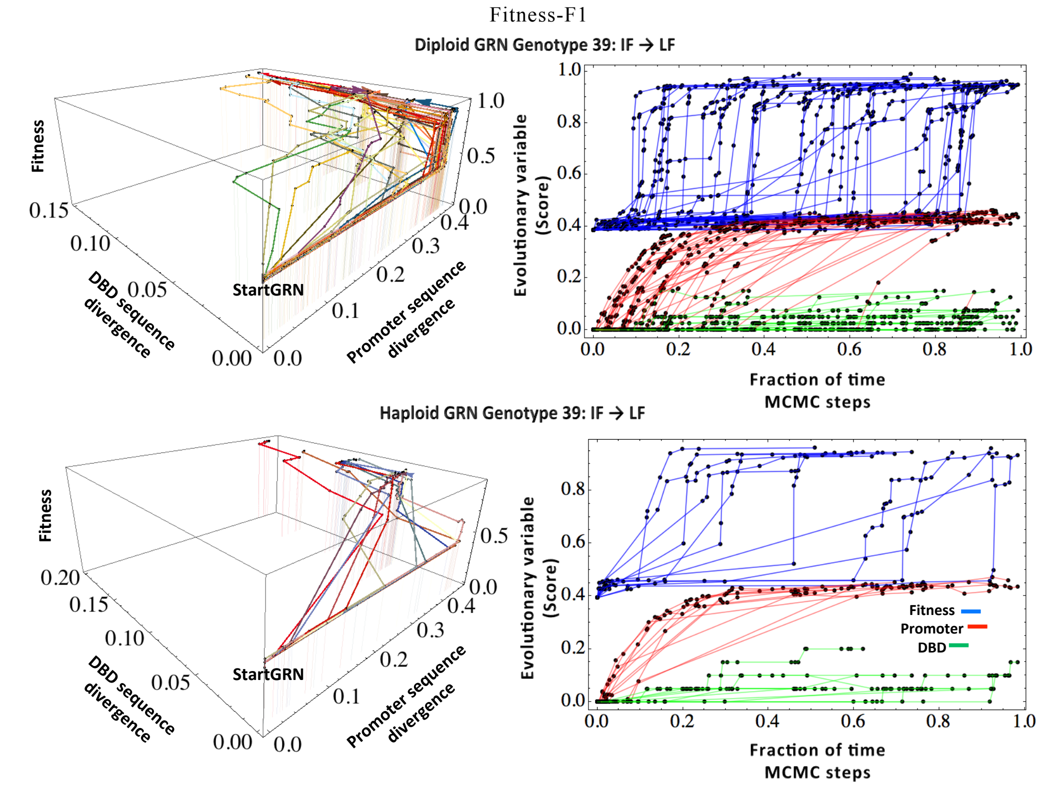


Figure C.18: Evolutionary phase space and sequence divergence plots: Fitness function FF2, GRN genotype 39 *Panels on the left depict the phase space of evolving GRNs. Coordinates shown represent the proportion of accumulated changes (with respect to start GRN configurations) in DBD encoding sequences and promoter regions (x, y coordinates), as well as their associated fitness score (z coordinate). Sequence divergence is assessed with respect to the start GRN configurations using a normalized Edit distance, which measures the percentage of dissimilarity (in terms of DNA sequence for promoters and of amino acids for DBDs) between a given ancestral sequence and mutant sequences sampled at a given time point (t_i) over the course of evolution. The concatenation of a sequence of triplets $(x(t_i), y(t_i), z(t_i))$ sampled at different time points over an evolution run describes a trajectory across the phase space. Triplets $(x(t_i), y(t_i), z(t_i))$ projected on the evolutionary phase space were sampled every time there was a fitness increment of $\Delta F \geq 0.001$ over the course of evolution toward a new optimum. Panels shown on the right illustrate the temporal sequence for individual components in the triplets $(x(t_i), y(t_i), z(t_i))$ recorded among different replicates. Time is depicted as the fraction of the total number of MCMC steps simulated. The y axis represents either percentage of dissimilarity between ancestral sequence and mutant sequences (for both promoter regions and DBD encoding sequences, separately) or the fitness value sampled at a given time point (t_i). Evolutionary simulations were performed using an implementation of the NEA (neutral-evolution-allowed) algorithm with FF2 (see subsection 4.1.6.2).*

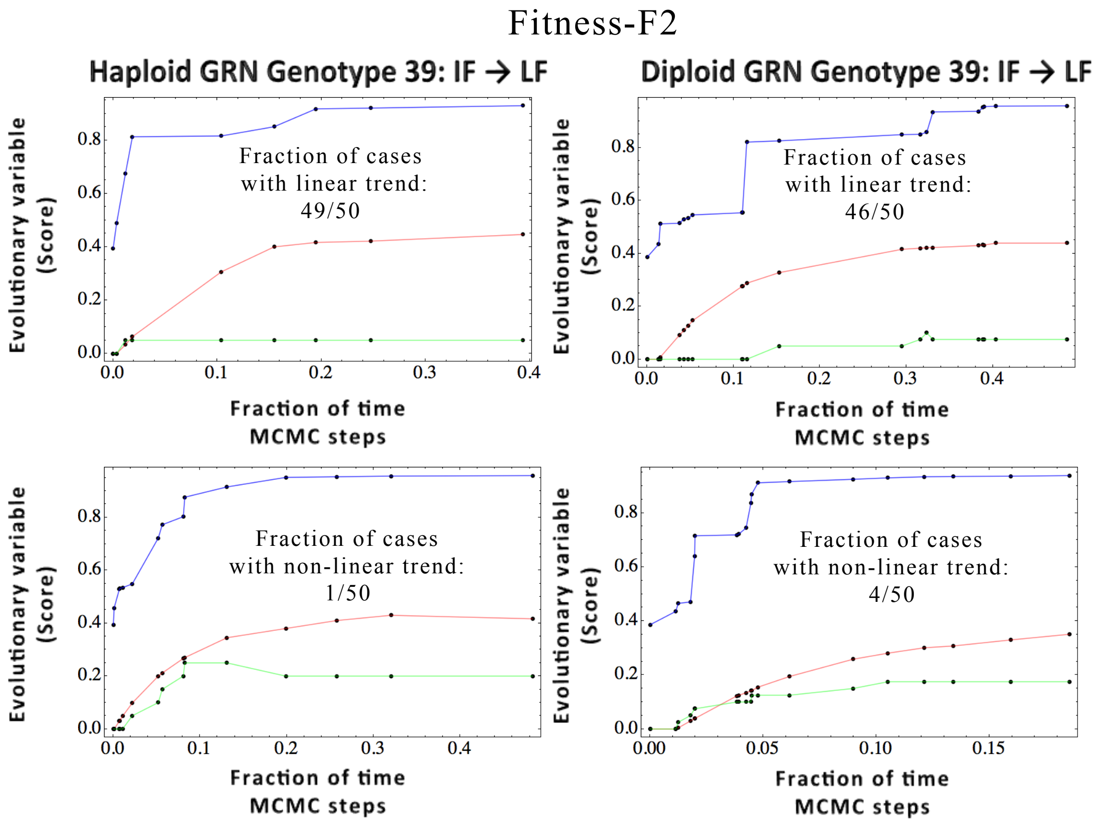


Figure C.19: Proportion of sequence divergence curves with linear vs. non-linear trends. *Temporal sequence for individual components in the triplets $(x(t_i), y(t_i), z(t_i))$ recorded among different replicates. Time is depicted as the fraction of the total number of MCMC steps simulated. The y axis represents either percentage of dissimilarity between ancestral sequence and mutant sequences (for both promoter regions and DBD encoding sequences, separately) or the fitness value sampled at a given time point (t_i). The proportions shown on each figure correspond to sequence divergence curves for DBDs. Evolutionary simulations were performed using an implementation of the NEA (neutral-evolution-allowed) algorithm with the FF1 (see subsection 4.1.6.2).*

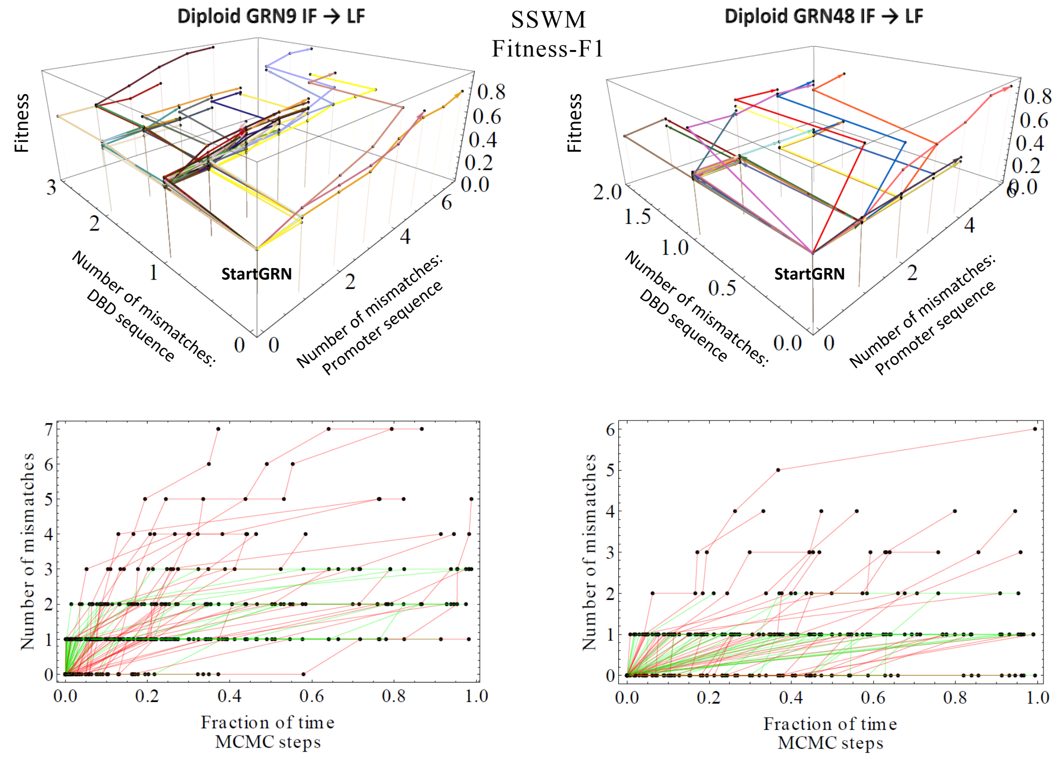


Figure C.20: Evolutionary phase space and sequence divergence plots: Evolution under the SSWM regime. *The panels shown on top of the figure depict the phase space of GRNs evolved under the SSWM regime using FF1 (see subsection 4.1.6.2), from start GRN configurations 9 and 48. Coordinates shown represent the proportion of accumulated changes (with respect to start GRN configurations) in DBD encoding sequences and promoter regions (x, y coordinates), as well as their associated fitness score (z coordinate). Sequence divergence is assessed with respect to the start GRN configurations using a normalized Edit distance, which measures the percentage of dissimilarity (in terms of DNA sequence for promoters and of amino acids for DBDs) between a given ancestral sequence and mutant sequences sampled at a given time point (t_i) over the course of evolution. The concatenation of a sequence of triplets $(x(t_i), y(t_i), z(t_i))$ sampled at different time points over an evolution run describes a trajectory across the phase space. Triplets $(x(t_i), y(t_i), z(t_i))$ projected on the evolutionary phase space were sampled every time there was a fitness increment of $\Delta F > 0$ over the course of evolution toward a new optimum. Panels shown at the bottom of the figure illustrate the temporal sequence for individual components in the triplets $(x(t_i), y(t_i), z(t_i))$ recorded among different replicates. Time is depicted as the fraction of the total number of MCMC steps simulated. The y axis represents either percentage of dissimilarity between ancestral sequence and mutant sequences (for both promoter regions and DBD encoding sequences, separately) or the fitness value sampled at a given time point (t_i).*

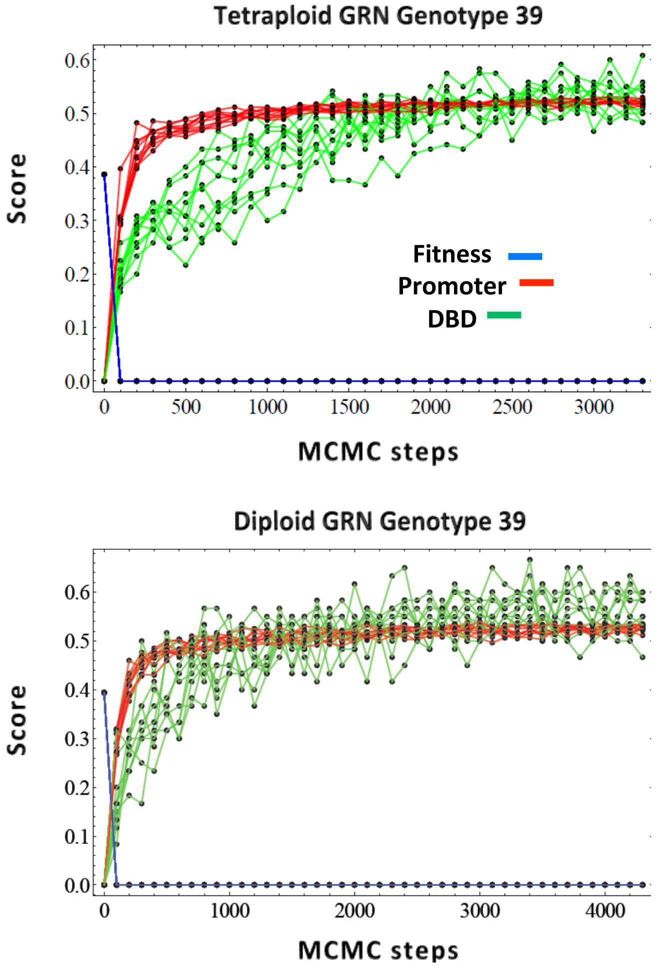


Figure C.21: Sequence divergence curves under relaxed selective constraints. *Divergence curves for promoter regions and DBD encoding sequences were generated by simulating random walks throughout sequence space under relaxed selective constraints (e.g. when the acceptance probability of any mutant genotype sampled at a given time point over the course of a walk through sequence space is 1, regardless of its phenotype). Time is depicted as the fraction of the total number of MCMC steps simulated. The score axis represents either percentage of dissimilarity between ancestral sequence and mutant sequences (for both promoter regions and DBD encoding sequences separately) or the fitness value sampled at a given time point (t_i). Sequence divergence is assessed using a normalized Edit distance.*

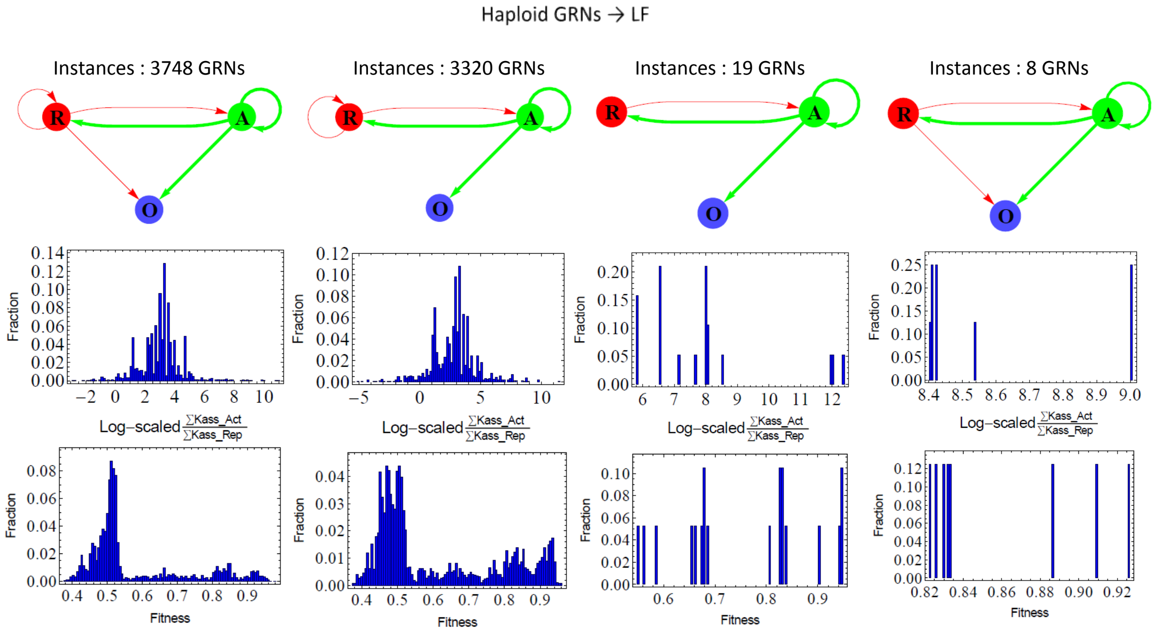


Figure C.22: Quantitative design features of haploid GRNs evolved toward an LF-type phenotypic optimum. *Type of regulatory wirings sampled over the course of evolution toward the LF-type phenotypic optimum. The thickness of the edges in the wirings is proportional to the aggregated DNA binding strength of a given TF overall possible binding sites on the promoter region of a target gene. Repressor and activator transcriptional regulators are shown in red and green, respectively; output genes are shown in blue. Distributions shown are for the ratio (in logarithmic scale) of aggregated DNA binding strength of activating TFs to the aggregated DNA binding strength of repressing TFs (middle panels), across all the solutions sampled over the course of evolution. The corresponding distributions of fitness scores are also shown (bottom panels).*

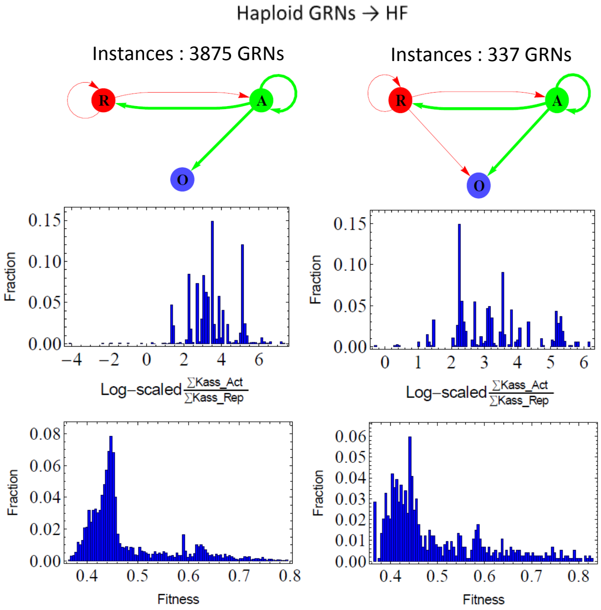


Figure C.23: Quantitative design features of haploid GRNs evolved toward an HF-type phenotypic optimum *Type of regulatory wirings sampled over the course of evolution toward the HF-type phenotypic optimum. The thickness of the edges in the wirings is proportional to the aggregated DNA binding strength of a given TF overall possible binding sites on the promoter region of a target gene. Repressor and activator transcriptional regulators are shown in red and green, respectively; output genes are shown in blue. Distributions shown are for the ratio (in logarithmic scale) of aggregated DNA binding strength of activating TFs to the aggregated DNA binding strength of repressing TFs (middle panels), across all the solutions sampled over the course of evolution. The corresponding distributions of fitness scores are also shown (bottom panels).*

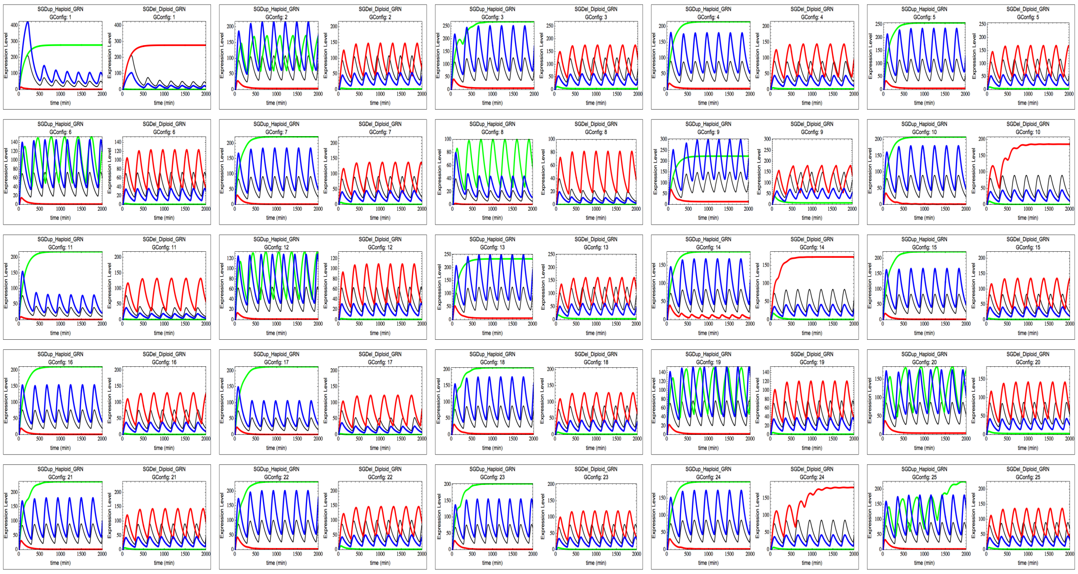


Figure C.24: Phenotypic responses of GRNs to dosage balance alteration: start GRNs set I. *Each framed panel in the figure depicts the phenotypic responses (dynamic expression patterns) of haploid (left-hand plot in each panel) and diploid (right-hand plot in each panel) GRNs to gene duplication and deletion, respectively. Phenotypic responses to duplication/deletion of activator (green color-coded profile), repressor (red color-coded profile) and output (blue color-coded profile) encoding genes are displayed in each plot. Each plot also displays the expression phenotype of the unperturbed system (black color-coded profile). The phenotypic readout of a GRN was always taken as the time varying concentration of the protein encoded by the downstream output gene. Phenotypic responses for start GRN configurations 1-25 are displayed in the figure.*

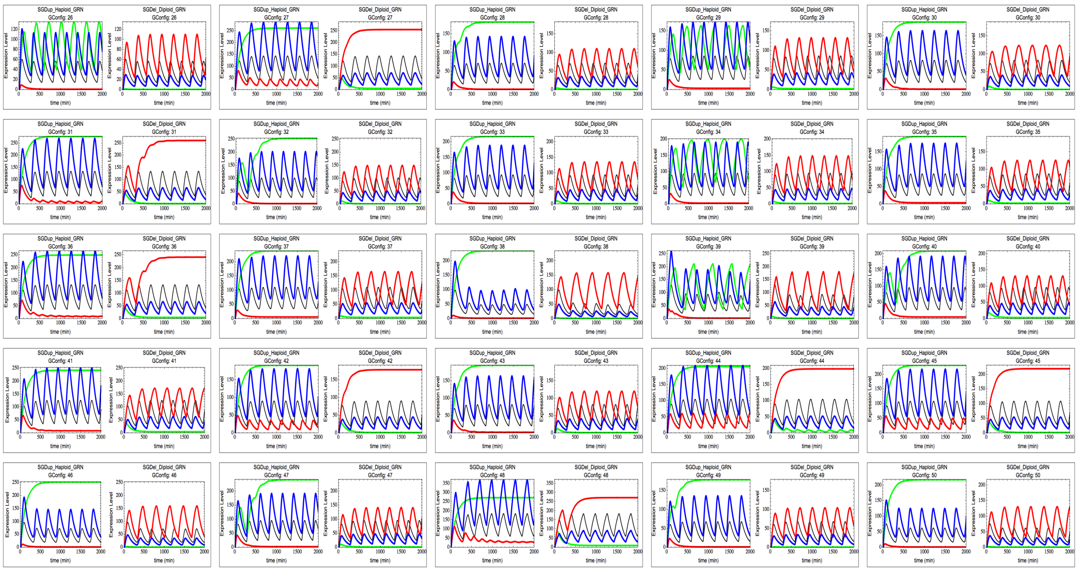


Figure C.25: Phenotypic responses of GRNs to dosage balance alteration: start GRNs set II. *Each framed panel in the figure depicts the phenotypic responses (dynamic expression patterns) of haploid (left-hand plot in each panel) and diploid (right-hand plot in each panel) GRNs to gene duplication and deletion, respectively. Phenotypic responses to duplication/deletion of activator (green color-coded profile), repressor (red color-coded profile) and output (blue color-coded profile) encoding genes are displayed in each plot. Each plot also displays the expression phenotype of the unperturbed system (black color-coded profile). The phenotypic readout of a GRN was always taken as the time varying concentration of the protein encoded by the downstream output gene. Phenotypic responses for start GRN configurations 26-50 are displayed in the figure.*

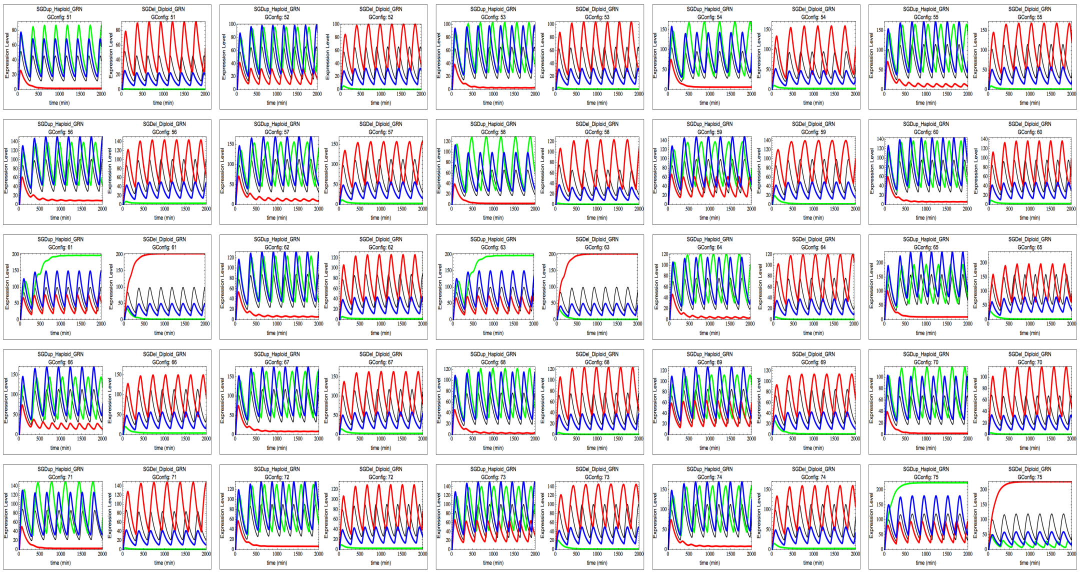


Figure C.26: Phenotypic responses of GRNs to dosage balance alteration: Ensemble I. *Each framed panel in the figure depicts the phenotypic responses (dynamic expression patterns) of haploid (left-hand plot in each panel) and diploid (right-hand plot in each panel) GRNs to gene duplication and deletion, respectively. Phenotypic responses to duplication/deletion of activator (green color-coded profile), repressor (red color-coded profile) and output (blue color-coded profile) encoding genes are displayed in each plot. Each plot also displays the expression phenotype of the unperturbed system (black color-coded profile). The phenotypic readout of a GRN was always taken as the time varying concentration of the protein encoded by the downstream output gene. Phenotypic responses for 25 randomly generated GRN configurations are displayed in the figure.*

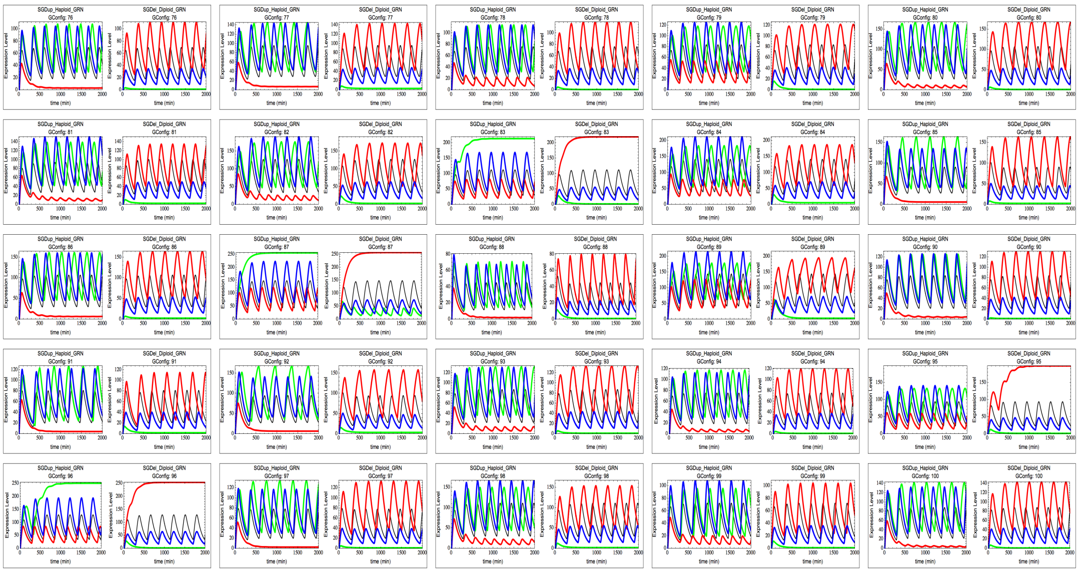


Figure C.27: Phenotypic responses of GRNs to dosage balance alteration: Ensemble II *Each framed panel in the figure depicts the phenotypic responses (dynamic expression patterns) of haploid (left-hand plot in each panel) and diploid (right-hand plot in each panel) GRNs to gene duplication and deletion, respectively. Phenotypic responses to duplication/deletion of activator (green color-coded profile), repressor (red color-coded profile) and output (blue color-coded profile) encoding genes are displayed in each plot. Each plot also displays the expression phenotype of the unperturbed system (black color-coded profile). The phenotypic readout of a GRN was always taken as the time varying concentration of the protein encoded by the downstream output gene. Phenotypic responses for another 25 randomly generated GRN configurations are displayed in the figure.*

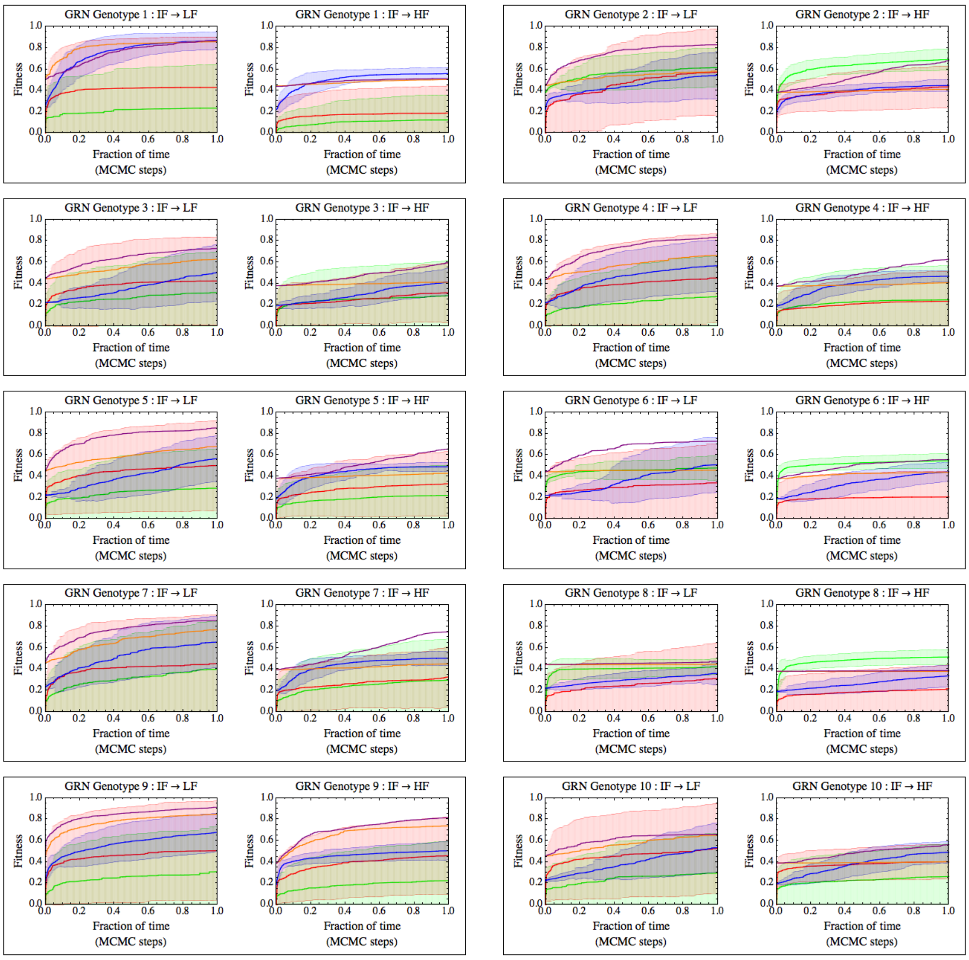


Figure C.28: Average fitness trajectories under dosage balance constraints: Fitness function FF1, set I. *Temporal sequences of fitness values recorded from 50 independent simulation replicates, using 10 different start GRN configurations, were averaged out to display the general trend of the adaptation process toward a new phenotypic optimum. Trajectories for evolving GRNs carrying an extra copy of the activator, repressor and output encoding genes are color-coded in green, red and blue, respectively. Bars along the trajectories represent standard deviations. For comparison purposes, the average fitness trajectories of the corresponding haploid (orange color-coded) and diploid (purple-color coded) GRN configuration are also displayed on the plots. Evolutionary simulations were performed using an implementation of the NEA (neutral-evolution-allowed) algorithm with the FF1 (see subsection 4.1.6.2. The number of MCMC steps simulated for GRNs carrying an extra gene copy were set as follows: GRNs carrying an extra copy of the activator or repressor gene were evolved for 890 (effective genome size) x 5 = 4450 MCMC steps; whereas GRNs carrying an extra copy of the output gene were evolved for 860 (effective genome size) x 5 = 4300 MCMC steps. The time (x) axis in the plots shown have been re-scaled to 1.*

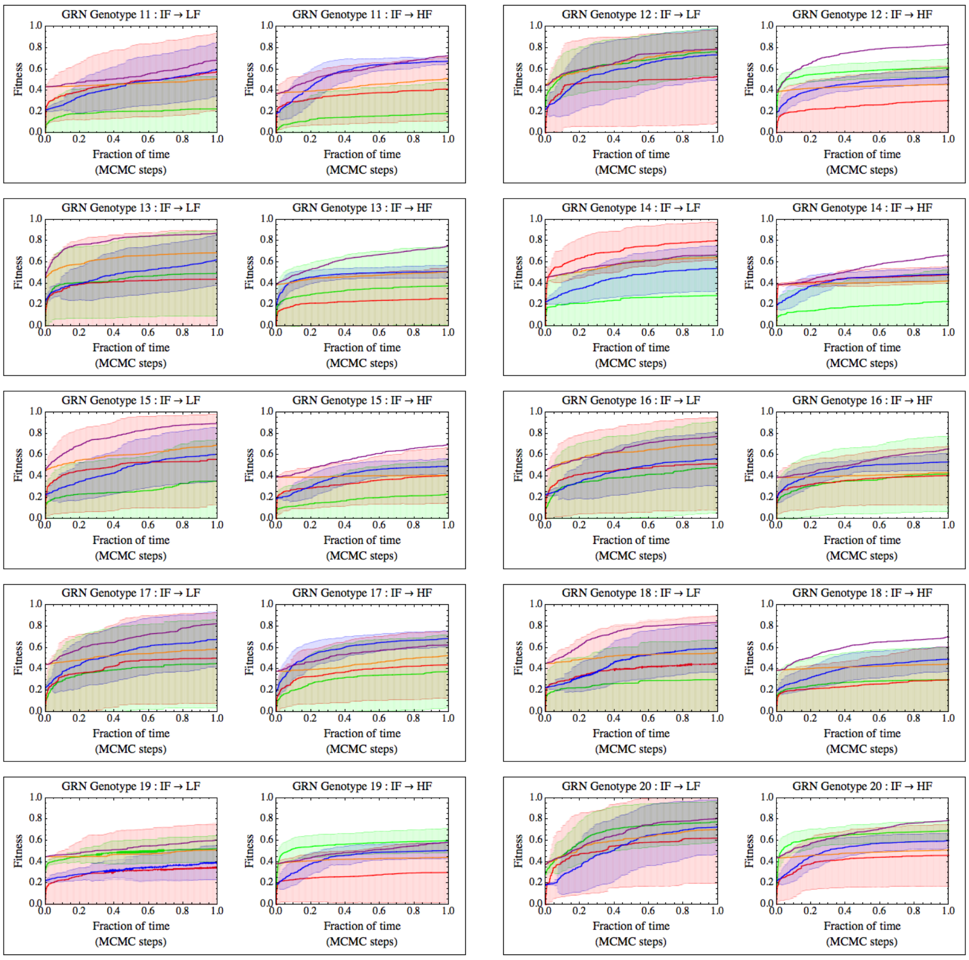
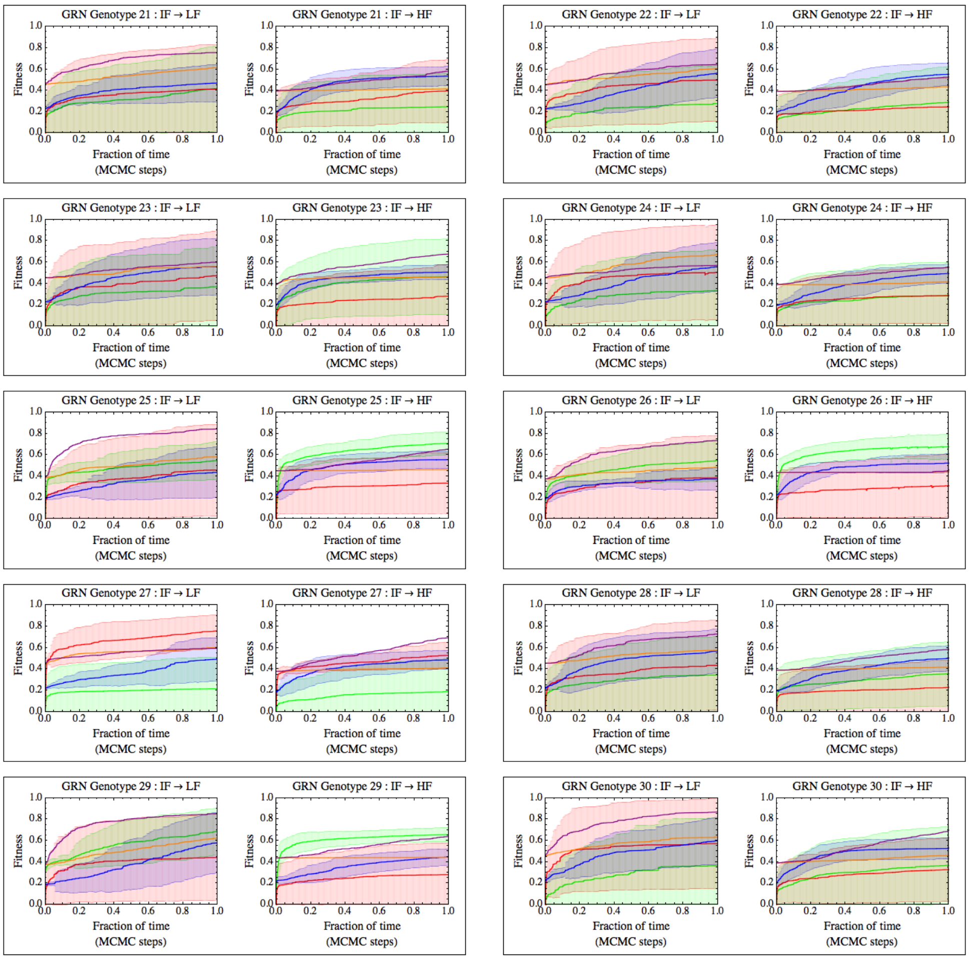


Figure C.29: Average fitness trajectories under dosage balance constraints: Fitness function FF1, set II. *Description is the same as in Figure C.28*



GRN Genotype 21 : IF → LF

GRN Genotype 21 : IF → HF

GRN Genotype 22 : IF → LF

GRN Genotype 22 : IF → HF

GRN Genotype 23 : IF → LF

GRN Genotype 23 : IF → HF

GRN Genotype 24 : IF → LF

GRN Genotype 24 : IF → HF

GRN Genotype 25 : IF → LF

GRN Genotype 25 : IF → HF

GRN Genotype 26 : IF → LF

GRN Genotype 26 : IF → HF

GRN Genotype 27 : IF → LF

GRN Genotype 27 : IF → HF

GRN Genotype 28 : IF → LF

GRN Genotype 28 : IF → HF

GRN Genotype 29 : IF → LF

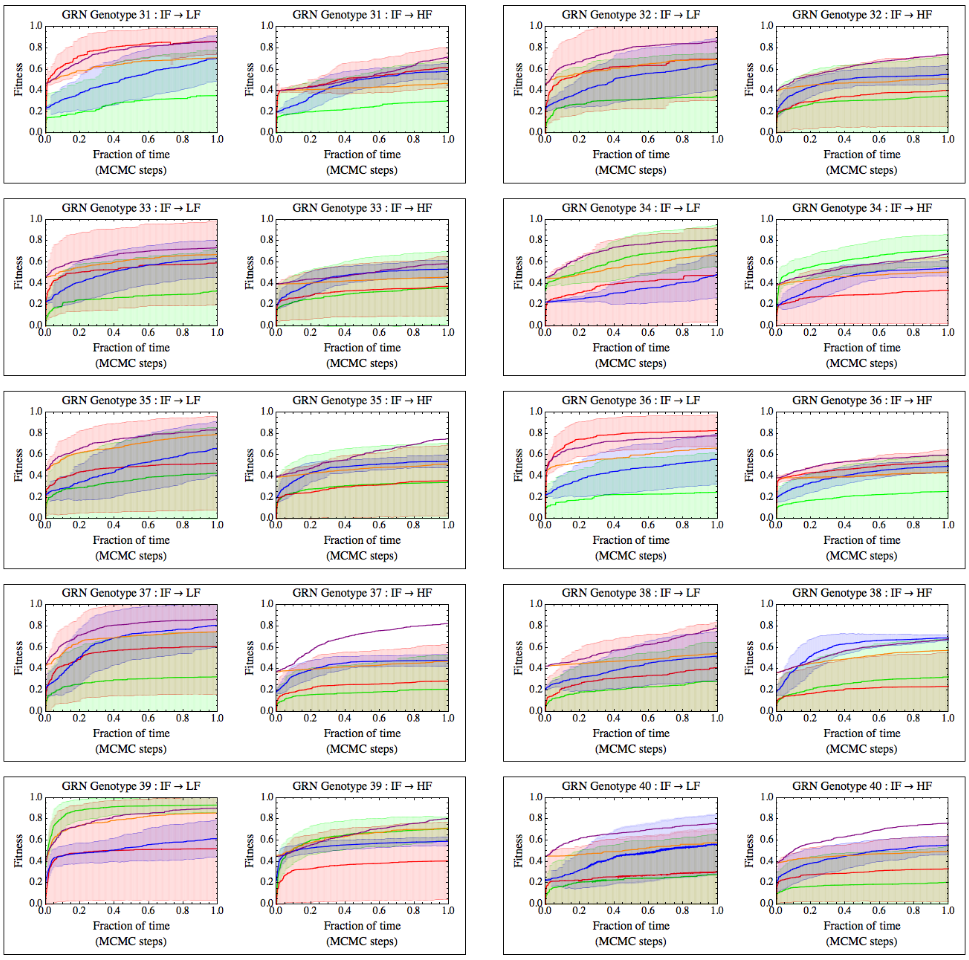
GRN Genotype 29 : IF → HF

GRN Genotype 30 : IF → LF

GRN Genotype 30 : IF → HF

Figure C.30: Average fitness trajectories under dosage balance constraints: Fitness function FF1, set III.

Description is the same as in Figure C.28



GRN Genotype 31 : IF → LF

GRN Genotype 31 : IF → HF

GRN Genotype 32 : IF → LF

GRN Genotype 32 : IF → HF

GRN Genotype 33 : IF → LF

GRN Genotype 33 : IF → HF

GRN Genotype 34 : IF → LF

GRN Genotype 34 : IF → HF

GRN Genotype 35 : IF → LF

GRN Genotype 35 : IF → HF

GRN Genotype 36 : IF → LF

GRN Genotype 36 : IF → HF

GRN Genotype 37 : IF → LF

GRN Genotype 37 : IF → HF

GRN Genotype 38 : IF → LF

GRN Genotype 38 : IF → HF

GRN Genotype 39 : IF → LF

GRN Genotype 39 : IF → HF

GRN Genotype 40 : IF → LF

GRN Genotype 40 : IF → HF

Figure C.31: Average fitness trajectories under dosage balance constraints: Fitness function FF1, set IV. *Description is the same as in Figure C.28*

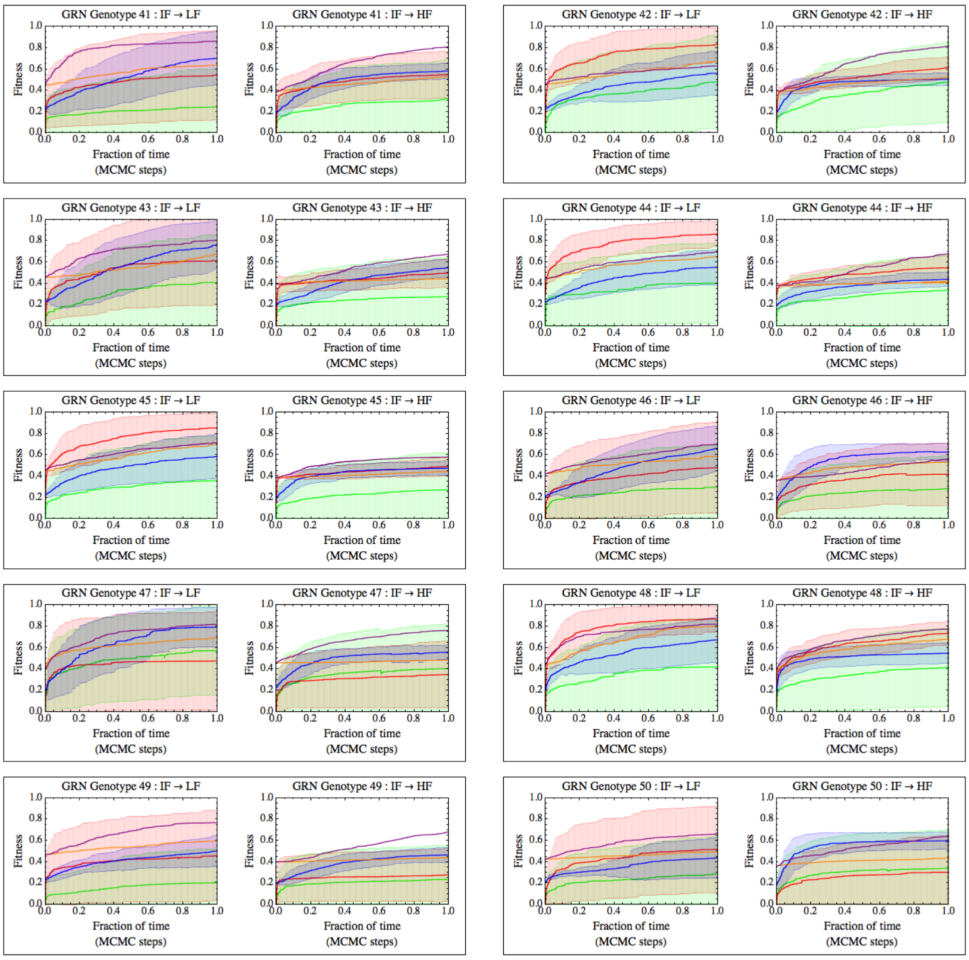


Figure C.32: Average fitness trajectories under dosage balance constraints: Fitness function FF1, set V. *Description is the same as in Figure C.28*

“I received the fundamentals of my education in school, but that was not enough. My real education, the superstructure, the details, the true architecture, I got out of the public library”

Isaac Asimov

D
Bibliography

Bibliography

[1] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.

[2] W.L Johannsen. *Arvelighedslrens elementer.* (The Elements of Heredity). Copenhagen., 1905.

[3] T. H. Morgan, Sturtevant H. J. M., and C. B. Bridges. *The Mechanism of Mendelian Heredity*. New York: Henry Holt, 1915.

[4] C. H. Waddington. Canalization of development and the inheritance of acquired characters. *Nature*, 150:563–565, 1942.

[5] J. Monod and F. Jacob. General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation. *Cold Spring Harbor Symposia on Quantitative Biology*, 26(0):389–401, 1961.

[6] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–467, 1969.

[7] G. Oster and P. Alberch. Evolution and Bifurcation of Developmental Programs. *Evolution*, 36(3):444, 1982.

[8] P. Alberch. From genes to phenotype: dynamical systems and evolvability. *Genetica*, 84(1):5–11, 1991.

[9] M. J. West-Eberhard. *Developmental plasticity and evolution.* Oxford University Press, New York, 2003.

[10] M. Pigliucci. Evolution of phenotypic plasticity: where are we going now? *Trends in Ecology & Evolution*, 20(9):481–486, 2005.

[11] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics*, 2(1):343–372, 2001.

[12] A. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2):101–113, 2004.

[13] K. Kaneko. *Life: An Introduction to Complex Systems Biology (Understanding Complex Systems)*. Springer-Verlag New York, Inc., September 2006.

[14] T. W. Grebe and J. Stock. Bacterial chemotaxis: The five sensors of a bacterium. *Current Biology*, 8(5):R154–R157, 1998.

[15] E. H. Davidson. Genomic Regulatory Systems In Development and Evolution. 2001.

[16] W. C. Spencer, G. Zeller, and J. D. Watson. A spatial and temporal map of C. elegans gene expression. *Genome Research*, 21(2):325–341, 2011.

[17] D. W. Knowles. Three-Dimensional Morphology and Gene Expression Mapping for the Drosophila Blastoderm. *Cold Spring Harbor Protocols*, 2012(2):150–161, 2012.

[18] B. Alberts. *Molecular Biology of the Cell*. Garland Science, 2008.

[19] O. Dahan, H. Gingold, and Y. Pilpel. Regulatory mechanisms and networks couple the different phases of gene expression. *Trends in Genetics*, 27(8):316–322, 2011.

[20] K. B. Singh. Transcriptional regulation in plants: the importance of combinatorial control. *Plant Physiology*, 118(4): 1111–1120, 1998.

[21] K Struhl. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, 98(1):1–4, 1999.

[22] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, 20(9):1377–1419, 2003.

[23] J. C. Perez and E. A. Groisman. Evolution of Transcriptional Regulatory Circuits in Bacteria. *Cell*, 138(2):233–244, 2009.

[24] D. J. Lee, S. D. Minchin, and S. J. W. Busby. Activating Transcription in Bacteria. *Annual Review of Microbiology*, 66(1): 125–152, 2012.

[25] H. Bolouri. Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9371–9376, 2003.

[26] S. Ben-Tabou de Leon and E. H. Davidson. Modeling the dynamics of transcriptional gene regulatory networks for animal development. *Developmental Biology*, 325(2):317–328, 2009.

[27] M. L. Howard and E. H. Davidson. cis-Regulatory control circuits in development. *Developmental Biology*, 271(1):109–118, 2004.

[28] L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell*, 122(6):947–956, 2005.

[29] L. Lopez-Maury, S. Marguerat, and B. Jurg. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, 9(8):583–593, 2008.

[30] S. Huang, , G. Eichler, Y. Bar-Yam, and D. E. Ingber. Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. *Physical Review Letters*, 94(12):128701, 2005.

[31] S B Carroll. *Endless Forms Most Beautiful: The New Science of Evo Devo*. W. W. Norton & Company, April 2006.

[32] E. H. Davidson. *The Regulatory Genome*. Gene Regulatory Networks In Development And Evolution. Academic Press, 2010.

[33] K. Kaufmann, A. Pajoro, and G. C. Angenent. Regulation of transcription in plants: mechanisms controlling developmental switches. *Nature Reviews Genetics*, 11(12):830–842, 2010.

[34] C. V. Kirchhamer, L. D. Bogarad, and E. H. Davidson. Developmental expression of synthetic cis-regulatory systems composed of spatial control elements from two differentgenes. *Proceedings of the National Academy of Sciences*, 93(24): 13849–13854, 1996.

[35] G. A. Wray and C. J. Lowe. Developmental Regulatory Genes and Echinoderm Evolution. *Systematic Biology*, 49(1):28–51, 2000.

[36] M. M. Kulkarni and D. N. Arnosti. Information display by transcriptional enhancers. *Development*, 130(26):6569–6575, 2003.

[37] P. Habets, A. F. Moorman, and V. M. Christoffels. Regulatory modules in the developing heart. *Cardiovascular Research*, 58(2):246–263, 2003.

[38] S. B. Carroll. Evolution at Two Levels: On Genes and Form. *PLoS Biology*, 3(7):e245, 2005.

[39] K. Kaufmann, J. M. Muiño, R. Jauregui, C. A. Airoldi, C. Smaczniak, P. Krajewski, and G. C. Angenent. Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. *PLoS Biology*, 7(4):e1000090, 2009.

[40] R. G. H. Immink, K. Kaufmann, and G. C. Angenent. The 'ABC' of MADS domain protein behaviour and interactions. *Seminars in Cell & Developmental Biology*, 21(1):87–93, 2010.

[41] F. Wellmer and J. L. Riechmann. Gene networks controlling the initiation of flower development. *Trends in Genetics*, 26 (12):519–527, 2010.

[42] S. B. Carroll. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*, 134(1):25–36, 2008.

[43] G. P. Wagner and V. J. Lynch. The gene regulatory logic of transcription factor evolution. *Trends in Ecology & Evolution*, 23(7):377–385, 2008.

[44] M. M. Kulkarni and D. N. Arnosti. cis-Regulatory Logic of Short-Range Transcriptional Repression in Drosophila melanogaster. *Molecular and Cellular Biology*, 25(9):3411–3420, 2005.

[45] S E Sultan. Phenotypic plasticity and plant adaptation. *Acta Botanica Neerlandica*, 44(4):363–383, 1995.

[46] W. R. Marcotte, S. H. Russell, and R. S. Quatrano. Abscisic acid-responsive sequences from the em gene of wheat. *The Plant Cell Online*, 1(10):969–976, 1989.

[47] J. Mundy, K. Yamaguchi-Shinozaki, and N. H. Chua. Nuclear proteins bind conserved elements in the abscisic acid-responsive promoter of a rice rab gene. *Proceedings of the National Academy of Sciences*, 87(4):1406–1410, 1990.

[48] C. Zou, K. Sun, J. D. Mackaluso, A. E. Seddon, R. Jin, M. F. Thomashow, and SH Shiu. Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 108(36):14992–14997, 2011.

[49] D. Wilson, V. Charoensawan, S. K. Kummerfeld, and S. A. Teichmann. DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, 36:D88–92, 2008.

[50] N. M. Luscombe, S. E. Austin, and H. M. Berman. An overview of the structures of protein-DNA complexes. *Genome Biol.*, 1(0), 2000.

[51] G. D Amoutzias, D. L. Robertson, Y. Van de Peer, and S. G. Oliver. Choose your partners: dimerization in eukaryotic transcription factors. *Trends in Biochemical Sciences*, 33(5):220–229, 2008.

[52] F. Messenguy and E. Dubois. Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. *Gene*, 316:1–21, 2003.

[53] K. Kaufmann and G. Melzer, R. Theissen. MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. *Gene*, 347(2):183–198, 2005.

[54] G. Gill and M. Ptashne. Mutants of GAL4 protein altered in an activation function. *Cell*, 51(1):121–126, 1987.

[55] S. Triezenberg. Structure and function of transcriptional activation domains. *Current Opinion in Genetics & Development*, 5(2):190–196, 1995.

[56] A. J. Stewart, S. Hannenhalli, and J. B. Plotkin. Why transcription factor binding sites are ten nucleotides long. *Genetics*, 192(3):973–85, 2012.

[57] U. Gerland, J. D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 99(19):12015–12020, 2002.

[58] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*, 15(2):116–124, 2005.

[59] M. Lässig. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC bioinformatics*, 8(Suppl 6):S7, 2007.

[60] M. D. Biggin. Animal transcription networks as highly connected, quantitative continua. *Dev Cell.*, 21(4):611–26, 2011.

[61] V. Mustonen, J. Kinney, C. G. Callan Jr., and M. Lassig. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc. Natl. Acad. Sci. USA*, 105(34):12376–12381, 2008.

[62] P. Dhaeseleer. What are DNA sequence motifs? *Nat. Biotechnol.*, 24(4):423–425, 2006.

[63] A. M. Sengupta, M. Djordjevic, and B. I. Shraiman. Specificity and robustness in transcription control networks. *Proc. Natl. Acad. Sci. USA*, 99(4):2072–2077, 2002.

[64] U. Gerland and T. Hwa. On the selection and evolution of regulatory DNA motifs. *J. Mol. Evol.*, 55:386–400, 2002.

[65] M. Babu, N. M. Luscombe, M. Nicholas, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283–291, 2004.

[66] N. Guelzim, S. Bottani, P. Bourgine, and Kepes F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31:60–63, 2002.

[67] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[68] T. I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, and I. Simon. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[69] R Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U Alon. Superfamilies of designed and evolved networks. *Science*, 303:1538–1542, 2004.

[70] U. Alon. Network motifs: theory and experimental approaches. *Nature reviews. Genetics*, 8(6):450–461, 2007.

[71] N. Yosef and A. Regev. Impulse Control: Temporal Dynamics in Gene Transcription. *Cell*, 144(6):886–896, 2011.

[72] E H Davidson. Gene regulatory networks and the evolution of animal body plans. *Science (New York, N.Y.)*, 311(5762): 796–800, 2006.

[73] M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–151, 2003.

[74] M. Levine and E. H. Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942, 2005.

[75] B. Prud'homme, N. Gompel, and S. B. Carroll. Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104 Suppl 1(suppl 1):8605–8612, 2007.

[76] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)*, 1975.

[77] J. Doebley and L. Lukens. Transcriptional regulators and the evolution of plant form. *The Plant Cell Online*, 10(7):1075–1082, 1998.

[78] J. R. Stone and G. A. Wray. Rapid evolution of cis-regulatory sequences via local point mutations. *Molecular Biology and Evolution*, 18(9):1764–1770, 2001.

[79] C. R. Landry, P. J. Wittkopp, C. H. Taubes, J. M. Ranz, A. G. Clark, and D. L. Hartl. Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. *Genetics*, 171(4):1813–1822, 2005.

[80] P. J. Wittkopp, B. K. Haerum, and A. G. Clark. Regulatory changes underlying expression differences within and between Drosophila species. *Nature genetics*, 40(3):346–350, 2008.

[81] H. E. Hoekstra and J. A. Coyne. The locus of evolution: evo devo and the genetics of adaptation. *International Journal of Organic Evolution*, 61(5):995–1016, 2007.

[82] V. J Lynch and G. P. Wagner. Resurrecting the role of transcription factor change in developmental evolution. *International Journal of Organic Evolution*, 62(9):2131–2154, 2008.

[83] I. Tirosh, S. Reikhav, A. A. Levy, and N. Barkai. A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation. *Science*, 324(5927):659–662, 2009.

[84] A. Tanay, A. Regev, and R. Shamir. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A*, 102(20):7203–8, 2005.

[85] B. B. Tuch, H. Li, and A. D. Johnson. Evolution of Eukaryotic Transcription Circuits. *Science (New York)*, 319(5871): 1797–1799, 2008.

[86] H. Li and A. D. Johnson. Evolution of Transcription Networks – Lessons from Yeasts. *Current Biology*, 20(17):R746–R753, 2010.

[87] D. A Thompson, S. Roy, M. Chan, M. P. Styczynsky, J. Pfiffner, C. French, A. Socha, A. Thielke, S. Napolitano, P. Muller, M. Kellis, J. H. Konieczka, I. Wapinski, A. Regev, and D. Tautz. Evolutionary principles of modular gene regulation in yeasts. *eLife*, 2(0):e00603–e00603, 2013.

[88] M. Lynch. The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics*, 8(10):803–813, 2007.

[89] S. Ohno. *Evolution by gene duplication*. Springer-Verlag, 1970.

[90] S. P Otto and J. Whitton. Polyploid incidence and evolution. *Annu Rev Genet.*, 34(1):401–437, 2003.

[91] J. S. Taylor and J. Raes. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.*, 38: 615–643, 2004.

[92] S. P. Otto. The Evolutionary Consequences of Polyploidy. *Cell*, 131(3):452–462, 2007.

[93] J. Zhang. Evolution by gene duplication: an update. *Trends Ecol. Evol.*, 18:192–198, 2003.

[94] C. L. Stebbins. *Variation and evolution in plants*. Columbia University Press, New York., 1950.

[95] W. Lewis. *Polyploidy in species populations*. In: Lewis W (ed.). Polyploidy: Biological Relevance. Plenum: New York, 1980.

[96] J. Ramsey and D. W. Schemske. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu Rev Ecol Syst.*, 29:467–501, 1998.

[97] D. Levin. *The Role of Chromosomal Change in Plant Evolution.* . Oxford University Press: New York, 2002.

[98] D. Gevers, K. Vandepoele, C. Simillion, and Y. Van de Peer. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol*, 12:148–54, 2004.

[99] M. Kellis, B.W. Birren, and E.S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–624, 2004.

[100] Y. Van de Peer, J. S. Taylor, and A. Meyer. Are all fishes ancient polyploids? *J Struct Funct Genom*, 3:65–73, 2003.

[101] J. P. Bogart. Evolutionary implications of polyploidy in amphibians and reptiles. *Basic Life Sci.*, 13, 1979.

[102] J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, and et al. Recent segmental duplications in the human genome. *Science*, 297, 2002.

[103] J F Wendel. Genome evolution in polyploids. *Plant Molecular Evolution*, 42(1):225–249, 2000.

[104] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science (New York)*, 290(5494): 1151–1155, 2000.

[105] S. De Bodt, S. Maere, and Y. Van de Peer. Genome duplication and the origin of angiosperms. *Trends Ecol Evol*, 20(11): 591–7, 2005.

[106] L. A. Meyers and D. A. Levin. On the abundance of polyploids in flowering plants. *Evolution*, 60(6):1198–1206, 2006.

[107] Y. Van de Peer, J. A. Fawcett, S. Proost, L. Sterck, and K. Vandepoele. The flowering world: a tale of duplications. *Trends Plant Sci*, 14(12):680–8, 2009.

[108] G. Blanc, K. Hokamp, and K. H. Wolfe. A recent polyploidy superimposed on older large-scale duplications in the arabidopsis genome. *Genome Res*, 13(2):137–44, 2003.

[109] C. Rizzon, L. Ponger, and Gaut B. S. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Comput Biol.*, 2(9):2:e115, 2006.

[110] S. Maere, S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15): 5454–5459, 2005.

[111] G. Blanc and K. H. Wolfe. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16(7):1667–78, 2004.

[112] C. Seoighea and C. Gehring. Genome duplication led to highly selective expansion of the arabidopsis thaliana proteome. *Trends Genet.*, 20:461464, 2004.

[113] C. L. Stebbins. *Processes of organic evolution*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[114] W. H. Wagner. Biosystematics and evolutionary noise. *Taxon*, 19:146–151, December 1970.

[115] C. L. Stebbins. *Chromosomal evolution in higher plants*. Addison- Wesley, Reading, 1971.

[116] R. J. Schultz. Role of polyploidy in the evolution of fishes. *Polyploidy: Biological Relevance: New York: Plenum*, 1980.

[117] Y. Van de Peer, S. Maere, and A. Meyer. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, 10(10):725–732, 2009.

[118] J. A. Fawcett, S. Maere, and Y. Van de Peer. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proceedings of the National Academy of Sciences*, 106(14):5737–5742, 2009.

[119] S. M. Shimeld and P. W. Holland. Vertebrate innovations. *Proc Natl Acad Sci USA*, (97):4449–4452, 2000.

[120] E.M. Mellgren and S.L. Johnson. The evolution of morphological complexity in zebrafish stripes. *Trends Genet*, (18): 128134, 2002.

[121] E.J. Stellwag. Are genome evolution, organism complexity and species diversity linked? *Integr Comp Biol.*, (44):358–365, 2004.

[122] M. Berenbrink, P. Koldkjaer, O. Kepp, and A. R. Cossins. Evolution of oxygen secretion in fishes and the emergence of a complex physiological system. *Science*, (307):17521757, 2005.

[123] J. A. Fawcett, Y. Van de Peer, and S. Maere. Significance and Biological Consequences of Polyploidization in Land Plant Evolution. *Plant Genome Diversity Volume 2*, (Chapter 17):277–293, 2013.

[124] S. P. Otto. The evolutionary consequences of polyploidy. *Cell*, 131(3):452–62, 2007.

[125] Z. Gu, L. M Steinmetz, X. Gu, C. Scharfe, R. W Davis, and WH. Li. Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421(6918):63–66, 2003.

[126] J. A. Birchler, H. Yao, S. Chudalayandi, D. Vaiman, and R. A. Veitia. Heterosis. *The Plant Cell Online*, 22(7):2105–2112, 2010.

[127] M. J. Hegarty and S. J. Hiscock. Genomic Clues to the Evolutionary Success of Polyploid Plants. *Current Biology*, 18: R435–R444, 2008.

[128] L. Comai. The advantages and disadvantages of being polyploid. *Nat Rev Genet*, 6(11):836–46, 2005.

[129] M. Lynch and B. Walsh. *The origins of genome architecture*. Sinauer Associates, 2007.

[130] P. S. Soltis and D. E. Soltis. The role of genetic and genomic attri- butes in the success of polyploids. *Proc. Natl. Acad. Sci. USA*, 97:7051–7057, 2000.

[131] D. R. Scannell, K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082):341–5, 2006.

[132] Matthew J Hegarty, Gary L Barker, Ian D Wilson, Richard J Abbott, Keith J Edwards, and Simon J Hiscock. Transcriptome Shock after Interspecific Hybridization in Senecio Is Ameliorated by Genome Duplication. *Current Biology*, 16(16):1652–1659, 2006.

[133] R A Rapp, Joshua A Udall, and J F Wendel. Genomic expression dominance in allopolyploids. *BMC biology*, 7(1):18, 2009.

[134] G. C Conant and K. H. Wolfe. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950, 2008.

[135] J B Walsh. How often do duplicated genes evolve new functions? *Genetics*, (139):421–428, 1995.

[136] M Kimura. The neutral theory of molecular evolution. 1983.

[137] S. Maere and Y. Van de Peer. *Duplicate Retention After Small- and Large-Scale Duplications*. Dittmar/Evolution After Gene Duplication. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2010.

[138] E B Lewis. *Pseudoallelism and gene evolution.* Cold Spring Harbor Symp Quant Biol., 1951.

[139] A. L. Hughes. The evolution of functionally novel proteins after gene duplication. *Proc R Soc Lond B Biol Sci*, 1994.

[140] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–45, 1999.

[141] V. J. Lynch, R. D. Leclerc, G. May, and G. P. Wagner. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics*, 43(11):1154–1159, 2011.

[142] H. Hendrickson, E.S. Slechta, U. Bergthorsson, D. I. Andersson, and J. R. Roth. Amplification-mutagenesis: evidence that directed adap- tive mutation and general hypermutability result from growth with a selected gene amplification. *Proc Natl Acad Sci USA*, (99):21642169, 2002.

[143] M.P. Francino. An adaptive radiation model for the origin of new gene functions. *Nat Genet*, (37):573577, 2005.

[144] A. L. Hughes. Gene duplication and the origin of novel proteins. *Proc. Natl Acad. Sci. USA*, 102:8791–8792, 2005.

[145] C. T. Hittinger and S. B. Carroll. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, 449(7163): 677–681, 2007.

[146] D. L. Des Marais and M. D. Rausher. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, (454):762–765, 2008.

[147] E. J. Chapman and M. Estelle. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proc Natl Acad Sci U S A*, 103:2730–2735, 2006.

[148] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, and et.al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418:387–391, 2002.

[149] R.S. Kamath, A.G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, B. N. Le, S. Moreno, M. Sohrmann, and et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, 421:231–237, 2003.

[150] A Wagner. Gene duplications, robustness and evolutionary innovations. *Bioessays*, 30(4):367–373, 2008.

[151] J. Brookfield. Can genes be truly redundant? *Curr Biol.*, 2:553–554, 1992.

[152] J. Cooke, M.A. Nowak, M. Boerlijst, and J. Maynard-Smith. Evolutionary origins and maintenance of redundant gene expression during metazoan development. *Trends Genet*, 13:360–364, 1997.

[153] M.A. Nowak, M.C. Boerlijst, J. Cooke, and J.M. Smith. Evolution of genetic redundancy. *Nature*, 388:167–171, 1997.

[154] G. C Conant and K. H. Wolfe. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Molecular Systems Biology*, 3(1), 2007.

[155] J. A. Birchler, U. Bhadra, M. P. Bhadra, and D. L. Auger. Dosage-Dependent Gene Regulation in Multicellular Eukaryotes: Implications for Dosage Compensation, Aneuploid Syndromes, and Quantitative Traits. *Developmental Biology*, 234(2): 275–288, 2001.

[156] R. A. Veitia. Gene Dosage Balance in Cellular Pathways Implications for Dominance and Gene Duplicability. *Genetics*, 168 (1):569–574, 2004.

[157] J. A. Birchler and R. A. Veitia. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, 109(37):14746–14753, 2012.

[158] R. A Veitia, S. Bottani, and Birchler J. A. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends in Genetics*, 24(8):390–397, 2008.

[159] J. A. Birchler and R. A. Veitia. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytologist*, 186(1):54–62, 2009.

[160] R A Veitia. Exploring the etiology of haploinsufficiency. *Bioessays*, 24(2):175–184, 2002.

[161] B. Papp, C. Pal, and L. D Hurst. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197, 2003.

[162] J. A. Birchler and D. L. Auger. Biological consequences of dosage dependent gene regulation in multicellular eukaryotes. *The Biology of Genetic Dominance*, 2004.

[163] W. Driever and C. Nusslein-Volhard. The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. *Cell*, 54:95–104, 1988.

[164] G. Struhl, K. Struhl, and P. M. Mcdonald. The gradient morphogen bicoid is a concentration-dependent transcriptional activator. *Cell*, 57:1259–1273, 1989.

[165] S. Di Talia, H. Wang, J. M. Skotheim, A. P. Rosebrock, B. Futcher, and F.R. Cross. Daughter- specific transcription factors regulate cell size control in budding yeast. *PLoS Biol.*, 7:e1000221, 2009.

[166] C.J. Zopf, K. Quinn, J. Zeidman, and N. Maheshri. Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Comput Biol.*, 9:e1003161., 2013.

[167] T. Galitski, A.J. Saldanha, C.A. Styles, E.S. Lander, and G.R. Fink. Ploidy regulation of gene expression. *Science*, 285: 251–254, 1999.

[168] M.B. Elowitz, A.J. Levine, E.D. Siggia, and P.S. Swain. Stochastic gene expression in a single cell. *Science*, 297:1183–1186, 2002.

[169] J.M. Pedraza and A. van Oudenaarden. Noise propagation in gene networks. *Science*, 307:1965–1969, 2005.

[170] J.A. Lee and J.R. Lupski. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*, 52:103121, 2006.

[171] J. Malone, D.Y. Cho, N. Mattiuzzo, C. Artieri, L. Jiang, R. Dale, H. Smith, J. McDaniel, S. Munro, M. Salit, J. Andrews, T. Przytycka, and B. Oliver. Mediation of drosophila autosomal dosage effects and compensation by network interactions. *Genome Biol.*, 13:R28, 2012.

[172] M. Acar, B. F. Pando, F. H. Arnold, M. B. Elowitz, and A. van Oudenaarden. A General Mechanism for Network-Dosage Compensation in Gene Circuits. *Science (New York)*, 329(5999):1656–1660, 2010.

[173] R. Song, , P. Liu, and M. Acar. Network-dosage compensation topologies as recurrent network motifs in natural gene networks. *BMC Syst Biol.*, 8(1):69, 2014.

[174] J. Yang, R. Lusk, and W. H. Li. Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci USA*, 100:15661–15665., 2003.

[175] Y. S. Lin, J. K. Hwang, and W. H. Li. Protein complexity, gene duplicability and gene dispensability in the yeast genome. *Gene*, 387:109–117, 2007.

[176] L. Li, , Y. Huang, X. Xia, and Z. Sun. Preferential duplication in the sparse part of yeast protein interaction network. *Mol Biol Evol.*, 23:2467–2473, 2006.

[177] M. Freeling. Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition. *Annu Rev Plant Biol.*, 60(1):433–453, 2009.

[178] M. Freeling. The evolutionary position of subfunctionalization, downgraded. *Genome Dyn*, 4:25–40, 2008.

[179] P. P. Edger and J. C. Pires. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research*, 17(5):699–717, 2009.

[180] S. Maere, S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A*, 102(15):5454–9, 2005.

[181] M. Prestel, C. Feller, and P. B. Becker. Dosage compensation and the global re-balancing of aneuploid genomes. *Genome Biol.*, 11(8):216, 2010.

[182] Z. J. Chen. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu Rev Plant Biol*, 58:377–406, 2007.

[183] F. M. Rosin and E. M. Kramer. Old dogs, new tricks: regulatory evolution in conserved genetic modules leads to novel morphologies in plants. *Dev Biol*, 332(1):25–35, 2009.

[184] K Geuten. Robustness and evolvability in the B-system of flower development. *Annals of botany*, 107(9):1545–1556, 2011.

[185] O. S. Soyer. *Evolutionary Systems Biology*. Springer, 2012.

[186] A. Wagner. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proceedings of the National Academy of Sciences of the United States of America*, 91(10):4387–4391, 1994.

[187] A. Wagner. Does Evolutionary Plasticity Evolve? *Evolution; international journal of organic evolution*, 50(3):1008, 1996.

[188] M. L. Siegal and A. Bergman. Waddington's canalization revisited: developmental stability and evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10528–10532, 2002.

[189] J. Masel. Genetic assimilation can occur in the absence of selection for the assimilating phenotype, suggesting a role for the canalization heuristic. *J. Evol. Biol.*, 17:1106–1110, 2004.

[190] J. Cotterell and J. Sharpe. An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. *Molecular Systems Biology*, 6(1), 2010.

[191] A. Spirov and D. Holloway. Using evolutionary computations to understand the design and evolution of gene and cell regulatory networks. *Methods*, 62(1):39–55, 2013.

[192] A. Bergman and M. L. Siegal. Evolutionary capacitance as a general feature of complex gene networks. *Nature*, 424(6948): 549–552, 2003.

[193] R. B. R. Azevedo, R. Lohaus, S. Srinivasan, K. K. Dang, and C. L. Burch. Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature*, 440(7080):87–90, 2006.

[194] R. Lohaus, C.L. Burch, and R.B.R. Azevedo. Genetic architecture and the evolution of sex. *J. Hered.*, 101:S142S157., 2010.

[195] O. C. Martin and A. Wagner. Effects of recombination on complex regulatory circuits. *Genetics*, 183:673684, 2009.

[196] J. Draghi and G. P. Wagner. The evolutionary dynamics of evolvability in a gene network model. *Journal of evolutionary biology*, 22(3):599–611, 2009.

[197] S. Ciliberti, O. C. Martin, and A. Wagner. Robustness Can Evolve Gradually in Complex Regulatory Gene Networks with Varying Topology. *PLoS Computational Biology*, 3(2):e15, 2007.

[198] E. Mjolsness, D. H. Sharp, and J. Reinitz. A connectionist model of development. *Journal of Theoretical Biology*, 152(4): 429–453, 1991.

[199] D. H Sharp and J. Reinitz. Prediction of mutant expression patterns using gene circuits. *BioSystems*, 47(1-2):79–90, 1998.

[200] J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K.N. Kozlov, E. Manu, Myasnikova, C. E. Vanario-Alonso, M. Samsonova, Sharp D. H., and J. Reinitz. Dynamic control of positional information in the early Drosophila embryo. *Nature*, 430(6997):368–371, 2004.

[201] I. Salazar-Ciudad, R. V. Solé, and S.A. Newman. Phenotypic and dynamical transitions in model genetic networks. I. Emergence of patterns and genotype-phenotype relationships. *Evol Dev.*, 3(2):84–94, 2001.

[202] I. Salazar-Ciudad, R. V. Solé, and S.A. Newman. Phenotypic and dynamical transitions in model genetic networks. II. Application to the evolution of segmentation mechanisms. *Evol Dev.*, 3(2):95–103, 2001.

[203] P. François, V. Hakim, and E. D Siggia. Deriving structure from evolution: metazoan segmentation. *Molecular Systems Biology*, 3(1), 2007.

[204] P. François and E. D. Siggia. Predicting embryonic patterning using mutual entropy fitness and in silico evolution. *Development*, 137(14):2385–2395, 2010.

[205] K. H. ten Tusscher and P. Hogeweg. The role of genome and gene regulatory network canalization in the evolution of multi-trait polymorphisms and sympatric speciation. *BMC Evolutionary Biology*, 9(1):159, 2009.

[206] K. H. ten Tusscher and P. Hogeweg. Evolution of Networks for Body Plan Patterning; Interplay of Modularity, Robustness and Evolvability. *PLoS Computational Biology*, 7(10):e1002208, 2011.

[207] J. A. Birchler, N. C. Riddle, D. L Auger, and R. A. Veitia. Dosage balance in gene regulation: biological implications. *Trends in Genetics*, 21(4):219–226, 2005.

[208] S. A. Frank. Population and quantitative genetics of regulatory networks. *Journal of Theoretical Biology*, 197(3):281–294, 1999.

[209] S. W. Omholt, E. Plahte, L. Øyehaug, and K. Xiang. Gene Regulatory Networks Generating the Phenomena of Additivity, Dominance and Epistasis. *Genetics*, 155(2):969–980, 2000.

[210] J. Peccoud, K. V. Velden, D. Podlich, C. Winkler, L. Arthur, and M. Cooper. The selective values of alleles in a molecular network model are context dependent. *Genetics*, 166(4):1715–1725, 2004.

[211] A. B. Gjuvsland, J. H. Ben, S. W. Omholt, and O. Carlborg. Statistical Epistasis Is a Generic Feature of Gene Regulatory Networks. *Genetics*, 175(1):411–420, 2007.

[212] H. Rajasingh, A. B. Gjuvsland, D. I. Våge, and S. W. Omholt. When parameters in dynamic models become phenotypes: a case study on flesh pigmentation in the chinook salmon (Oncorhynchus tshawytscha). *Genetics*, 179(2):1113–1118, 2008.

[213] A. B. Gjuvsland, J. O. Vik, D. A. Beard, P. J. Hunter, and S. W. Omholt. Bridging the genotype–phenotype gap: what does it take? *The Journal of Physiology*, 591(8):2055–2066, 2013.

[214] W. A. Lim. Designing customized cell signalling circuits. *Nature Reviews Molecular Cell Biology*, 11(6):393–403, 2010.

[215] J. B. Caleb, A. H. Andrew, G. P. Sergio, and A. L. Wendell. Rewiring Cells: Synthetic biology as a tool to interrogate the organizational principles of living systems. *Annual Review of Biophysics*, 39(1):515–537, 2010.

[216] W. A. Lim, C. M. Lee, and C. Tang. Design Principles of Regulatory Networks: Searching for the Molecular Algorithms of the Cell. *Molecular Cell*, 49(2):202–212, 2013.

[217] A. Crombach, M. A. Garcia-Solache, and J. Jaeger. Evolution of early development in dipterans: Reverse-engineering the gap gene network in the moth midge Clogmia albipunctata (Psychodidae). *Biosystems*, S0303-2647(14):00084–7, 2014.

[218] A. H. Chau, J. M. Walter, J. Gerardin, C. Tang, and W. A. Lim. Designing Synthetic Regulatory Networks Capable of Self-Organizing Cell Polarization. *Cell*, 151(2):320–332, 2012.

[219] S. Rosenfeld. Mathematical descriptions of biochemical networks: stability, stochasticity, evolution. *Prog Biophys Mol Biol.*, 106(2):400–9, 2012.

[220] E. Alm and A. P. Arkin. Biological networks. *Current Opinion in Structural Biology*, 13(2):193–202, April 2003.

[221] T Dobzhansky. Nothing in biology makes sense except in the light of evolution. *The american biology teacher*, 1973.

[222] T. J. Kawecki, R. E. Lenski, D. Ebert, B. Hollis, I. Olivieri, and M. C. Whitlock. Experimental evolution. *Trends in Ecology & Evolution*, 27(10):547–560, 2012.

[223] J. E. Barrick and R. E. Lenski. Genome dynamics during experimental evolution. *Nature Reviews Genetics*, 14(12):827–839, 2013.

[224] M. Lynch and B. Walsh. *Genetics and analysis of quantitative traits*. Sunderland, MA: Sinauer Associates, 1998.

[225] L. Loewe. A framework for evolutionary systems biology. *BMC systems biology*, 3(1):27, 2009.

[226] J. A. Draghi, T. L. Parsons, G. Wagner, and J. B. Plotkin. Mutational robustness can facilitate adaptation. *Nature*, 463(7279): 353–355, 2010.

[227] A. J. Stewart, T. L. Parsons, and J. B. Plotkin. Environmental robustness and the adaptability of populations. *Evolution*, 66 (5):1598–1612, 2012.

[228] B. C. Goodwin. Development and evolution. *Journal of Theoretical Biology*, 97(1):43–55, 1982.

[229] J. Jaeger, D. Irons, and N. Monk. The Inheritance of Process: A Dynamical Systems Approach. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 318(8):591–612, 2012.

[230] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the sixth international congress of genetics*, 1932.

[231] N. Kashtan, N. Noor, and U. Alon. Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13711–13716, 2007.

[232] E. S. Snitkin and D. Segrè. Epistatic Interaction Maps Relative to Multiple Metabolic Phenotypes. *PLoS Genetics*, 7(2): e1001294, 2011.

[233] S. Ciliberti, O. C. Martin, and A. Wagner. Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13591–13596, 2007.

[234] O. C. Martin and A. Wagner. Multifunctionality and Robustness Trade-Offs in Model Genetic Circuits. *Biophysical Journal*, 94(8):2927–2937, 2008.

[235] C. Espinosa-Soto and A. Wagner. Specialization Can Drive the Evolution of Modularity. *PLoS Computational Biology*, 6 (3):e1000719, 2010.

[236] J. Macía, R. V Solé, and S. F. Elena. The causes of epistasis in genetic networks. *Evolution; international journal of organic evolution*, 66(2):586–596, 2012.

[237] M. Inoue and K. Kaneko. Cooperative Adaptive Responses in Gene Regulatory Networks with Many Degrees of Freedom. *PLoS Computational Biology*, 9(4):e1003001, 2013.

[238] N. Kashtan and U. Alon. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13773–13778, 2005.

[239] J. L Payne and A. Wagner. Constraint and Contingency in Multifunctional Gene Regulatory Circuits. *PLoS Computational Biology*, 9(6):e1003071, 2013.

[240] J. Clune, J. B. Mouret, and H. Lipson. The evolutionary origins of modularity. *Proceedings of the Royal Society B: Biological Sciences*, 280(1755):20122863, 2013.

[241] P. François and V. Hakim. Design of genetic networks with specified functions by evolution in silico. *Proceedings of the National Academy of Sciences of the United States of America*, 101(2):580–585, 2004.

[242] G. Rodrigo, J. Carrera, and A. Jaramillo. Computational design of synthetic regulatory networks from a genetic library to characterize the designability of dynamical behaviors. *Nucleic acids research*, 39(20):e138–e138, 2011.

[243] J. Carrera, S. F. Elena, and A. Jaramillo. Computational design of genomic transcriptional networks with adaptation to varying environments. *Proceedings of the National Academy of Sciences of the United States of America*, 109(38):15277–15282, 2012.

[244] T. D. Cuypers and P. Hogeweg. Virtual genomes in flux: an interplay of neutrality and adaptability explains genome expansion and streamlining. *Genome Biology and Evolution*, 4(3):212–229, 2012.

[245] A. Warmflash, P. François, and E. D. Siggia. Pareto evolution of gene networks: an algorithm to optimize multiple fitness objectives. *Physical Biology*, 9(5):056001, 2012.

[246] J. Cotterell and J. Sharpe. Mechanistic Explanations for Restricted Evolutionary Paths That Emerge from Gene Regulatory Networks. *PloS one*, 8(4):e61178, 2013.

[247] K. Roh, F. R. P. Safaei, J. P. Hespanha, and S. R. Proulx. Evolution of transcription networks in response to temporal fluctuations. *Evolution; international journal of organic evolution*, 67(4):1091–1104, 2013.

[248] M. B. Cooper, M. Loose, and J. F. Y. Brookfield. Evolutionary modelling of feed forward loops in gene regulatory networks. *BioSystems*, 91(1):231–244, 2008.

[249] Bhavin S Khatri, Tom C B McLeish, and Richard P Sear. Statistical mechanics of convergent evolution in spatial patterning. *Proceedings of the National Academy of Sciences of the United States of America*, 106(24):9564–9569, June 2009.

[250] S. Dasmahapatra. Model of haplotype and phenotype in the evolution of a duplicated autoregulatory activator. *Journal of theoretical biology*, 325:83–102, 2013.

[251] K. Bullaughey. Changes in selective effects over time facilitate turnover of enhancer sequences. *Genetics*, 187(2):567–582, 2011.

[252] M. Pujato, T. MacCarthy, A. Fiser, and A. Bergman. The Underlying Molecular and Network Level Mechanisms in the Evolution of Robustness in Gene Regulatory Networks. *PLoS computational biology*, 9(1):e1002865, 2013.

[253] M. Andrecut, D. Cloud, and S. A. Kauffman. Monte Carlo simulation of a simple gene network yields new evolutionary insights. *Journal of Theoretical Biology*, 250(3):468–474, 2008.

[254] I. Tagkopoulos, YC. Liu, and S. Tavazoie. Predictive Behavior Within Microbial Genetic Networks. *Science*, 320(5881): 1313–1317, 2008.

[255] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.

[256] F. J. Poelwijk, D. J Kiviet, D. M. Weinreich, and S. J. Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386, 2007.

[257] R. Gordon. Evolution escapes rugged fitness landscapes by gene or genome doubling: The blessing of higher dimensionality. *Computers & Chemistry*, 18(3):325–331, 1994.

[258] Z. Burda, A. Krzywicki, O. C. Martin, and M. Zagorski. Motifs emerge from function in model gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 108(42):17263–17268, 2011.

[259] N. Kashtan, A. E. Mayo, T. Kalisky, and U. Alon. An Analytically Solvable Model for Rapid Evolution of Modular Structure. *PLoS computational biology*, 5(4):e1000355, 2009.

[260] M. Parter, N. Kashtan, and U. Alon. Facilitated Variation: How Evolution Learns from Past Environments To Generalize to New Environments. *PLoS computational biology*, 4(11):e1000206, November 2008.

[261] A. Crombach and P. Hogeweg. Evolution of Evolvability in Gene Regulatory Networks. *PLoS computational biology*, 4(7): e1000112, 2008.

[262] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(supp): C47–C52, 1999.

[263] N. Kashtan, M. Parter, E. Dekel, A. E. Mayo, and U. Alon. Extinctions in heterogeneous environments and the evolution of modularity. *Evolution; international journal of organic evolution*, 63(8):1964–1975, 2009.

[264] E. A. Variano and H. Lipson. Networks, Dynamics, and Modularity. *Physical Review Letters*, 92(18):188701, 2004.

[265] R. Calabretta, S. Nolfi, D. Parisi, and G. P. Wagner. Duplication of Modules Facilitates the Evolution of Functional Specialization. *Artif Life*, 6(1):69–84, 2006.

[266] G. Wagner, P. Mihaela, and J. M. Cheverud. The road to modularity. *Nature Reviews Genetics*, 8(12):921–931, 2007.

[267] M. Lynch. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 104 Suppl 1(Supplement 1):8597–8604, 2007.

[268] A. Wagner. Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators. *Biological Cybernetics*, 74(6):557–567, 1996.

[269] C. D. Meiklejohn and D. L. Hartl. A single mode of canalization. *Trends in Ecology & Evolution*, 17(10):468–473, 2002.

[270] P. D. Sniegowski and H. A. Murphy. Evolvability. *Current Biology*, 16, 2006.

[271] V. Mozhayskiy and I. Tagkopoulos. Guided evolution of in silico microbial populations in complex environments accelerates evolutionary rates through a step-wise adaptation. *BMC bioinformatics*, 13 Suppl 10(Suppl 10):S10, 2012.

[272] A. Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275 (1630):91–100, 2008.

[273] J. Masel and M. V. Trotter. Robustness and Evolvability. *Trends in Genetics*, 26(9):406–414, 2010.

[274] J. L. Payne and A. Wagner. The robustness and evolvability of transcription factor binding sites. *Science*, 343(6173): 875–877, 2014.

[275] M. Aldana, E. Balleza, S. Kauffman, and O. Resendiz. Robustness and evolvability in genetic regulatory networks. *Journal of Theoretical Biology*, 245(3):433–448, 2007.

[276] C. Torres-Sosa, S. Huang, and M. Aldana. Criticality Is an Emergent Property of Genetic Networks that Exhibit Evolvability. *PLoS Computational Biology*, 8(9):e1002669, 2012.

[277] C. F. Steiner. Environmental Noise, Genetic Diversity and the Evolution of Evolvability and Robustness in Model Gene Networks. *PloS One*, 7(12):e52204, 2012.

[278] D. A. Pechenick, J. H. Moore, and J. L. Payne. The influence of assortativity on the robustness and evolvability of gene regulatory networks upon gene birth. *Journal of Theoretical Biology*, 330:26–36, 2013.

[279] S. L Martin and B. C. Husband. Whole Genome Duplication Affects Evolvability of Flowering Time in an Autotetraploid Plant. *PloS one*, 7(9):e44784, 2012.

[280] S Ohno. *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag., 1970.

[281] J. Arjan G. M. de Visser, T. F. Cooper, and S. F. Elena. The causes of epistasis. *Proceedings of the Royal Society B: Biological Sciences*, 278(1725):3617–3624, 2011.

[282] M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538, 2012.

[283] F. J. Poelwijk, S. Tănase-Nicola, D. J. Kiviet, and S. J. Tans. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of Theoretical Biology*, 272(1):141–144, 2011.

[284] A. W. Covert III, R. E Lenski, C. O. Wilke, and C. Ofria. Experiments on the role of deleterious mutations as stepping stones in adaptive evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 110(34): E3171–E3178, 2013.

[285] D. M. Weinreich and L. Chao. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution; international journal of organic evolution*, 59(6):1175–1182, 2005.

[286] J. C. Chen and M. Conrad. A multilevel neuromolecular architecture that uses the extradimensional bypass principle to facilitate evolutionary learning. *Physica D: Nonlinear Phenomena*, 75(1-3):417–437, 1994.

[287] M. Conrad. Towards High Evolvability Dynamics. Introduction. *Evolutionary systems*, (Chapter 4):33–43, 1998.

[288] P. A. Cariani. Extradimensional bypass. *BioSystems*, 64(1-3):47–53, 2002.

[289] S. Gavrilets. Evolution and speciation on holey adaptive landscapes. *Trends in Ecology & Evolution*, 12(8):307–312, 1997.

[290] S. Gavrilets. A Dynamical Theory of Speciation on Holey Adaptive Landscapes. *The American Naturalist*, 154(1):1–22, 1999.

[291] F. J. Poelwijk, M. G. J. de Vos, and S. J. Tans. Tradeoffs and Optimality in the Evolution of Gene Regulation. *Cell*, 146(3): 462–470, 2011.

[292] M. Dragosits, V. Mozhayskiy, S. Quinones Soto, J. Park, and I. Tagkopoulos. Evolutionary potential, cross-stress behavior and the genetic basis of acquired stress resistance in Escherichia coli. *Molecular Systems Biology*, 9(1), 2013.

[293] D. L. Des Marais and M. D. Rausher. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, 454(7205):762–765, 2008.

[294] K. Voordeckers, C. A. Brown, K. Vanneste, E. van der Zande, A. Voet, S. Maere, and K. J. Verstrepen. Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. *PLoS Biology*, 10(12):e1001446, 2012.

[295] R. Huang, F. Hippauf, D. Rohrbeck, M. Haustein, K. Wenke, and J. Feike. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proceedings of the National Academy of Sciences*, 109(8):2966–2971, 2012.

[296] T. Sikosek, H. S. Chan, and E. Bornberg-Bauer. Escape from Adaptive Conflict follows from weak functional trade-offs and mutational robustness. *Proceedings of the National Academy of Sciences of the United States of America*, 109(37): 14888–14893, 2012.

[297] E. J Chapman and M. Estelle. Mechanism of Auxin-Regulated Gene Expression in Plants. *Annu Rev Genet.*, 43:265–85, 2009.

[298] T. Vernoux, G. Brunoud, E. Farcot, V. Morin, H. Van den Daele, J. Legrand, M. Oliva, P. Das, A. Larrieu, D. Wells, Y. Guédon, L. Armitage, F. Picard, S. Guyomarc'h, C. Cellier, G. Parry, R. Koumproglou, J. H. Doonan, M. Estelle, C. Godin, S. Kepinski, M. Bennett, L. De Veylder, and J. Traas. The auxin signalling network translates dynamic input into robust patterning at the shoot apex. *Molecular Systems Biology*, 7(1), 2011.

[299] C. Finet, A. Berne-Dedieu, C. P. Scutt, and F. Marlétaz. Evolution of the ARF Gene Family in Land Plants: Old Domains, New Tricks. *Molecular Biology and Evolution*, 30(1):45–56, 2013.

[300] T. J Guilfoyle and G. Hagen. Auxin response factors. *Current Opinion in Plant Biology*, 10(5):453–460, 2007.

[301] M. Sauer, Robert S., and Kleine-Vehn J. Auxin: simply complicated. *Journal of Experimental Botany*, 64(9):2565–2577, 2013.

[302] D. Inzé and L. De Veylder. Cell Cycle Regulation in Plant Development. *Annu Rev Genet.*, 40(1):77–105, 2006.

[303] L. De Veylder, T. Beeckman, and Dirk Inze. The ins and outs of the plant cell cycle. *Nature Reviews Molecular Cell Biology*, 8(8):655–665, 2007.

[304] S. Komaki and K. Sugimoto. Control of the Plant Cell Cycle by Developmental and Environmental Cues. *Plant and Cell Physiology*, 53(6):953–964, 2012.

[305] C. Smaczniak, R. G. H. Immink, G. C. Angenent, and K. Kaufmann. Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development*, 139(17):3081–3098, 2012.

[306] A. Becker. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Phylogenetics and Evolution*, 29(3):464–489, 2003.

[307] N. Pabón-Mora, B. A. Ambrose, and A. Litt. Poppy APETALA1/FRUITFULL Orthologs Control Flowering Time, Branching, Perianth Identity, and Fruit Development. *Plant physiology*, 158(4):1685–1704, 2012.

[308] L. M. Zahn, H. Kong, J. H. Leebens-Mack, S. Kim, P. S. Soltis, L. L. Landherr, D. E. Soltis, C. W. dePamphilis, and H. Ma. The Evolution of the SEPALLATA Subfamily of MADS-Box Genes A Preangiosperm Origin With Multiple Duplications Throughout Angiosperm History. *Genetics*, 169(4):2209–2223, 2005.

[309] A. Veron. Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. *Molecular Biology and Evolution*, 24(3):670–678, 2007.

[310] M. Mondragón-Palomino and G. Theißen. MADS about the evolution of orchid flowers. *Trends in Plant Science*, 13(2): 51–59, 2008.

[311] M. R. Khan, JY. Hu, S. Riss, C. He, and H. Saedler. MPF2-like-a MADS-box genes control the inflated Calyx syndrome in Withania (Solanaceae): roles of Darwinian selection. *Molecular Biology and Evolution*, 26(11):2463–2473, 2009.

[312] Y. Y. Chang, N. H. Kao, and CH. Yang. Characterization of the Possible Roles for B Class MADS Box Genes in Regulation of Perianth Formation in Orchid. *Plant physiology*, 152(2):837–853, 2010.

[313] M. Mondragón-Palomino and G. Theißen. Conserved differential expression of paralogous DEFICIENS- and GLOBOSA-like MADS-box genes in the flowers of Orchidaceae: refining the ‘orchid code’. *The Plant Journal*, 66(6):1008–1019, 2011.

[314] HL. Lee and V. F. Irish. Gene Duplication and Loss in a MADS Box Gene Transcription Factor Circuit. *Molecular Biology and Evolution*, 28(12):3367–3380, 2011.

[315] D. Vekemans, S. Proost, K. Vanneste, H. Coenen, T. Viaene, P. Ruelens, S. Maere, Y. Van de Peer, and K. Geuten. Gamma Paleohexaploidy in the Stem Lineage of Core Eudicots: Significance for MADS-Box Gene and Species Diversification. *Molecular Biology and Evolution*, 29(12):3793–3806, 2012.

[316] P. Ruelens, R. A. de Maagd, S. Proost, G. Theißen, K. Geuten, and K. Kaufmann. FLOWERING LOCUS C in monocots and the tandem origin of angiosperm-specific MADS-box genes. *Nature Communications*, 4, 2013.

[317] L. Vandesteene, L. López-Galvis, K. Vanneste, R. Feil, S. Maere, W. Lammens, F. Rolland, J. E. Lunn, N. Avonce, T. Beeckman, and P. Van Dijck. Expansive evolution of the trehalose-6-phosphate phosphatase gene family in Arabidopsis. *Plant physiology*, 160(2):884–896, 2012.

[318] Y. Jiao, N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, Y. Hu, H. Liang, P. S. Soltis, D. E. Soltis, S. W. Clifton, S. E. Schlarbaum, S. C. Schuster, H. Ma, J. Leebens-Mack, and C. W. dePamphilis. Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97–100, 2011.

[319] Y. Van de Peer, J. A Fawcett, S. Proost, L. Sterck, and K. Vandepoele. The flowering world: a tale of duplications. *Trends in Plant Science*, 14(12):680–688, 2009.

[320] D. E. Soltis. Polyploidy and angiosperm diversification. *American Journal of Botany*, 96(1):336–348, 2009.

[321] R. A. Veitia. Paralogs in Polyploids: One for All and All for One? *The Plant Cell Online*, 17(1):4–11, 2005.

[322] S. De Bodt, S. Maere, and Y. Van De Peer. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*, 20(11):591–597, 2005.

[323] M. Freeling and B.C. Thomas. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome research*, 16(7):805–814, 2006.

[324] K. Vanneste, S. Maere, and Y. Van de Peer. Tangled up in two: A burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Proc Roy Soc B-Biol Sci*, 5(369):1648, 2014.

[325] DY. Chao, B. Dilkes, H. Luo, A. Douglas, E. Yakubova, B. Lahner, and D. E. Salt. Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. *Science*, 341(6146):658–659, 2013.

[326] M. J. A. van Hoek and P. Hogeweg. Metabolic adaptation after whole genome duplication. *Molecular Biology and Evolution*, 26(11):2441–2453, 2009.

[327] M. W. Hahn. Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *The Journal of heredity*, 100(5):605–617, 2009.

[328] R. V. Solé and S. Valverde. Spontaneous emergence of modularity in cellular networks. *Journal of The Royal Society Interface*, 5(18):129–133, 2008.

[329] A. A. Neyfakh, N. N. Baranova, and L. J. Mizrokhi. A system for studying evolution of life-like virtual organisms. *Biol Direct*, 2006.

[330] M. E. Tsuda and M. Kawata. Evolution of Gene Regulatory Networks by Fluctuating Selection and Intrinsic Constraints. *PLoS Computational Biology*, 6(8):e1000873, 2010.

[331] E. H. Davidson. Evolutionary bioscience as regulatory systems biology. *Developmental Biology*, 357(1):35–40, 2011.

[332] F. J. Poelwijk, D. J. Kiviet, and S. J. Tans. Evolutionary Potential of a Duplicated Repressor-Operator Pair: Simulating Pathways Using Mutation Data. *PLoS Computational Biology*, 2(5):e58, 2006.

[333] H. Janssens, S. Hou, J. Jaeger, AR. Kim, E. Myasnikova, D. Sharp, and J. Reinitz. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. *Nature Genetics*, 38(10):1159–1165, October 2006.

[334] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451(7178):535–540, 2008.

[335] C. A. Martinez, K. A. Barr, AR. Kim, and J. Reinitz. A synthetic biology approach to the development of transcriptional regulatory models and custom enhancer design. *Methods*, 62(1):91–98, 2013.

[336] J. Ramsey. Polyploidy and ecological adaptation in wild yarrow. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7096–7101, 2011.

[337] R. Hermsen, S. Tans, and P. R. ten Wolde. Transcriptional Regulation by Competing Transcription Factor Modules. *PLoS Computational Biology*, 2(12):e164, 2006.

[338] J Reinitz, S Hou, and D H Sharp. Transcriptional Control in Drosophila. *Complexus*, 1(2):54–64, 2003.

[339] N. E. Buchler, U. Gerland, and Hwa T. On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):5136–5141, 2003.

[340] J. A Granek and N. D. Clarke. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biology*, 6(10):R87, 2005.

[341] Y. Mandel-Gutfreund and H. Margalit. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Research*, 26(10):2306–2312, 1998.

[342] E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, and E. Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, 30(6):521–530, 2012.

[343] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.

[344] R. Gutiérrez, G. C. MacIntosh, and P. J. Green. Current perspectives on mRNA stability in plants: multiple levels and mechanisms of control. *Trends in Plant Science*, 4(11):429–438, 1999.

[345] R. Narsai, K. A. Howell, A. H. Millar, N. O'Toole, I. Small, and J. Whelan. Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. *The Plant Cell Online*, 19(11):3418–3436, 2007.

[346] V. Iyer and K. Struhl. Absolute mRNA levels and transcriptional initiation rates in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 93(11):5208–5212, 1996.

[347] . 1997.

[348] T. Galitski, A. J. Saldanha, C. A. Styles, E. S. Lander, and G. R. Fink. Ploidy Regulation of Gene Expression. *Science (New York)*, 285(5425):251–254, 1999.

[349] CY. Wu, P. A. Rolfe, D. K. Gifford, and G. R. Fink. Control of Transcription by Cell Size. *PLoS Biology*, 8(11):e1000523, 2010.

[350] . *Science Signaling*, 4(176):ra38, 2011.

[351] A. V. Morozov, J.J. Havranek, D. Baker, and E. D. Siggia. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, 33((18)):5781–5798, 2005.

[352] X. He, M. A. Samee, C. Blatti, and S. Sinha. Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression. *PLoS Comput Biol.*, 6 (9):pii:e1000935, 2010.

[353] L. Bai and A. V. Morozov. Gene regulation by nucleosome positioning. *Trends Genet.*, 26((11)):476–483, 2010.

[354] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nat Genet.*, 31(1):69–73, 2002.

[355] J. M. Raser and E. K. O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, 2005.

[356] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna. Universally Sloppy Parameter Sensitivities in Systems Biology Models. *PLoS computational biology*, 3(10):e189, 2007.

[357] G. von Dassow, E. Meir, E. M. Munro, and G. M. Odell. The segment polarity network is a robust developmental module. *Nature*, 406(6792):188–192, 2000.

[358] E. Meir, G. von Dassow, E. Munro, and G. M. Odell. Robustness, Flexibility, and the Role of Lateral Inhibition in the Neurogenic Network. *Current Biology*, 12(10):778–786, 2002.

[359] T. Siggers and R. Gordân. Protein-DNA binding: complexities and multi-protein codes. *Nucleic acids research*, 42(4): 2099–2111, 2013.

[360] K. U. Winter, C. Weiser, K. Kaufmann, A. Bohne, C. Kirchner, A. Kanno, H. Saedler, and G. Theissen. Evolution of class B floral homeotic proteins: obligate heterodimerization originated from homodimerization.. *Mol Biol Evol.*, 19(5):587–96, 2002.

[361] P. M. Soltis, D. E. Soltis, S. Kim, A. Chanderbali, and M. Buzgo. Expression of floral regulators in basal angiosperms and the origin and evolution of ABC-function. *Adv Bot Res*, 44:483–506, 2006.

[362] N. E. Buchler, U. Gerland, and T. Hwa. Nonlinear protein degradation and the function of genetic circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9559–9564, July 2005.

[363] C. T. Harbison and et.al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[364] O. Purcell, N. J. Savery, C. S. Grierson, and M. di Bernardo. A comparative analysis of synthetic genetic oscillators. *Journal of The Royal Society Interface*, 7(52):1503–1524, 2010.

[365] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120, 1980.

[366] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol and Evol.*, 11(5):725–736, 1994.

[367] J. H. Gillespie. Molecular Evolution Over the Mutational Landscape. *International Journal of Organic Evolution*, 38(5): 1116, 1984.

[368] H. A. Orr. The population genetics of adaptation: The adaptation of DNA sequences. *Evolution; international journal of organic evolution*, 56(7):1317–1330, 2002.

[369] A. Wagner. The role of robustness in phenotypic adaptation and innovation. *Proceedings of the Royal Society B: Biological Sciences*, 279(1732):1249–1258, 2012.

[370] M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja. Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, 6(1):117, 2005.

[371] M Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–9, 1962.

[372] Y. Van de Peer, S. Maere, and A. Meyer. The evolutionary significance of ancient genome duplications. *Nat Rev Genet*, 10 (10):725–32, 2009.

[373] L. Cui, P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, J. E. Carlson, K. Arumuganathan, A. Barakat, V. A. Albert, H. Ma, and C. W. dePamphilis. Widespread genome duplications throughout the history of flowering plants. *Genome research*, 16(6):738–749, 2006.

[374] K. Vanneste, G. Baele, S. Maere, and Y. Van de Peer. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res.*, 24(8):1334–47, 2014.

[375] M. D. Purugganan, S. D. Rounsley, R. J. Schmidt, and M. F. Yanofsky. Molecular evolution of flower development: diversification of the plant MADS box regulatory gene family. *Genetics*, 140:345–356, 1995.

[376] E. M. Kramer, R. L. Dorit, and V. F. Irish. Molecular evolution of petal and stamen development: gene duplication and divergence within the APETALA3 and PISTILLATA MADS-box gene lineages. *Genetics*, 149:765–783, 1998.

[377] D. L. Remington, T. J. Vision, T. J. Guilfoyle, and J. W. Reed. Contrasting Modes of Diversification in the Aux/IAA and ARF Gene Families. *Plant Physiol.*, 135(3):1738–1752, 2004.

[378] S. A Teichmann and M. M. Babu. Gene regulatory network growth by duplication. *Nature genetics*, 36(5):492–496, 2004.

[379] A. Presser, M. B. Elowitz, M. Kellis, and R. Kishony. The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication. *Proceedings of the National Academy of Sciences*, 105(3):950–954, 2008.

[380] G. D. Amoutzias, D. L. Robertson, S. G. Oliver, and E. Bornberg-Bauer. Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO reports*, 5(3):274–279, 2004.

[381] E. Grotewold. Plant metabolic diversity: a regulatory perspective. *Trends in Plant Science*, 10(2):57–62, 2005.

[382] M. Bekaert, P. P. Edger, and J. C. Pires. Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints. *The Plant Cell*, 23(5):1719–28, 2011.

[383] R. De Smet and Y. Van de Peer. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology*, 15(2):168–176, 2012.

[384] D. A. Levin. Polyploidy and Novelty in Flowering Plants. *American Naturalist*, 122(1):1–25, 1983.

[385] N. Arrigo and M. S. Barker. Rarely successful polyploids and their legacy in plant genomes. *Current Opinion in Plant Biology*, 15(2):140–146, 2012.

[386] A. Madlung. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity*, 110(2):99–104, 2012.

[387] M. te Beest. The more the better? The role of polyploidy in facilitating plant invasions. *Annals of botany*, 109(1):19–45, 2012.

[388] A. Meyer and M. Schartl. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Current opinion in cell biology*, 11(6):699–704, 1999.

[389] K. D. Crow and P. W. Gunter. Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. What is the role of genome duplication in the evolution of complexity and diversity? *Molecular Biology and Evolution*, 23(5):887–892, 2006.

[390] M. Eric Schranz, S. Mohammadin, and P. P. Edger. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current Opinion in Plant Biology*, 15(2):147–153, 2012.

[391] M. Conrad. The geometry of evolution. *BioSystems*, 24(1):61–81, 1990.

[392] J. Gutiérrez and S. Maere. Modeling the evolution of molecular systems from a mechanistic perspective. *Trends in Plant Science*, 19(5):292–303, 2014.

[393] D. B. Kell. Genotype-phenotype mapping: genes as computer programs. *Trends in Genetics*, 18(11):555–559, 2002.

[394] P. N. Benfey. From genotype to phenotype: systems biology meets natural variation. *Science (New York)*, 320(5875):495–497, 2008.

[395] G. C. Conant. Rapid reorganization of the transcriptional regulatory network after genome duplication in yeast. *Proceedings of the Royal Society B: Biological Sciences*, 277(1683):869–876, 2010.

[396] J. R Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M Jacobs, B. Bolival, N. Assad-Garcia, J. I Glass, and M. W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.

[397] D. Heckmann, S. Schulze, A. Denton, U. Gowik, P. Westhoff, A. P. M. Weber, and M. J. Lercher. Predicting C4 Photosynthesis Evolution: Modular, Individually Adaptive Steps on a Mount Fuji Fitness Landscape. *Cell*, 153(7):1579–1588, 2013.

[398] R. Kassen and T. Bataillon. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature genetics*, 38(4):484–488, 2006.

[399] A. Eyre-Walker and P. D. Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8): 610–618, 2007.

[400] L. T. MacNeil and A. J. M. Walhout. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome research*, 21(5):645–657, 2011.

[401] S. A. Kauffman. *The Origins of Order*. Self-organization and Selection in Evolution. Oxford University Press, 1993.

[402] A. R. Fisher. The Genetical Theory Of Natural Selection: a Review. *J. Hered.*, 21:340–356, 1930.

[403] M. Eigen and R. Winkler-Oswatitsch. *Steps towards life: a perspective on evolution*. Oxford University Press, 1992.

[404] J. Franke, A. Klözer, J. A. de Visser, and J. Krug. Evolutionary Accessibility of Mutational Pathways. *PLoS Computational Biology*, 7(8):e1002134, 2011.

[405] A. E. Tsong, M. G. Miller, R. M. Raisner, and A. D. Johnson. Evolution of a Combinatorial Transcriptional Circuit. *Cell*, 115(4):389–399, 2003.

[406] H. H. McAdams, B. Srinivasan, and A. P. Arkin. The evolution of genetic regulatory systems in bacteria. *Nature Reviews Genetics*, 5(3):169–178, 2004.

[407] C. N. Henrichsen, E. Chaignat, and A. Reymond. Copy number variants, diseases and gene expression. *Human Molecular Genetics*, 18(R1):R1–8, 2009.

[408] YC. Tang and A. Amon. Gene Copy-Number Alterations: A Cost-Benefit Analysis. *Cell*, 152(3):394–405, 2013.

[409] J. R. Lupski. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics*, 14(10):417–422, 1998.

[410] K. Inoue and J. R. Lupski. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet*, 3(1):199–242, 2003.

[411] D. F. Conrad, T. D. Andrews, N. P. Carter, M. E. Hurles, and J. K. Pritchard. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics*, 38(1):75–81, 2006.

[412] G. H Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, N. P. Carter, C. Lee, and A. C. Stone. Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, 39(10):1256–1260, 2007.

[413] J. L. Santos, E. Saus, S. V. Smalley, L. R. Cataldo, G. Alberti, J. Parada, M. Gratacòs, and X. Estivill. Copy number polymorphism of the salivary amylase gene: implications in human nutrition research. *Journal of nutrigenetics and nutrigenomics*, 5(3):117–131, 2012.

[414] S. Koskiniemi, S. Sun, O. G. Berg, and D. I. Andersson. Selection-Driven Gene Loss in Bacteria. *PLoS Genetics*, 8(6): e1002787, 2012.

[415] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, and C. Beazley. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*, 315(5813):848–853, 2007.

[416] J. A. Klappenbach, J. M. Dunbar, and T. M. Schmidt. rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology*, 66(4):1328–1333, 2000.

[417] B. S. Stevenson, B. S. Stevenson, T. M. Schmidt, and T. M. Schmidt. Life History Implications of rRNA Gene Copy Number in Escherichia coli. *Applied and Environmental Microbiology*, 70(11):6670–6677, 2004.

[418] E. Chaignat, E. A. Yahya-Graison, C. N. Henrichsen, J. Chrast, F. Schütz, S. Pradervand, and A. Reymond. Copy number variation modifies expression time courses. *Genome research*, 21(1):106–113, 2011.

[419] R. Namba, T. M. Pazdera, R. L. Cerrone, and J. S. Minden. Drosophila embryonic pattern repair: how embryos respond to bicoid dosage alteration. *Development*, 124(7):1393–403, 1997.

[420] A. Wagner. Circuit topology and the evolution of robustness in two-gene circadian oscillators. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11775–11780, August 2005.

[421] C. A. Giurumescu, P. W. Sternberg, and A. R. Asthagiri. Predicting Phenotypic Diversity and the Underlying Quantitative Molecular Transitions. *PLoS Computational Biology*, 5(4):e1000354, 2009.

[422] R. A. Veitia. A sigmoidal transcriptional response: cooperativity, synergy and dosage effects. *Biological Reviews*, 78(1): 149–170, 2003.

[423] Y. Mileyko, R. I. Joh, and J. S. Weitz. Small-scale copy number variation and large-scale changes in gene expression. *Proceedings of the National Academy of Sciences*, 105(43):16659–16664, 2008.

[424] J. E. Baggs, T. S. Price, L. Di Tacchio, S. Panda, G. A. Fitzgerald, and J. B. Hogenesch. Network features of the mammalian circadian clock. *PLoS Biol.*, 7(3):e52, 2009.

[425] Richard Oberdorf and Tanja Kortemme. Complex topology rather than complex membership is a determinant of protein dosage sensitivity. *Molecular Systems Biology*, 5(1), 2009.

[426] K. Kaizu, H. Moriya, and H. Kitano. Fragilities Caused by Dosage Imbalance in Regulation of the Budding Yeast Cell Cycle. *PLoS Genetics*, 6(4):e1000919, 2010.

[427] R. A. Veitia. A generalized model of gene dosage and dominant negative effects in macromolecular complexes. *The FASEB Journal*, 24(4):994–1002, 2010.

[428] B. Bost and R. A. Veitia. Dominance and interloci interactions in transcriptional activation cascades: Models explaining compensatory mutations and inheritance patterns. *Bioessays*, 36(1):84–92, 2014.

[429] M. R. Atkinson, M. A. Savageau, J. T. Myers, and A. J. Ninfa. Development of Genetic Circuitry Exhibiting Toggle Switch or Oscillatory Behavior in Escherichia coli. *Cell*, 113(5):597–607, 2003.

[430] M. Tigges, T. T. Marquez-Lago, J. Stelling, and M. Fussenegger. A tunable synthetic mammalian oscillator. *Nature*, 457 (7227):309–312, 2009.

[431] Leonidas Bleris, Zhen Xie, David Glass, Asa Adadey, Eduardo Sontag, and Yaakov Benenson. Synthetic incoherent feed-forward circuits show adaptation to the amount of their genetic template. *Molecular Systems Biology*, 7(1):519–519, 2011.

[432] O. S. Soyer and C. J. Creevey. Duplicate retention in signalling proteins and constraints from network dynamics. *Journal of Evolutionary Biology*, 23(11):2410–2421, 2010.

[433] G. C. Conant, J. A. Birchler, and J. C. Pires. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol.*, 19:91–98, 2014.

[434] A. Munteanu, M. Constante, M. Isalan, and R. V. Solé. Avoiding transcription factor competition at promoter level increases the chances of obtaining oscillation. *BMC Syst Biol.*, 17:4–66, 2010.

[435] S. Dasmahapatra. Model of haplotype and phenotype in the evolution of a duplicated autoregulatory activator. *Journal of Theoretical Biology*, 325:83–102, 2013.

[436] H. Chen, L. Xu, and Z. Gu. Regulation dynamics of WGD genes during yeast metabolic oscillation. *Mol. Biol. Evol.*, 25 (12):2513–2516, 2008.

[437] K. Trachana, L. J. Jensen, and P. Bork. Evolution and regulation of cellular periodic processes: a role for paralogues. *EMBO reports*, 11(3):233–238, 2010.

[438] S. Shi, A. Hida, O. P. McGuinness, D. H. Wasserman, S. Yamazaki, and C. H. Johnson. Circadian clock gene Bmal1 is not essential; functional replacement with its paralog, Bmal2. *Curr Biol.*, 20(4):316–21, 2010.

[439] N. Pavelka, G. Rancati, and R. Li. Dr Jekyll and Mr Hyde: role of aneuploidy in cellular adaptation and cancer. *Current opinion in cell biology*, 22(6):809–815, 2010.

[440] B. Conrad and S. E. Antonarakis. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annual review of genomics and human genetics*, 8(1):17–35, 2007.

[441] H. A. Orr. The genetic theory of adaptation: a brief history. *Nat Rev Genet.*, 6(2):119–27, 2005.

[442] R. Kafri, A. Bar-Even, and Y. Pilpel. Transcription control reprogramming in genetic backup circuits. *Nature genetics*, 37 (3):295–299, 2005.

[443] T. Makino and A. McLysaght. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences*, 107(20):9270–9274, 2010.

[444] R. De Smet, K. L. Adams, K. Vandepoele, Marc C. E. Van M., S. Maere, and Y. Van de Peer. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110(8):2898–2903, 2013.

[445] R. A. Veitia, S. Bottani, and J. A. Birchler. Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. *Trends in Genetics*, 29(7):385–393, 2013.

[446] Y. Mileyko and J. S. Weitz. Bifurcation Analysis of Gene Regulatory Circuits Subject to Copy Number Variation. *SIAM Journal on Applied Dynamical Systems*, 9(3):799–826, 2010.

[447] J. Ihmels, S. Bergmann, M. Gerami-Nejad, I. Yanai, M. McClellan, J. Berman, and N. Barkai. Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science*, 309(5736):938–940, 2005.

[448] R. Sopko, D. Huang, N. Preston, G. Chua, B. Papp, K. Kafadar, M. Snyder, S. G. Oliver, M. Cyert, T. R Hughes, C. Boone, and B. Andrews. Mapping Pathways and Phenotypes by Systematic Gene Overexpression. *Molecular cell*, 21(3):319–330, 2006.

[449] Z. Ni, ED. Kim, M. Ha, E. Lackey, J. Liu, Y. Zhang, Q. Sun, and Z. J. Chen. Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature*, 457(7227):327–331, 2009.

[450] J K. Kim and D. B. Forger. A mechanism for robust circadian timekeeping via stoichiometric balance. *Molecular Systems Biology*, 8(1), 2012.

[451] S. G. Peisajovich. Evolutionary Synthetic Biology. *ACS Synthetic Biology*, 1(6):199–210, 2012.

[452] Y. Yokobayashi, R. Weiss, and F. H. Arnold. Directed evolution of a genetic circuit. *Proc Natl Acad Sci U S A*, 99(26): 16587–91, 2002.

[453] D. G. et.al Gibson. Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science*, 319 (5867):1215–20, 2008.

[454] J. S. Dymond and et.al. Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature*, 477(7365):471–476, 2011.

[455] J. W. Ellefson, A. J. Meyer, R. A. Hughes, J. R. Cannon, J. S. Brodbelt, and A. D. Ellington. Directed evolution of genetic parts and circuits by compartmentalized partnered replication. *Nat Biotechnol.*, 32(1):97–101, 2014.

[456] C. Pal, B. Papp, and G. Posfai. The dawn of evolutionary genome engineering. *Nat Rev Genet.*, 15(7):504–12, 2014.

[457] J. Wang, L. Tian, HS. Lee, N. E. Wei, H. Jiang, B. Watson, A. Madlung, T. C. Osborn, R. W. Doerge, L. Comai, and Z. J. Chen. Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. *Genetics*, 172(1):507–517, 2006.

[458] Z. J. Chen and Z. Ni. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays*, 28(3):240–252, 2006.

[459] Z. J. Chen. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. *Annual review of plant biology*, 58(1):377–406, 2007.

[460] J. J. Doyle, L. E. Flagel, A. H. Paterson, R. A. Rapp, D. E. Soltis, P. S. Soltis, and J. F. Wendel. Evolutionary Genetics of Genome Merger and Doubling in Plants. *Annu Rev Genet.*, 42(1):443–461, 2008.

[461] L. Comai, A. P. Tyagi, K. Winter, R. Holmes-Davis, Y. Reynolds, S. H. Stevens, and B. Byers. Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. *The Plant Cell Online*, 12(9):1551–1568, 2000.

[462] H. Shaked, K. Kashkush, H. Ozkan, M. Feldman, and A. A. Levy. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *The Plant Cell Online*, 13(8): 1749–1759, 2001.

[463] N. C. Riddle and J. A. Birchler. Effects of reunited diverged regulatory hierarchies in allopolyploids and species hybrids. *Trends in Genetics*, 19(11):597–600, 2003.

[464] P. S. Soltis and D. E. Soltis. The Role of Hybridization in Plant Speciation. *Annu Rev Plant Biol.*, 60(1):561–588, 2009.

[465] Z. J. Chen. Genomic and epigenetic insights into the molecular bases of heterosis. *Nature Reviews Genetics*, 14(7):471–482, 2013.

[466] F. Hochholdinger and N. Hoecker. Towards the molecular basis of heterosis. *Trends in Plant Science*, 12(9):427–432, 2007.

[467] X. Shi, D. WK. Ng, C. Zhang, L. Comai, W. Ye, and Z. J. Chen. Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. *Nature Communications*, 3:950, 2012.

[468] F. Mitelman. Recurrent chromosome aberrations in cancer. *Mutation Research*, 462:247–53, 2000.

[469] S. Frohling and H. Dohner. Chromosomal abnormalities in cancer. *New England Journal of Medicine*, 359:722–34, 2008.

[470] M. N. Raber and B. Barlogie. *DNA flow cytometry of human solid tumors. In: Melamed MR, Lindmo T MM, editors. Flow cytometry and sorting*. Wiley-Liss, New York, 1990.

[471] L. M. Merlo, L. S. Wang, J. W. Pepper, P. S. Rabinovitch, and C. C. Maley. Polyploidy, aneuploidy and the evolution of cancer. *Advances in Experimental Medicine and Biology*, 676:1–13, 2010.

[472] S. Huang. Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells. *Cancer Metastasis Rev.*, 32:423–448, 2013.

[473] E. Wang, J. Zou, N. Zaman, L. K. Beitel, M. Trifiro, and M. Paliouras. Cancer systems biology in the genome sequencing era: part 2, evolutionary dynamics of tumor clonal networks and drug resistance. *Semin Cancer Biol.*, 23(4):286–292, 2013.

[474] E. Wang, N. Zaman, , J. S. McGee, S. andMilanese, A. Masoudi-Nejad, and M. O'Connor-McCourt. Predictive genomics: A cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin Cancer Biol.*, S1044-579X(14):00050–9, 2014.

[475] Q. Cui, Yun Ma, M. Jaramillo, H. Bari, A. Awan, S. Yang, , S. Zhang, L. Liu, M. Lu, M. O'Connor-McCourt, E. O. Purisima, and E. Wang. A map of human cancer signaling. *Mol Syst Biol.*, 3:152, 2007.

[476] M. Cloutier and E. Wang. Dynamic modeling and analysis of cancer cellular network motifs. *Integr Biol (Camb).*, 3(7): 724–32, 2011.

[477] N. Eldredge and S. J. Gould. On punctuated equilibria. *Science*, 276(5311):338–341, 1997.

[478] H. H. Heng, J. B. Stevens, S. W. Bremer, G. Liu, B. Y. Abdallah, and C. J. Ye. Evolutionary mechanisms and diversity in cancer. *Advances in Cancer Research*, 112:217–253, 2011.