Theoretical analyses of the expansion of gene families implicated in dominant diseases

Giulia Malaguti

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Physique

École doctorale : « La physique, de la particule à la matière condensée »

réalisée

à l'Institut Curie

présentée par

Giulia MALAGUTI

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Analyses théoriques de l'expansion des familles de gènes impliqués dans des maladies dominantes

Theoretical analyses of the expansion of gene families implicated in dominant diseases

soutenue le 17 Octobre 2014

devant le jury composé de :

M^{me}	Christine DILLMANN	Rapporteur
M.	Thomas LENORMAND	Rapporteur
M.	Silvio FRANZ	Examinateur
M^{me}	Aleksandra WALCZAK	Examinateur
M.	Martin WEIGT	Examinateur
M.	Hervé ISAMBERT	Directeur de thèse

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Physique

École doctorale : « La physique, de la particule à la matière condensée »

réalisée

à l'Institut Curie

présentée par

Giulia MALAGUTI

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Analyses théoriques de l'expansion des familles de gènes impliqués dans des maladies dominantes

Theoretical analyses of the expansion of gene families implicated in dominant diseases

soutenue le 17 Octobre 2014

devant le jury composé de :

M^{me}	Christine DILLMANN	Rapporteur
M.	Thomas LENORMAND	Rapporteur
M.	Silvio FRANZ	Examinateur
M^{me}	Aleksandra WALCZAK	Examinateur
M.	Martin WEIGT	Examinateur
M.	Hervé ISAMBERT	Directeur de thèse

Analyses théoriques de l'expansion des familles de gènes impliqués dans des maladies dominantes

Les familles de gènes impliqués dans les cancers et autres maladies génétiques se sont beaucoup élargies *via* deux Duplications Globales de Génome (DGG) qui ont eu lieu à l'origine des vertébrés. La rétention des copies de ces gènes implique une susceptibilité plus grande aux maladies génétiques et constitue une énigme du point de vue de l'évolution. Dans cette thèse, nous avons généralisé des modèles classiques de génétique des populations pour révéler le mécanisme non-adaptatif qui a conduit à cette conservation de gènes potentiellement délétères chez les vertébrés. Nous avons résolu un modèle déterministe haploïde, nous avons étendu ce modèle à des génomes diploïdes et nous avons analysé les effets de taille finie des populations et de la sélection positive par une approche stochastique. Les résultats montrent, en accord avec les données génomiques du cancer chez l'homme, que les copies DGG susceptibles aux mutations délétères dominantes sont conservées indirectement *via* la sélection de purification dans les espèces post-DGG, qui présentent nécessairement une incompatibilité de ploïdie avec la population pre-DGG. Les résultats obtenus en étendant des méthodes avancées d'inférence bayésienne, quantifiant les effets causaux directs, soutiennent l'hypothèse d'une influence directe de la susceptibilité aux mutations délétères dominantes sur la rétention des copies DGG. Ces résultats révèlent le mécanisme d'évolution non-adaptatif responsable de la rétention de gènes DGG susceptibles aux mutations délétères dominantes et notre extension de méthodes d'inférence bayesienne ouvre la voie à la quantification des relations causales directes dans un large ensemble de problématiques.

Mots clés: modèles de génétique des populations; duplication globale de génome; spéciation ; mutations délétères dominantes; méthodes d'inférence bayesiénne; effets causaux directs.

Theoretical analyses of the expansion of gene families implicated in dominant diseases

Gene families implicated in cancers and other genetic diseases have been greatly expanded through two rounds of whole-genome duplication (WGD) that occurred at the onset of jawed vertebrates. However, such gene duplicates are expected to lead to an enhanced susceptibility to genetic diseases, and thus their retention represents an evolutionary puzzle from a natural selection perspective. In this thesis, we have expanded classical population genetics models to reveal the non-adaptive mechanism through which such potentially deleterious ohnologs (WGD-duplicated genes) were retained in the vertebrate genomes. We have solved a deterministic haploid model, we have considered extensions to diploid genotypes, and we have analyzed population size effects and the impact of positive selection through a stochastic approach. The results demonstrate, consistently with available human cancer genome data, that ohnologs prone to dominant deleterious mutations are indirectly selected through purifying selection in post-WGD species, arisen through the ploidy incompatibility between post-WGD individuals and the rest of the pre-WGD population. Extending advanced Bayesian inference methods to quantify direct and indirect causal effects, we have found further supporting evidences for the direct role of the gene susceptibility to deleterious mutations on ohnolog retention. Our findings rationalize the evolutionary mechanism responsible for the expansion of ohnologs prone to dominant deleterious mutations, highlighting the role of WGD-induced speciation. Our extension of Bayesian inference methods paves the way for the identification of direct causal relationships in a huge variety of problems.

Keywords: population genetics models; whole-genome duplication; speciation; dominant deleterious mutations; Bayesian inference methods; direct causal effects.

Contents

List of Figures

List of Tables

Abbreviations & Definitions

WGD	Whole Genome Duplication
Ohnolog	Gene retained from Whole Genome Duplications
SSD	Small Scale Duplication
MY	Million Years

Part I

Introduction

Chapter 1

Preamble

1.1 Thesis summary

Gene families prone to dominant deleterious mutations and implicated in cancers and other genetic diseases have been greatly expanded during the course of vertebrate evolution, unlike most other vertebrate genes without known deleterious mutations. In particular, their expansion can be traced back to two rounds of whole-genome duplication (WGD) that occurred at the onset of jawed vertebrates, some 500 MY ago. However, the duplication of these genes is expected to lead to an enhanced susceptibility to genetic diseases, and their retention represents an evolutionary puzzle from a natural selection perspective.

In order to rationalize this striking evolutionary outcome, we will model the long term evolution of gene families prone to dominant deleterious mutations, revealing the non-adaptive evolutionary mechanism that could have led to their surprising expansion. In particular, we will propose a consistent population genetics model to analyze the impact of the mode of duplication on the selection process of gene duplicates after WGD or small scale duplications (SSD). The retention of gene duplicates after WGD is related to their propensity to acquire dominant deleterious mutations. This is because WGD events, when successful, induce a speciation through the ploidy incompatibility between post-WGD individuals and the rest of the pre-WGD population. Such WGD-induced speciation then leads to the initial fixation of all gene duplicates and the long term retention of gene duplicates prone to dominant deleterious mutations through the indirect effect of purifying selection. In this context, we will solve a deterministic haploid model for the retention of gene duplicates, we will include extensions to diploid genotypes and analyze the fixation rates of gene duplicates through stochastic simulations, taking into account finite population size effects and the impact of positive selection. Finally, we will show that WGD duplicates prone to dominant deleterious mutations are indirectly selected through purifying selection in post-WGD species. The results will highlight the long-term evolutionary mechanism behind the surprising accumulation of WGD duplicates prone to

dominant deleterious mutations, consistently with cancer genome data on the prevalence of human oncogenes with WGD duplicates.

In order to further investigate the effect of the mode of duplication on the retention of gene duplicates susceptible to dominant deleterious mutations, we will consider other genomic properties that are correlated to the mode of duplication. These correlations suggest direct statistical associations among these properties, however many correlations do not result from direct effects in the underlying causal pathways but are in fact mediated by the indirect effect of third properties. To quantify direct and indirect effects, we will consider the Mediation Analysis, a method proposed by Pearl and typically used in social sciences and epidemiology. It is an advanced inference method to investigate causal pathways, specifically aiming at assessing the importance of a "mediator" Z in transmitting the indirect effect of a variable X on a response variable Y. In order to deal with complex causal graphs and analyze several correlated variables, we will extend the Mediation framework to the case of many intermediate variables between X and Y. The results will allow to bring further supporting evidences for the importance of the direct role of the susceptibility to dominant deleterious mutations, among other a priori possible genomic properties, on the retention of WGD duplicates and test alternative hypotheses proposed earlier to explain the retention of human genes coming from WGD.

Our findings rationalize the evolutionary mechanism responsible for the observed expansion of WGD duplicates prone to dominant deleterious mutations, and highlight the key roles of WGD-induced speciation and purifying selection after WGD events on the retention of gene duplicates. The direct influence of the gene susceptibility to dominant deleterious mutations on the retention of ohnologs is further supported by the application of our extension of the Mediation framework, which paves the way for the identification of direct causal relationships in various problems dealing with correlated data.

1.2 Organization of the thesis

This manuscript is organized in five parts. Part I introduces the subject of matter of this thesis, focusing on the mechanisms of formation of new genes by duplication and the evolutionary fate of duplicated genes. In particular, we will discuss the surprising expansion by WGD during vertebrate evolution of gene families susceptible to a specific kind of mutations, dominant deleterious mutations, which represents the source of inspiration of this study (Chapter 2).

Part II presents the two approaches used in this work. First, we will propose a consistent model in the context of population genetics to compare the evolutionary fate of SSD with WGD gene duplicates, taking into account their propensity to acquire dominant deleterious mutations (Chapter 3). Then, we will develop an extension of the Mediation Analysis framework to quantify direct and indirect causal effects in a network of multiple causally related variables (Chapter 4).

Part III contains the results obtained. First, we will present the predictions of our

population genetics model for the retention of SSD *vs* WGD duplicates susceptible to dominant deleterious mutations, and we will compare them with the reported retention of gene families with oncogenic properties in human (Chapter 5). Then, we will apply our extension of Pearl's Mediation Analysis to reach a global perspective, disentangling the direct from the indirect causal effects of multiple genomic properties on the retention of WGD duplicates in the human genome (Chapter 6).

Part IV contains general discussions and perspectives of this work, in relation to current theories and studies (Chapter 7).

Part V contains the details of the stochastic approach used to analyze the population genetics model introduced in Part II (Appendix A) and an overview of Pearl's theory of the *do*-calculus, the statistical framework at the basis of the Mediation Analysis (Appendix B).

1.3 Publications resulted/forthcoming from this thesis

(1) **G. Malaguti**, P.P. Singh and H. Isambert, "On the retention of gene duplicates prone to dominant deleterious mutations", *Theoretical Population Biology* (2014) 93: 38-51.

(2) P.P. Singh, S. Affeldt, **G. Malaguti** and H. Isambert, "Human dominant disease genes are enriched in paralogs originating from whole genome duplication", *PLoS Comput. Biol.* (2014) 10.7: e1003754.

(3) **G. Malaguti** and H. Isambert, "Extension of the Causal Mediation Analysis to partially directed acyclic graphs and its application to genomic data" [In Preparation].

Chapter 2

Evolution by gene duplication

The importance of gene duplication as the major evolutionary force for the creation of new genes has been primarily highlighted by Susumu Ohno [1] and Masatoshi Nei [2] in the late 1960's, despite limited early experimental observations and genomic information. The initial concepts and the theoretical framework of the evolution after gene duplication was laid down by Susumu Ohno in his seminal book, "Evolution by gene duplication" [3], and now gene duplication has been firmly established as the primary force of evolution.

In the introduction of this thesis we will discuss the main mechanisms of formation of new genes by duplication, the evolutionary fate of duplicated genes and the constraints underlying their retention or loss, considering the impact of the mode of duplication. We will focus, in particular, on whole-genome duplication events, highlighting their role during vertebrate evolution. We will also consider the case of genes susceptible to a specific kind of mutations, dominant deleterious mutations, whose surprising expansion by whole genome duplication during vertebrate evolution is the source of inspiration of this study.

2.1 Evolution through WGD and SSD

Different mechanisms can result in the duplication of regions of the genome, ranging from an individual gene to the entire genome. In particular, a special attention will be given to whole genome duplication events (WGD), genetic accidents that have been demonstrated to occur more frequently than traditionally expected and are thought to play a critical role during vertebrate evolution.

2.1.1 Mechanisms of gene duplication

The advent of genome sequencing has allowed to reveal the widespread occurrence of gene duplication events. Many genes in every sequenced eukaryotic genome have considerable sequence similarity and are clearly the products of gene duplication [4–11]. Gene du-

plicates can arise in many different ways [10,11], including unequal crossing over [1] [13,14], break-induced replication [2] and gene conversion [3] during the repair of broken chromosomes [18], slippage during recombination [4] [20], horizontal transfer [5] and other transpositions [6] [24–26], as well as temporary polyploidy [7] including an effective doubling of the whole genome [29]. Therefore, a gene duplication event can involve genomic regions ranging from a single gene, to large genomic segment and eventually to the entire genome.

Most of the above mechanisms generate duplicated regions ranging from a few base pairs to a large genomic segment, typically arranged in tandem [8]. Throughout this thesis, we will refer to them as small scale duplicates (SSD). By contrast, polyploidy can give rise to the duplication of the entire genome, so-called whole genome duplication (WGD). Such a genome duplication can be achieved by two mechanisms, auto- and allo-polyploidy [28]. Autopolyploidy, can occur by incomplete chromosome segregation, cytokinesis defects or fusion of two cells of the same organism during early development, leading to a polyploid embryo. In case of allopolyploidy, two cells from different but closely related organisms can fuse and give rise to an organism with whole genome duplication [28].

2.1.2 Frequent occurrence of WGD during evolution

Unlike SSD events, WGD events are evolutionary accidents providing the simultaneous duplication of the entire genome of an organism and their impact on evolution has been controversial for a long time.

A change in ploidy is traditionally expected to be deleterious and an evolutionary dead-end [29–31]. It is often argued that the evolutionary success of polyploids is hampered

1. Crossing over is the exchange of DNA between the two homologous chromosomes e.g. the maternal and paternal chromosome during meiosis. During this process, the homologous regions on the two aligned chromosomes break and then reconnect to create variations by double stranded breaks. However, if the chromosomes are misaligned, this may result in a duplication of the genomic segment on one chromosome and a deletion in the other [12].

2. If only one chromosome end at the break has homology with sequences elsewhere in the genome, then the defect could be repaired in the break-induced replication (BIR) pathway. In this case a single stranded tail can invade a homologous duplex DNA molecule and restart DNA replication at the replication fork, leading to duplication of one chromosome arm [15].

3. If both chromosome ends at the break have homology to sequences on an unbroken chromosome that can serve as a template, then repair may proceed by gene conversion. In this case, the damaged sequence is replaced with the homologous sequence such that the two sequences become identical after the conversion event [16,17].

4. Replication slippage is a mechanism involving slipped-strand mispairing by which the number of short, tandemly repeated sequences increases or decreases when DNA is replicated [19].

5. Horizontal gene transfer (HGT) is the physical transmission of DNA between different genomes, in a way other than reproduction; it is known to occur also between different species [21,22].

6. Transposition refers to the ability of genes to change position on chromosomes, a process in which a transposable element is removed from one site and inserted into a second site in the DNA, often resulting in its duplication [23].

7. Polyploidy is the presence of more than two paired sets of chromosomes in an organism [27,28].

8. Duplicated genes are directly adjacent to each other in the chromosome.

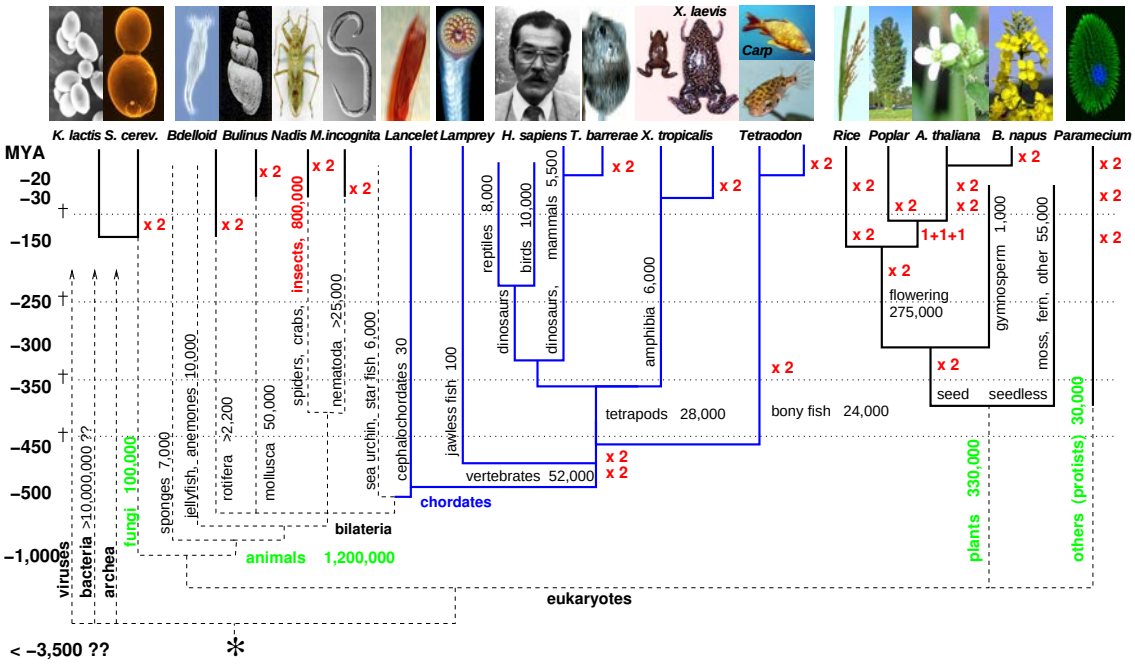

Figure 2.1: The occurrence of WGD events on the tree of life. The occurrence of WGD events during evolution is represented through the red symbol x2 (the symbol 1+1+1 represents a palaeo-hexaploid genome in early eudicots, that could be the result of a true hexaploidization event or successive genome duplications [33]). Polyploidy is widespread in plants, facilitated by selfing, asexuality, and perenniality, and it is rarer in animals [29]. However, two rounds of WGD occurred at the onset of jawed vertebrates, some 500 MY ago [34–36], creating the conditions for the evolution of vertebrate complexity and contributing to the evolutionary diversification in plants and animals.

by the inefficiency of selection when multiple alleles are present at each gene. Indeed, the spread of a favorable allele from a given frequency is slower at higher ploidy levels, because the selective effects of an allele are partially off-set by the presence of alternate alleles [30]. It was also believed that animals, unlike plants, should not tolerate polyploidy due to their usual mode of sexual reproduction [29, 31, 32], although asexual reproduction (such as parthenogenesis) also exists in animals.

By contrast, in the late 1960s, Susumu Ohno proposed that genome duplications are a significant mechanism of evolution even in the animal genomes [1, 3]. The increasing amount of genome sequences data and the state-of-the-art approaches to their analysis have now established polyploidy as a major evolutionary mechanism in all major eukaryotes — from unicellular eukaryotes, fungi, plants to animals (Fig.2.1). Polyploidy is especially common in plants: the common ancestor of all the extant angiosperms has undergone a tetraploidy event [37], and almost all major plant lineages have subsequently undergone multiple polyploidy events. Successive WGDs have also occurred in many animal genomes, as in annelids, flatworms, mollusks, insects, amphibians [29]. Most importantly, most vertebrates are now known to descend from a single lineage that experienced two consecutive

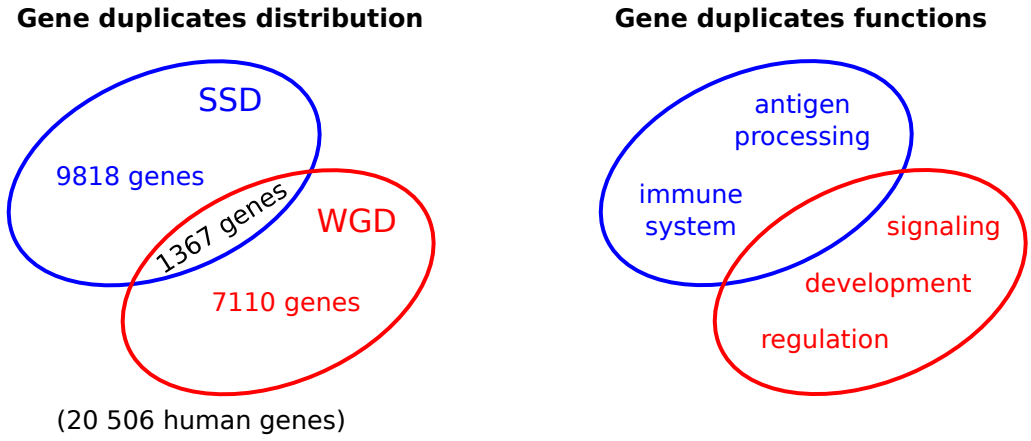

Figure 2.2: Antagonist retention patterns of SSD *vs* WGD duplicates. The figure highlights the antagonist patterns of retention during evolution of gene duplicates coming from SSD and WGD events in the human genome. WGD duplicates tend to be *not* duplicated through SSD and *vice versa* [40] (left). WGD duplicates have been preferentially retained in specific gene classes associated with organismal complexity (development, signal, regulation) while the retention of SSD duplicates is related to different functional categories (antigen processing, immune system) [41,42] (right).

WGDs soon after the divergence from other chordates about 500 MY ago [34–36] (this is the long debated "2R hypothesis" [1, 3]). Similarly, all bony fishes, which make up about 90% of extant fishes, are now known to derive from a single species that doubled its genome about 300 MY ago (*i.e.* the "3R hypothesis" [38, 39]).

Although, in the short term, polyploidy leads to a population bottleneck (related to the obligate speciation owing to the difference in ploidy between pre- and post-WGD individuals) and possible competition with their diploid ancestors, the frequent occurrence of WGD events and the success of polyploid organisms strongly suggest that whole genome duplication is a dynamic process that has contributed to the evolutionary diversification in plants and animals [29].

2.1.3 Antagonist retention pattern of SSD and WGD duplicates

Most genes belong to gene families which have undergone consecutive gene duplication events [11]. However, a duplication event is usually followed by the loss of the duplicated genes through non-functionalization (see Sec.2.2). In particular, in the case of the 2R WGD events, the majority of the resulting gene duplicates are subsequently lost. Nevertheless, the analysis of the few duplicated genes retained in the genome discloses an interesting retention pattern related to the mode of duplication.

Recent studies have revealed that SSD and WGD duplicates have been retained during evolution following antagonist patterns (Fig. 2.2). Evidence has accumulated that WGD duplicates have been preferentially retained in specific functional gene classes associated with higher organismal complexity, such as signaling pathways, transcription networks,

and developmental genes (for example, nervous system, morphogenesis) [42–46]. This is the case of the basic set of genes involved in development and signaling that was already present in chordates, but WGD events resulted in the specific expansion of these gene families in vertebrates, leading to the evolution of the neural crest [36], the vertebrate skeleton [47] and brain structures [36]. By contrast, gene duplicates coming from SSD are strongly biased toward different functional categories, such as antigen processing, immune response, and metabolism [42]. SSD and WGD duplicates also differ in their gene expression and protein network properties [48, 49]. Moreover, recent genome-wide analyses have shown that WGD duplicates in the human genome have experienced fewer SSD than genes *not* coming from WGD events and tend to be refractory to copy number variation (CNV) caused by polymorphism of small segmental duplications in human populations [41]. All these findings highlight the antagonist retention patterns of WGD and SSD gene duplicates and suggest the relevance of WGD for vertebrate evolution.

2.1.4 Critical role of WGD in evolution

Recent studies on the retention of duplicated genes suggest the critical role of duplication events (especially WGD) during evolution, as outlined in the previous section 2.1.3. However, the importance of gene duplication in supplying raw genetic material to biological evolution has been recognized since the 1930s [50]. In particular, duplication of genes (and their subsequent functional divergence) can now be considered the major evolutionary mechanism to generate new genes and rewire cellular pathways and networks [11]. Without gene duplication, the plasticity of a genome in adapting to changing environments would be severely limited.

In the late 1960s Susumu Ohno outlined the potential role of gene duplication as the driving force behind the evolution of increasingly complex organisms. He suggested that the huge boost in complexity in vertebrates was facilitated by the sudden increase in the availability of genetic material through WGD events, which was subsequently modeled by evolution in the following millions of years [1, 3]. Indeed, while SSD duplicates provide a continuous flux of genetic material, WGD events can favor unique evolutionary innovations, implying the simultaneous duplication of many genes at once. However, compelling evidence supporting this hypothesis (the so called “2R hypothesis”) remained elusive for a long time. Only recent genome wide studies have confirmed the occurrence of these WGD events at the origin of vertebrates [34–36] and WGD events have now been firmly established in almost all major eukaryotic lineages [51] (see Sec.2.1.2). Therefore, the two rounds of WGDs in the early vertebrate lineage are now credited with creating the conditions for the evolution of vertebrate complexity. Due to the pioneering works of Susumu Ohno, the genes retained from WGD events are now referred to as “ohnologs” [3, 52].

2.2 Evolutionary fate of gene duplicates

Gene duplicates arise frequently, either via local or genome-wide events. In particular, genome-wide analyses have estimated the average rate of origin of new gene duplicates to be of the order of 0.01 per gene per million years [7]. However, the majority of duplicated genes appears to be transient and only a minority is retained in the genome [6,11], leading to a still ongoing debate about the evolutionary mechanisms and constraints governing the retention of gene duplicates.

Newly duplicated genes (called *paralogs*) are assumed to initially have fully overlapping redundant functions [6,53,54]. In the absence of any advantage for this redundancy and due to the frequent occurrence of genetic degenerative mutations [55], it is commonly thought that one copy will usually become silenced by the random accumulation of degenerative mutations [3, 56–60]. Degenerative mutations disrupt the structure and the function of the gene such that it gradually becomes a *pseudogene*, which is either unexpressed or functionless [6, 54]. After a long evolutionary time, pseudogenes will either be deleted from the genome or become so diverged from the parental genes that they are no longer identifiable and the traces of duplication are lost [11]. Observations from the genomic databases for several eukaryotic species suggest that the vast majority of gene duplicates are silenced within a few million years [6]. Therefore, *non-functionalization*, the stochastic silencing of one copy, is considered to be the most likely fate of a duplicated gene (*e.g.* about $80 - 90\%$ of WGD duplicates are estimated to be lost from the genome through non-functionalization). However, it is now known that most eukaryotic genomes harbor large numbers of functional gene duplicates, many of which originated tens to hundreds of millions of years ago [61–64]. Different evolutionary mechanisms have been proposed to explain the preservation of duplicate genes, and the most credited ones are elucidated in the following sections.

2.2.1 Neo-functionalization

In the field, *neo-functionalization* has initially been suggested as the mechanism by which gene duplicates can permanently escape mutational degeneration. It is based on the idea that the initial functional redundancy of gene duplicates will allow one copy to acquire through mutation a new beneficial function that permanently preserves it in the population by natural selection, while the other copy will retain the original function [3, 65, 66]. However, clear and well studied examples of neofunctionalization are difficult to find [67, 68] and in many cases a related function that exists in some other genes in the genome, rather than an entirely new function, appears after gene duplication [11]. Therefore, the retention of duplicated genes through the evolution of truly 'new' functions, non-existent in the genome before duplication, is expected to be very rare.

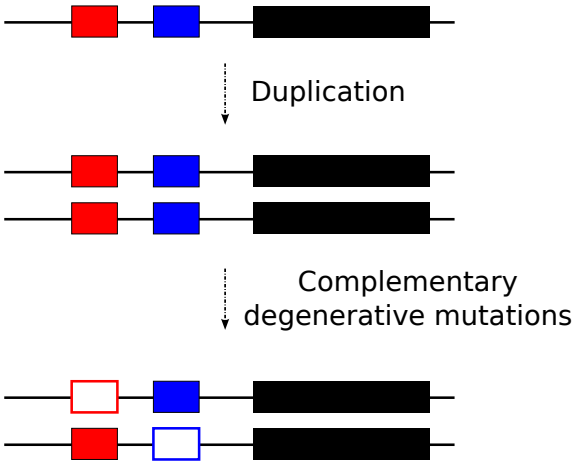

Figure 2.3: Retention of gene duplicates through subfunctionalization. The figure represents the simple case of a gene with two independently mutable subfunctions (depicted as regulatory regions by the two boxes in red and blue), which are spatially non-overlapping with each other and with the coding region (depicted as a black rectangle). Solid boxes denote intact regions of a gene, while open boxes denote degenerative mutations. After duplication, the two copies have lost single, non-overlapping subfunctions, and therefore complement each other. Because both copies are now essential to perform the ancestral gene function, they are retained in the genome. Figure adapted from [11, 53, 64].

2.2.2 Subfunctionalization

Experimental observations suggested that gene duplication could have allowed the specialization of separate α and β globin genes [69], and Jensen and Byng [70, 71] hypothesized that two enzymes specialized to catalyze two separate reactions may evolve, after gene duplication, from an enzyme capable of catalyzing both reaction. Therefore, another mechanism for the retention of duplicated genes was initially proposed [67] and then extensively studied [53, 54, 64]. This mechanism, called *subfunctionalization*, originates from the fact that genes often have several functions, each of which may be controlled by different DNA regulatory elements or binding partners [67]. Subfunctions are then defined as a specific subset of a gene's function that are often complementary. A subfunction might involve the expression of a gene in a specific tissue, cell lineage, or developmental stage [53]. According to this model, degenerative mutations in regulatory subfunctions can help the retention of duplicate genes, in the absence of any positive selection. Indeed, if duplicated genes loose different regulatory subfunctions through degenerative mutations, none of the paralogs alone can provide the original function. The only way to preserve the ancestral gene function is to complement each other by retaining the full set of original subfunctions. As a consequence, the task of the ancestral gene is partitioned and each gene duplicate will be preserved, provided the different subfunctions are essential for survival and/or reproduction (Fig. 2.3). Moreover, also several variants of this model have been proposed. For example, the partitioning of the ancestral gene functions may be driven by positive

selection, since each gene duplicate can acquire specific mutational refinements [67].

Numerous examples now exist for the partitioning of ancestral gene tasks, mostly involving developmental genes [38, 53, 72–74]. All these observations and the fact that degenerative mutations are much more frequent than beneficial mutations suggest that loss-of-subfunction mutations may play a relevant role in the evolutionary fate of gene duplicates.

2.2.3 Buffering against deleterious mutations

The presence of gene duplicates may confer robustness against deleterious mutations, since the duplicates can compensate each other's function, behaving as a backup mechanism [75]. This idea was initially proposed based on experimental studies [76, 77]. However, buffering alone should only rarely lead to the preservation of a pair of genes, since it requires that they are completely redundant in function [75, 78].

2.2.4 The dosage balance hypothesis

The dosage balance hypothesis has been proposed to explain the distinct properties of SSD and WGD duplicates, whose antagonist retention pattern has recently become apparent (see Sec. 2.1.3).

Evidence from a variety of data suggests that in multicellular eukaryotes the stoichiometric relationship of the components of regulatory complexes affects target gene expression. This mechanism sets the level of gene expression and, as a consequence, the phenotypic characteristics [79, 80]. This concept has been successively extended from regulatory to all protein complexes [81]. Therefore, the relative dosage (*i.e.* the amount of protein expressed) of genes belonging to the same complex or to the same metabolic pathway plays an important role for the proper formation and functioning of cellular assemblies and must be preserved [45, 75, 79–83] (Fig. 1 in [83]).

Since a WGD event implies a simultaneous duplication of the entire set of genes of an organism, it leads to the initial preservation of dosage balance constraints. As a consequence, it was supposed that the complete set of genes whose products participate in protein–protein interactions tend to be retained to prevent the loss of only one gene that would lead to the deleterious effects of dosage imbalance [81]. By contrast, duplication through SSD of only one of the interacting partners leads to dosage imbalance and has been proposed to be opposed by natural selection [45, 75, 82]. In particular, studying the yeast duplicates, Papp et al. [81] observed that WGD-retained genes are somewhat enriched in protein complexes and suggested that an imbalance in the components of protein complexes leads to lower fitness [9]. Since both the loss of a WGD duplicate and the dupli-

9. The fitness of an individual is a measure of the degree of its adaptation to its environment, and is defined as the reproductive contribution of an individual to the next generation. In mathematical terms, it corresponds to the expected number of offspring that reach adulthood.

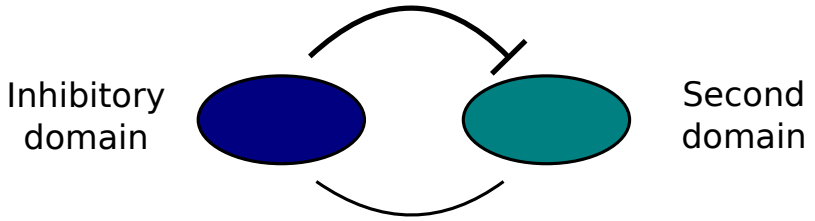

Figure 2.4: The regulatory mechanism of autoinhibition. An autoinhibitory domain (the protein region in dark violet) modulates the activity of a second, separable domain (dark green) to negatively regulate the protein activity. This regulatory system restricts the signaling pathway response only to specific signals. If the autoinhibitory domain is mutated, this negative control is lost and the protein activity can be strongly enhanced, causing huge deleterious effects. Figure adapted from [86].

cation through SSD in protein complexes are thought to be opposed by selection during evolution, the dosage balance hypothesis has been frequently invoked to explain the biased retention of SSD and WGD genes in a variety of organisms such as yeast [81], *Paramecium* [84], *Arabidopsis* [43] and human [41], by seeking enrichment of protein complexes in WGD duplicates.

2.3 Dominant deleterious mutations

Besides beneficial and degenerative mutations (whose evolutionary effects are outlined in Sec.2.2), another kind of mutations has an impact on the long term evolution of duplicated genes. This is the case of dominant deleterious mutations, leading to genetic disorders associated with a single defective allele (in one of the two possible variants at one locus in diploid genomes, while diseases caused by recessive mutations require the two alleles of the gene to be affected, as the production of a functional protein from one allele is often sufficient to satisfy physiological requirements) [85].

For example, some genes are characterized by the propensity to acquire deleterious *gain-of-function* mutations, that lead to constitutively active mutants with dominant deleterious phenotypes. This means that the activity of the gene (once mutated) is typically enhanced (gain) and cannot be masked by possible other non-mutated functional copies of the gene, thus causing huge dominant deleterious effects often associated to diseases in human. Indeed, a mutation dominantly-acting that occurs in a single gene is sufficient to cause the disease state.

Dominant deleterious mutations are related, in particular, to oncogenes and genes with autoinhibitory domains [86]. Autoinhibitory domains are regions of a protein that inhibit the function of other domains of the same protein, through intramolecular interactions (Fig. 2.4). The discovery of this regulatory pathway has been guided by the experimental observation of the enhanced activity of a particular protein domain in the absence of some other region of the protein, indicating the deletion of the autoinhibitory domain [86]. The

precise regulation of protein activities is essential for normal growth and development and autoinhibition is a widespread phenomenon that plays a key role in the regulation of proteins by facilitating the response to signaling pathways. Indeed, the autoinhibitory domain is an on-site repressor that restrains the targeted domain in a secure off state, preventing spurious inappropriate activation of a signaling pathway. Therefore, this regulatory system allows for response only to appropriate signals. If mutated, the autoinhibitory domain can be disrupted and loose its inhibitory function, leading to an enhanced uncontrolled protein activity and eventually to activities resulting in tumor genesis and cancer progression induced by gain-of-function (or dominant negative) rather than loss-of-function mutations [86].

2.3.1 The great expansion of dominant deleterious gene families

Dominant deleterious mutations, that induce dominant deleterious effects and thus are lethal or drastically reduce the fitness over the lifespan of organisms, must also have an impact on the long term evolution of the genes susceptible to these mutations, on timescales relevant for the evolution of the genome (e.g., $>$ 10–100 MY). Indeed, dominant disease genes (genes that harbor disease-causing dominant deleterious mutations) in human have been shown to be under strong purifying selection [85, 87, 88].

However, many vertebrate disease genes are phylogenetically ancient [88–90] and their orthologs [10] also cause severe genetic disorders in extant invertebrates [91–93]. Moreover, it has been revealed that disease gene families implicated in cancers and severe genetic diseases have been greatly expanded by duplication in the course of vertebrate evolution (e.g. [94, 95]), mainly through the two rounds of WGD that occurred at the origin of vertebrates [40, 41, 90] (see Sec. 2.1.2). In particular, Singh et al. [40] found a strong association between the retention of human ohnologs from vertebrate WGDs and their reported susceptibility to dominant deleterious mutations (see Fig. 10.1A in [96]). Comparing vertebrate with invertebrate genomes, this surprising result corresponds to an expansion of deleterious gene families in vertebrates of about $100 - 200\%$ genes, that is significantly higher than the global vertebrate genomes expansion of about $20 - 50\%$ genes [40]. As an example, Fig. 1 in [97] shows the duplication of the single orthologous locus Ras in *Drosophila* fruit fly into three Ras loci that have been retained in typical vertebrates, KRas, HRas, and NRas, that present permanently activating mutations in 20%–25% of all human tumors [98]. Interestingly, their duplication can be traced back to the two rounds of WGD in the early vertebrate lineage and led to a substantial expansion of the Ras-Ral signaling pathway [97].

All these observations imply an evolutionary puzzle from a natural selection perspective. Indeed, while gene duplicates are thought to confer some mutational robustness against loss-of-function mutations [99–102], the duplication of genes prone to dominant

10. Orthologs are genes in different species that share a common ancestral gene due to a speciation event.

deleterious mutations is expected to lead to an enhanced susceptibility to genetic diseases and, hence, be opposed by purifying selection [85, 87, 88]. Yet, surprisingly, gene families prone to dominant deleterious mutations have been greatly expanded through the two rounds of WGD dating back from the onset of jawed vertebrates, unlike most other vertebrate gene families [40, 41, 90].

2.4 Objectives

Recent studies have shown that gene families prone to dominant deleterious mutations, frequently implicated in cancers and other genetic diseases in vertebrates and thus expected to be opposed by purifying selection, have been, on the contrary, greatly expanded in the course vertebrate evolution, as discussed in the previous section. This evolutionary puzzle opens the way for the investigation of the selective mechanism responsible for this evolutionary outcome, relating the mode of duplication to the propensity to acquire dominant deleterious mutations. Indeed, as discussed above, these disease gene families have been greatly expanded through the two rounds of WGD that occurred at the origin of vertebrates. By contrast, gene families lacking such a susceptibility to dominant deleterious mutations have been more typically expanded through SSD [40]. This implies that the mode of duplication through SSD or WGD events directly impacts the selection process of gene duplicates, depending on their specific susceptibility to genetic mutations.

The peculiar retention pattern of WGD duplicates, which have a strong association with human diseases, has frequently been suggested to result from dosage balance constraints [41] (see Sec. 2.2.4). However, extensive statistical analysis combining multiple properties of human genes (such as dosage balance constraints, association to cancers and genetic diseases and expression levels) have recently demonstrated that the retention of WGD duplicates in vertebrates is more directly related to their susceptibility to dominant deleterious mutations than to dosage balance constraints or expression levels [40].

Singh et al. [40] investigated the evolutionary causes responsible for the expansion of gene families prone to dominant deleterious mutations in vertebrates and proposed a simple qualitative evolutionary model that accounts for their antagonistic retention pattern after WGD and SSD events. In particular, they argued that the enhanced retention of ohnologs susceptible to dominant deleterious mutations is a consequence of the speciation event triggered by WGD and the ensuing purifying selection in post-WGD species. The first aim of our study is to rationalize, through a quantitative approach and from an evolutionary perspective, these intriguing observations and verify the consistency of the resulting predictions with human cancer genome data. From a wider perspective, we want to investigate the effect of speciation on the selection of specific gene duplicates, that has been largely overlooked so far. For this purpose, we propose a consistent population genetics model taking into account the impact of the mode of duplication (WGD *vs* SSD) on the retention of gene duplicates prone to dominant deleterious mutations.

In [40], Singh et al. also demonstrated, using a causal inference analysis, that the retention of many ohnologs suspected to be dosage balanced is in fact an indirect effect of their higher susceptibility to deleterious mutations. Moreover, a number of studies have now shown that many genomic properties appear to be correlated to some extent, suggesting many indirect statistical associations [41, 103, 104]. Our next aim is to further analyze ohnolog retention from a broader perspective, including all possible properties known to be correlated to it, and uncover if statistically significant correlations may result from indirect rather than direct associations. In order to go beyond statistical correlations and quantify direct and indirect causal effects, we will extend advanced inference methods to analyze the multiple, direct, and indirect causes underlying the evolution of gene duplicates.

Part II

Materials & Methods

Chapter 3

Population genetics approach

The great yet counterintuitive expansion of gene families prone to dominant deleterious mutations, introduced in Sec. 2.3.1, calls for a consistent evolutionary model to rationalize this apparent evolutionary oddity. A quantitative approach dealing with evolutionary forces and mechanisms acting at the genetic level of a population needs to be established in the framework of population genetics. Indeed, population genetics is concerned with the study of the genetic composition of populations under the joint action of genetic and evolutionary factors such as natural selection, mutation, recombination, mating structure, population structure, migration. In particular, it investigates the dynamics of genetic variation within species under the interactions of these mechanisms, which lead to evolutionary changes, adaptation, and speciation in populations. The field of population genetics was founded in the late 1920s and early 1930s by Fisher, Haldane, and Wright, who set the mathematical framework of theoretical investigations for the subsequent decades.

In this context, we will propose a quantitative model to rationalize the great expansion of gene families prone to dominant deleterious mutations during vertebrate evolution, focusing on the effect of speciation on the selection of gene duplicates. We will explicitly take into account the mode of duplication (WGD *vs* SSD) and the gene propensity to acquire dominant deleterious mutations. Both analytical and stochastic approaches will allow to demonstrate that the enhanced retention of ohnologs prone to dominant deleterious mutations is an indirect consequence of the initial speciation induced by WGD and the ensuing purifying selection in post-WGD species. We will further verify the consistency of the resulting predictions with human cancer genome data (see Chapter 5).

3.1 Hypothesis: a qualitative model recently proposed

A new hypothesis has been recently suggested to explain the great expansion of gene families prone to dominant deleterious mutations [40]. Starting from the evidence that human disease genes have been mainly retained from the two rounds of WGD at the onset of vertebrates, Singh et al. [40] proposed a qualitative evolutionary model that shows

how this peculiar retention pattern is a consequence of the speciation event triggered by WGD and the ensuing purifying selection in post-WGD species. In short, ohnologs have been suggested to be retained in the vertebrate genome not because they initially brought selective advantages, but because they were more susceptible to detrimental than nonfunctional mutations, thereby preventing their rapid elimination from the genomes of surviving individuals following WGD transitions. A similar proposition had been made in an article comment by Gibson and Spring [105].

Figure 4 in [40] depicts all the possible evolutionary scenarios following either a SSD or a WGD duplication event occurring in the genome of one or a few individuals in an initial population, highlighting the different outcomes between SSD and WGD scenarios in the case of dominant deleterious mutations (scenario C). The critical difference between WGD and SSD events is the obligate speciation after WGD owing to the difference in ploidy between pre- and post-WGD individuals and the subsequent interbreeding incompatibility, resulting in a post-WGD population where all individuals carry twice as many genes as their pre-WGD relatives. By contrast, a SSD event does not typically imply a speciation event and thus only a few individuals in the post-SSD population carry a single small duplicated region. The figure then outlines the three mutation/selection scenarios focusing on a single gene duplicate in the genomes of post-SSD or post-WGD populations:

(A) Beneficial mutations are expected to rarely occur but can then spread and become eventually fixed in the new population after both duplication events, although the bottleneck in population size following WGD limits in practice the efficacy of adaptation in post-WGD species.

(B) Neutral mutations mainly lead to the random non-functionalization of one copy of redundant gene duplicates and, therefore, to their elimination following both SSD and WGD events. In rare cases, neutral mutations can also result in the retention of both duplicate copies through subfunctionalization (see Sec. 2.2).

(C) Dominant deleterious mutations favor the elimination of the individuals harboring them, through purifying selection. However, this typically leads to opposite outcomes in post-SSD and post-WGD populations. In post-SSD populations, dominant deleterious mutations will tend to eliminate SSD duplicates *before* they have time to reach fixation, bringing the population back to the initial situation. By contrast, in post-WGD populations where all ohnologs have been initially fixed through the WGD-induced speciation, purifying selection will indirectly favor the retention of ohnologs prone to dominant deleterious mutations, since all surviving individuals still present functional copies (not yet mutated, thus not yet deleterious) of these genes.

In order to rationalize this intriguing evolutionary outcome, we propose some quantitative consistent models in the context of population genetics. We will compare the evolutionary fates of SSD with WGD duplicates, taking into account their propensity to acquire dominant deleterious mutations. We will finally show that these models support the idea that the enhanced retention of ohnologs prone to dominant deleterious mutations is an

indirect consequence of the initial speciation induced by WGD and the ensuing purifying selection in post-WGD species.

3.2 Population genetics models: a deterministic approach

In order to analyze the retention of gene duplicates originated from either a SSD or WGD event, we first propose a simple, deterministic model that is analytically tractable and represents an approximation of the discrete dynamics of a population at the genetic level in the limit of large population size. We assume the case of two duplicated loci in a haploid population to limit the number of two-locus combinations. Extensions to diploid models will be then considered. For the sake of simplicity, we further assume a population of fixed size N and uncoupled mutation/selection dynamics. Using a continuous time model, we are implicitly assuming that generations overlap and that reproduction is always occurring. Despite these simplifying assumptions, the asymptotic solutions of this deterministic population genetics model allow to capture the main evolutionary process responsible for the different retention of SSD *versus* WGD duplicates caused by deleterious gain-of-function mutations.

3.2.1 Simple haploid deterministic model

We start to consider the gene duplication event $A \rightarrow AA_o$ for the initial locus A in a haploid genome. The only difference between the two mode of duplication, SSD and WGD, concerns the initial condition for the fraction ϵ of individuals with duplicated loci in a population of size N. The SSD scenario corresponds to the gene duplication event $A \rightarrow AA_o$ occurring in the genome of a *single* (or few) individual(s) of the initial population, leading to $\epsilon \simeq 1/N \ll 1$. Instead, the WGD scenario corresponds to the gene duplication $A \rightarrow AA_o$ occurring in the genome of *all* individuals in the newly arisen population through the speciation event triggered by WGD, implying $\epsilon = 1$.

Since the newly duplicated gene is usually assumed to be initially functionally redundant [6, 53, 54, 64], we assign to the initial (unstable) genotype with two redundant duplicated loci AA_o a neutral fitness parameter $\omega_o = 1$, indicating the absence of any selective advantage/disadvantage for the gene duplicates. Then, we consider three possible mutation-selection scenarios, corresponding to the emergence, through mutations, of three different phenotypes from the initial genotype AA_o, as illustrated in Fig. 1 in Malaguti et al. [106]. These three scenarios correspond to the main alternative evolutionary fates for gene duplicates traditionally credited in the literature and analyzed in classical models [6, 11, 53, 54, 64], but include also the case of gain-of-function mutations that lead to an enhanced activity of the gene and are associated to a dominant deleterious phenotypes. Thus, this model allows to specifically address the long term evolution of gene families implicated in cancers and other severe genetic diseases and investigate the selective mechanisms responsible for their observed great expansion. In particular, the

evolutionary outcomes of gene duplicates classically studied are the following (see Sec. 2.2 for further details),

(i) *non-functionalization*: one gene copy becomes silenced by the accumulation of degenerative mutations, while the other (fully functional) copy is retained in the genome. In our model, this corresponds to a neutral phenotype due to a loss-of-function (with mutation rate ν_-) of one of the duplicate. The corresponding genotype is indicated as AA_-, and associated with a neutral fitness parameter $\omega_- = 1$ since the copy retained can fully perform the ancestral gene function without any selective advantage/disadvantage;

(ii) *subfunctionalization*: both copies may be preserved due to complementary loss-of-subfunction mutations, leading to the partition of the tasks of the ancestral gene function and the necessary joint retention of both gene copies to fully perform the original function. In our model, this corresponds to a neutral (or possibly beneficial) phenotype due to the retention of the two non-equivalent duplicated loci through subfunctionalization (with mutation rate $\nu_\star$). The corresponding genotype is indicated as $AA_\star$, and associated with a neutral (or possibly beneficial) fitness parameter $\omega_\star = 1 + s_\star \geq 1$;

(iii) *neofunctionalization*: one copy may acquire a novel beneficial function while the other retains the original function. However, the incidence of beneficial mutations is negligible compared to the frequency of other mutations and we do not explicitly include this possibility in our model, although it can be implicitly lumped together with subfunctionalization with $s_\star > 0$.

In addition to these classical evolutionary scenarios for gene duplicates, we specifically include the case of *gain-of-function* mutations (see Sec. 2.3). In our model, constitutive gain-of-function mutations occur (with mutation rate ν_+) in one of the duplicate genes, leading to a dominant deleterious phenotype. The associated genotype is indicated as AA_+. Since dominant deleterious mutations drastically reduce the fitness of the individual, they correspond to a deleterious phenotype characterized by a fitness decrement $\omega_+ = 1 - s_+ < 1$ (see Fig. 1 in Malaguti et al. [106].

Subfunctionalization of the duplicated loci ($AA_\star$) implies, in principle, degenerative mutations at both loci, which can no longer perform the full function of the ancestral gene A [64, 67]. By contrast, loss-of-function (AA_-) and gain-of-function (AA_+) genotypes involve mutations at a single locus and are assumed to retain a fully functional copy of the ancestral gene A at the other locus. However, while this functional copy can compensate the deleterious effect of loss-of-function mutations in AA_-, resulting in neutral fitness $\omega_- = 1$, it is unable to mask the deleterious effects of the enhanced activity induced by gain-of-function mutations in AA_+, resulting in a fitness decrement $\omega_+ = 1 - s_+ < 1$. For the sake of simplicity, in this haploid model we will not distinguish on which copy the loss-of-function and gain-of-function mutations occur. In particular, we assume that the loss-of-function mutations on either duplicate copy lead to a genotype equivalent to the

ancestral one with a single gene copy, $AA_- \equiv A_-A \equiv A$.

We now consider the simplest deterministic population genetics model to analyze the fixation of SSD *versus* WGD duplicates. We note $\phi_o(t)$, $\phi_+(t)$, $\phi_-(t)$, $\phi_\star(t)$ the fractions of individuals in the population with the corresponding genotypes for the duplicated loci, AA_o, AA_+, AA_- and $AA_\star$. The equations linking these genotypes and modeling their evolution over time through an uncoupled mutation/selection dynamics are

$$\begin{aligned} d_t\phi_\circ &= (w_\circ - \bar{w})\phi_\circ - (\nu_+ + \nu_- + \nu_\star)\phi_\circ \\ d_t\phi_+ &= (w_+ - \bar{w})\phi_+ + \nu_+\phi_\circ \\ d_t\phi_- &= (w_- - \bar{w})\phi_- + \nu_-\phi_\circ \\ d_t\phi_\star &= (w_\star - \bar{w})\phi_\star + \nu_\star\phi_\circ \end{aligned} \tag{3.1}$$

where $\bar{w}(t) = \sum_i w_i\phi_i(t)$ is the average fitness of the population. The first term on the right-hand side of each equation represents the fitness contribution to the change in allele frequency over time, while the second term is the independent effect of genetic mutations. In particular, noting $S = \sum_i \phi_i$, one can check that $d_tS = \bar{w}(1-S)$ leads to the expected constant, $S(t) = 1$ at all time, providing that $S(t=0) = 1$ is taken as initial condition. The initial fraction of individuals with duplicated loci AA_o, $\epsilon = \phi_o(0)$, allows to discriminate the WGD ($\epsilon = 1$) from the SSD ($\epsilon \simeq 1/N \ll 1$) scenario.

In the case of neutral fixable genotypes through subfunctionalization for the duplicated loci ($w_\star = w_- = 1$), the average fitness of the population can be expressed as $\bar{w} = \sum_i w_i\phi_i = 1 - s_+\phi_+$, and the system of equations becomes

$$\begin{aligned} d_t\phi_\circ &= s_+\phi_+\phi_\circ - (\nu_+ + \nu_- + \nu_\star)\phi_\circ \\ d_t\phi_+ &= (s_+\phi_+ - s_+)\phi_+ + \nu_+\phi_\circ \\ d_t\phi_- &= s_+\phi_+\phi_- + \nu_-\phi_\circ \\ d_t\phi_\star &= s_+\phi_+\phi_\star + \nu_\star\phi_\circ \end{aligned} \tag{3.2}$$

The asymptotic solutions for this deterministic model assuming neutral fixable genotypes will be analytically obtained, showing a different retention of WGD *versus* SSD duplicates for genes prone to dominant deleterious mutations (see Sec. 5.1). Extensions to adaptive selection of duplicates with $\omega_\star > 1$ will be obtained in section 5.2.3 through simulations of the stochastic approach outlined in section 3.3.

3.2.2 Extension to diploid models

The extension to a diploid population is essential to gain a more general perspective on the long term evolution of gene duplicates prone to dominant deleterious mutations. Indeed, we want to specifically address the issue of the great expansion of disease gene families in vertebrates, and test the consistency of the theoretical predictions with the available data about WGD duplicates belonging to gene families with oncogenic proper-

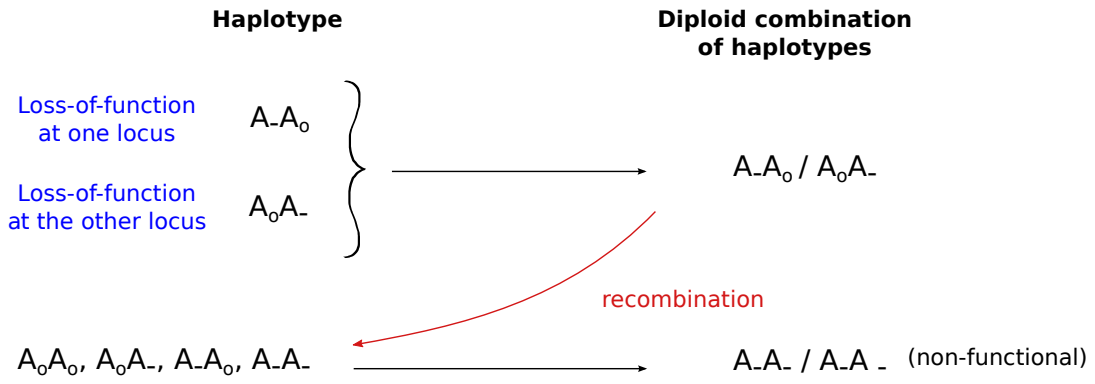

Figure 3.1: Breaking of symmetry in the divergence of multiple alleles at duplicated loci. The figure depicts the specific effect of reciprocal gene loss at duplicated loci on haplotypes (left side) and the corresponding diploid combinations of haplotypes (right side) in an evolving population. The recombination process ultimately leads to a non-functional diploid genotype $A\_A\_/A\_A\_$.

ties and responsible for a broad range of primary tumors in human. However, the extension of the previous haploid model to a diploid model including epistatic interaction and recombination between four different alleles at each duplicated locus implies a combinatorial proliferation of two-locus diploid genotypes, such as $A_\circ A\_/A_+A_\circ$, $A_\circ A_\star/A_+A\_$, etc.

In addition to this multiplicity of diploid states, we also expect further complications due to the process of duplication-driven speciation proposed by Werth et al. and Lynch et al. in [107, 108]. These authors suggest a simple genetic mechanism for the allopatric origin of new species, namely that it is driven by gene duplication and degenerative mutations, which are much more common than beneficial mutations. In particular, it is based on the idea that interbreeding barriers between individuals having lost different copies of the duplicated loci are at the origin of duplication-driven speciation. Indeed, the reciprocal gene loss at duplicated loci (*i.e.* $A\_A_\circ$ or $A_\circ A\_$) leads to the diploid combination of haplotypes with reciprocal loss of duplicates, $A_\circ A\_/A\_A_\circ$, which readily recombines to yield a double mutant haplotype, $A\_A\_$, and ultimately a non-functional diploid genotype $A\_A\_/A\_A\_$ (Fig. 3.1). This non-functional diploid genotype effectively lowers the interspecific compatibility between individuals coming from subpopulations carrying primarily the $A_\circ A\_$ or the $A\_A_\circ$ haplotype (Fig. 3.2). Similar subpopulation structures are expected to arise from independent breakings of symmetry in the divergence of multiple alleles at duplicated loci, such as with the two functional haplotypes $A\_A_\circ$ and $A_\circ A_\star$ (with functional $A_\circ$, non-functional $A\_$ and sub-functional $A_\star$) which lead, after recombination, to the non-functional diploid genotype, $A\_A_\star/A\_A_\star$.

In order to circumvent these complications in analyzing the retention of a single or two gene copies with multiple alleles at duplicated loci, we will in fact consider only one breaking of symmetry and reciprocal gene loss scenario below, while keeping in mind that alternative scenarios can exist and possibly co-exist as different subpopulations or species. This assumption amounts to simplify the actual two-locus four-allele diploid system into an ef-

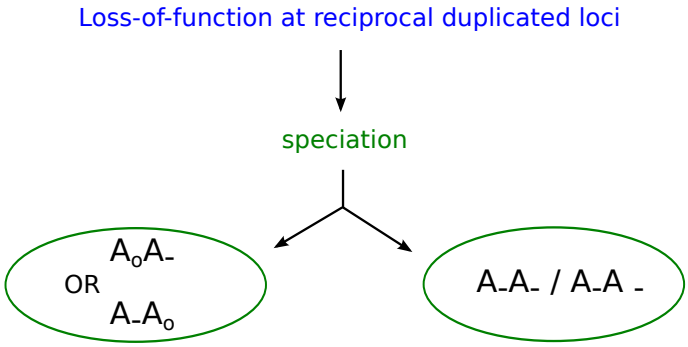

Figure 3.2: Duplication-driven speciation. Interbreeding barriers between the non-functional diploid genotype A_-A_-/A_-A_- originated by reciprocal gene loss at duplicated loci (Fig. 3.1) and individuals having lost different copies of the duplicated loci (belonging to subpopulations carrying primarily the $A_\circ A_-$ or the $A_-A_\circ$ haplotype) are responsible for the formation of new species.

fective one-locus four-allele diploid system, based on the four haplotypes introduced earlier, *i.e.* $AA_\circ$, AA_-, AA_+ and $AA_\star$. If we further assume, for the sake of simplicity, that there is no difference between maternal and paternal inherited haplotypes, we are left to consider only ten diploid combinations of these haplotypes, *i.e.* $AA_\circ/AA_\circ$, $AA_\circ/AA_-$, $AA_\circ/AA_+$, $AA_\circ/AA_\star$, AA_-/AA_-, AA_-/AA_+, $AA_-/AA_\star$, AA_+/AA_+, $AA_+/AA_\star$, $AA_\star/AA_\star$.

We can now study the effects of the dominant deleterious phenotype caused by the AA_+ haplotype, assuming, otherwise, a neutral fitness for all diploid combinations without AA_+. This leads to the following marginal fitness for each haplotype, obtained averaging over the fitnesses of all genotypes in which that haplotype can be found,

$$w_i = w_i^\circ(1-\phi_+) + w_i'\phi_+,$$

where $w_i^\circ = 1$ and $w_i' = 1-hs_+$ for $i = \circ, -, \star$ and $w_+^\circ = 1-hs_+$ and $w_+' = 1-s_+$, where h is the dominance coefficient of the heterozygous diploid genotypes including one haplotype AA_+. In particular, $h = 1$ corresponds to a simple dominant deleterious mutant, while $h = 1/2$ corresponds to a co-dominant deleterious mutant with additive deleterious effects for the AA_+/AA_+ diploid genotype. This leads to the average marginal fitness of the population,

$$\bar{w} = \sum_i \phi_i w_i = 1 - 2hs_+\phi_+ + s_+(2h-1)\phi_+^2$$

and the relative marginal fitness for each haplotype,

$$\begin{aligned} w_+ - \bar{w} &= -hs_+ + s_+(3h-1)\phi_+ - s_+(2h-1)\phi_+^2 \\ w_i - \bar{w} &= hs_+\phi_+ - s_+(2h-1)\phi_+^2, \end{aligned}$$

for $i = \circ, -, \star$. Thus, if the fraction of dominant deleterious haplotype AA_+ remains small in the population, $\phi_+ \ll 1$, as expected and confirmed by simulations (see Sec. 5.2.1),

we retrieve the same population genetics system as for the haploid model studied earlier, Eqs. (3.1,3.2), in the case of dominant deleterious mutations ($h = 1$) or in the case of incomplete dominance ($0 < h < 1$), if the fitness decrement is rescaled as $s_+ \to s_+ h^{-1}$.

Hence, with these simplifications, the two-locus, four-allele diploid system of duplicated loci behaves essentially like a one-locus, four-allele haploid system. This is the population genetics model that we will further consider to study the stochastic effects in populations of finite size and more general coupled mutation/selection dynamics.

3.3 Population genetics models: a stochastic approach for small populations

The allele frequency states of a finite population are discrete and a proper analysis of their change over time requires a stochastic treatment, especially if the population size is small - as it can be in concomitance of bottleneck events. The first population genetics stochastic model of an evolving population dates from the 1930s and was introduced independently by Fisher [109] and Wright [110]. In this model, the number of individuals in the population is taken to be constant from one generation to the next, and each gene in one generation is an exact copy of a gene randomly chosen with replacement from the previous generation. A variant of the Wright–Fisher model was introduced by Moran [111]. The Moran model, unlike the Wright–Fisher model, has overlapping generations: it is a birth-and-death stochastic process in which at each step one individual is chosen to reproduce and one individual (perhaps the same) is chosen to die. Thus, after few sampling events, each gene need not have been replaced, and several might not have survived very long at all.

The proper equation for a stochastic treatment of allele frequencies in the continuous time limit of these models is a discrete (in allele frequency) master equation. The master equation of the Moran model is a special case called a one-step master equation, based on transitions for a single stochastic step. Instead, the Wright–Fisher model is quite difficult to analyze, especially when mutations are allowed for. Therefore, a partial differential equation approximating the master equation describing it, the Fokker-Planck or diffusion equation, is usually used. It describes the change over time of the probability distribution of allele frequencies under the influence of different evolutionary forces. The use of the diffusion equation in problems related to population genetics was first suggested by Kolmogorov to Wright [112] and was successfully applied by Kimura [113] to genetic drift. The diffusion equation is an approximation of the discrete master equation governing the dynamics of a stochastic system for large populations: if the size N of the population is sufficiently large to neglect terms smaller than 1/N, then the discrete master equation can be written as a continuous (in allele frequency) partial differential equation. Since the resolution of partial differential equation is much more advanced than discrete equations, the diffusion equation has been proved very popular and has become a standard technique

of population genetics theory. However, this is an approximation of order $1/N$ and is not suitable for small populations. Moreover, it assumes that selection and mutation are sufficiently weak, of order $1/N$. Finally, the original Kimura equation is a forward equation and important quantities such as the fixation probabilities of absorbing states cannot be computed directly, but one has to resort to the accompanying backward equation [114,115].

Therefore, we consider in our study a stochastic approach based on a general one-step process master equation formalism to describe the dynamics of a population of finite fixed size N with more than two alleles. Then, we will reduce to the specific case of the four alleles introduced in section 3.2.1. Finally, we will perform simulations of the stochastic population genetics models corresponding to the master equation that we will introduce, in order to analyze the stochastic effects of finite population sizes on the retention of SSD and WGD duplicates.

3.3.1 General approach for K alleles

We use a one-step process master equation formalism between $K \geqslant 2$ alleles in order to treat the stochastic allele frequency changes in a finite population. Moreover, this approach enables to include more realistic coupled mutation/selection dynamics such as the Moran model [116] in addition to the uncoupled mutation/selection dynamics that we initially considered for the haploid deterministic model in section 3.2.1.

We consider a population of finite size N in the context of one-locus haploid systems. The generic one-step process master equation for K alleles $A_1,\cdots,A_K$ ($K \geqslant 2$) governing the probability, $P(n_1,\cdots,n_K,t)$, of observing n_i individuals with allele A_i at time t (with $\sum_{i=1}^{K} n_i = N$), is

$$\frac{\partial P(\{n_k\},t)}{\partial t} = \sum_{i,j=1}^{K} (\mathbb{E}_i^{-1}\mathbb{E}_j^{1} - 1) W_{ij}(\{n_k\}) P(\{n_k\},t)$$

where $\mathbb{E}^{\pm 1}$ is the "step operator" [117] such that $\mathbb{E}_i^{\pm 1} f(n_i) = f(n_i \pm 1)$. Each subpopulation j of size n_j has transition rates from allele j to allele i that can be expressed in terms of the numbers of individuals with the different alleles as,

$$W_{ij}(n_1,\cdots,n_K) = \frac{n_j}{N} \sum_k \beta_{ik}^{(j)} n_k$$

where n_j/N is the probability that one individual with the allele j is randomly chosen to die and $\beta_{ik}^{(j)} n_k$ is the rate at which one individual with allele k is chosen to reproduce and mutate into the allele i, given that an individual with allele j has been chosen to die. This general expression enables to include both coupled and uncoupled mutation/selection dynamics depending on the definition of the reproduction/mutation rates $\beta_{ik}^{(j)}$. In particular, three main population genetics models have been studied in the literature: two models with coupled mutation/selection processes correspond to the first and second Moran models [116] with mutations occurring either before or after selection, respectively.

The first Moran model essentially selects on the lifespan of adults rather than their reproductive success, while the second Moran model amounts to a gametic selection independent of death rate, see Appendix A.1. By contrast, the uncoupled mutation/selection model outlined in the section 3.2.1 amounts to use, as model parameters, "average" mutation rates $\bar{\nu}_{ij} = \sum_{k\neq i} \beta_{ik}^{(j)} \phi_k$, for $j \neq i$, and "average" selection rates $\bar{w}_i = \sum_j \beta_{ii}^{(j)} \phi_j$ and $\bar{w}^{(i)} = \sum_k \beta_{kk}^{(i)} \phi_k$, see Appendix A.1. Uncoupled mutation/selection models have been frequently used in recent years for multiallelic systems [118–122].

These different mutation/selection models can then be applied to study the fixation of gene duplicates following either a SSD or a WGD event. To this end, we consider the multiallelic model with the four different alleles introduced earlier in section 3.2.1 ($K = 4$, Fig. 1 in Malaguti et al. [106]) corresponding to the initial (unstable) duplicate state $AA_\circ$ as well as the three alleles arising through mutations from $AA_\circ$, namely, $AA_- \equiv A$, AA_+, and $AA_\star$. The rates of mutations from j to i then correspond to $\bar{\nu}_{ij} = \bar{\nu}_{i\circ(i\neq\circ)} = \nu_i$ with $i = \star, -, +$. It is worth noting that when all fitness parameters are neutral except for the fitness disadvantage of dominant deleterious mutants (*i.e.* $w_\circ = w_- = w_\star = 1$ and $w_+ = 1 - s_+$, where $s_+ \ll 1$), the two coupled mutation/selection models by Moran [116] lead to very similar deterministic equation systems as the uncoupled mutation/selection model of Eqs. (3.1,3.2) in the large population size limit ($N \gg 1$), see Appendix A.2. Thus, the deterministic solutions for allele frequencies are only slightly affected by the details of the stochastic models. Beyond this observation, our main interest in the master equation formalism remains in the stochastic effects encompassed in the full distribution, solution of the master equation. However, they are not accessible analytically in the case of four alleles. Yet, stochastic simulations directly corresponding to the master equation detailed above allow to analyze the effects of finite population sizes on the retention of SSD and WGD duplicates. Moreover, extensions to adaptive selection of duplicates with $w_\star > 1$ can be directly assessed through simulations, as we will discuss in section 5.2.3.

3.3.2 Stochastic simulations

We have performed stochastic simulations of the birth, death and mutation processes for the three population genetics models corresponding to the one-step process master equation detailed in section 3.3.1. For each of the four alleles $k = \{AA_\circ, AA_+, AA_-, AA_*\}$, we keep track of a random variable $n_k(t)$ representing the number of individuals with allele k and fitness w_k at time t. We subdivide one generation into small time steps of length δt and update the frequency of each allele after every such time step.

We first consider the model with uncoupled selection and mutation, corresponding to Eqs. (3.1,3.2) in the deterministic limit of large population size. At each time step, the number of offspring b_k with allele k is obtained from a binomial distribution with mean $n_k w_k \delta t$. We then randomly remove a number of individuals d_k from the sub-population of allele k, so as to keep the overall population size constant, $\sum_k n'_k = \sum_k (n_k + b_k - d_k) = N$, where $n'_k = n_k + b_k - d_k \geqslant 0$ corresponds to the updated size of the sub-population k,

after birth and death steps. Finally, the stochastic mutations are generated independently from the selection process for the $n'_{AA_\circ}$ individuals in the unstable duplicate allele class $AA_\circ$ with mutation probability $p'_k = \nu_k \delta t$ from allele $AA_\circ$ to allele k where ν_k is the corresponding mutation rate per generation. The sub-population sizes are then updated to $n_{AA_\circ}(t+\delta t) = n'_{AA_\circ} - \sum_k m_k$ for the $AA_\circ$ allele and to $n_k(t+\delta t) = n'_k + m_k$ for $k = AA_-, AA_+, AA_\star$, where m_k represents the number of individuals mutated from allele $AA_\circ$ to allele k. The time step δt is typically chosen in the range of $0.01 - 0.1$ generation.

In the case of the Moran models with coupled mutation/selection dynamics, the transition W_{ij} removes one individual from class j (*i.e.* $j \to j-1$) and replicates one individual of class i (*i.e.* $i \to i+1$) in the time step δt, taking into account the coupling between birth, death and mutation at the same time. At each time step the transition rates $W_{ij}(n_1, \cdots, n_K) = (n_j/N) \sum_k \beta_{ik}^{(j)} n_k$ are computed using the coefficients $\beta_{ik}^{(j)}$ from the corresponding Moran models, with either mutations before selection (model 1) or mutations after selection (model 2). The transition $j \to i$ is then chosen stochastically according to its rate W_{ij} leading to the population updates $n_j = n_j - 1$ and $n_i = n_i + 1$, and a time increment $\delta t = (\sum_{ij} W_{ij})^{-1}$ summed over all possible transitions. In the case of the Moran model controlling death rate, we choose the death rate $\lambda_k = w_k^{-1}$, that approximates to $\lambda_k \simeq 1$ for $k = AA_\circ, AA_\star, AA_-$ and $\lambda_k \simeq 1 + s_+$ for $k = AA_+$, in the limit of small s_+, $0 < s_+ \ll 1$.

The mutation and selection parameters of the models are chosen in agreement with the available estimates in the literature [123]. The total mutation rate in the germline of vertebrates such as mouse or human is of the order of $1 - 4 \times 10^{-8}$ per nucleotide site per generation [123]. Taking an average gene length of 1000 to 1500 nt leads to an average mutation rate of $\nu_f = 4 \times 10^{-5}$ mutation per gene per generation. As the rate of sub-functionalization $\nu_\star$ is expected to be a small fraction of ν_f, we assume $\nu_\star = \nu_f/10 = 4 \times 10^{-6}$ per gene per generation. This corresponds to a fixation rate of about 10% of typical duplicates after WGD according to the solution of the deterministic system of Eqs. 3.2 (see Result section 5.1, Eq. (5.1) with $\Pi_e^{\mathrm{WGD}} = \epsilon = 1$ and $\nu_f \gg \nu_+$), in agreement with the average retention of ohnologs from each round of WGD at the origin of vertebrates, see Sec. 5.3 and [40, 106]. In addition, we assume that the local rates of gain-of-function and loss-of-function mutations vary depending on the gene local susceptibility to gain-of-function *versus* loss-of-function mutations at each position with a constant averaged sum across all genes, $\nu_+ + \nu_- = \nu_f - \nu_\star = 3.6 \times 10^{-5}$ per gene per generation. Hence, in the following, we will simply assign increasing values to ν_+, while keeping the sum $\nu_- + \nu_+$ fixed. The selective disadvantage s_d of a deleterious allele is known to be typically in the range of $s_d \simeq 10^{-3}/10^{-2}$ [123], thus, the value of the selection coefficient s_+ for the dominant deleterious mutant AA_+ is chosen as $s_+ = 0.05$ to emphasize its large deleterious phenotypic effect. Finally, we start either with a single individual with a SSD duplicate, leading to $\epsilon = 1/N$ for the SSD scenario, or with all individuals with WGD duplicates, leading to $\epsilon = 1$ for the WGD scenario.

Chapter 4

Mediation Analysis approach

In the previous chapter, we have proposed a population genetics model to investigate the evolutionary mechanism responsible for the observed biased retention of ohnologs in vertebrate genomes, discussed in Sec.2.3.1. In particular, we have related ohnolog retention to the gene propensity to acquire dominant deleterious mutations, frequently implicated in cancers and genetic diseases. Therefore, we have rationalized from a theoretical perspective previous observations suggesting that the susceptibility to dominant deleterious mutations is a critical factor in the retention of ohnologs after WGD events [40].

However, multiple genomic and functional properties are known to affect the retention of genes after duplication. In particular, evolutionary constraints to maintain balance in the relative dosage of the interacting sub-units in macromolecular complexes have been argued to underlie the observed antagonistic retention pattern of genes in different functional categories after WGD and SSD events (see Sec. 2.2.4). Therefore, in order to further analyze ohnolog retention from a broader perspective and assess the relative contribution of each of these properties to ohnolog retention, we need a framework where the relative causal influence of a property could be quantitatively evaluated.

Rather than studying two-properties correlations, we have decided to disentangle the direct from the indirect effects of many genomic properties on ohnolog retention through the Mediation Analysis theory, guided by the approach of Judea Pearl [124–127]. Indeed, the Mediation framework, developed in the context of causal inference analysis, aims at uncovering, beyond statistical correlations, causal pathways along which changes in multivariate variables are transmitted from a cause, X, to an effect, Y. More specifically, the Mediation Analysis assesses the importance of a mediator, Z, in transmitting the indirect effect of the variable X on the response variable Y.

The Mediation Analysis approach has already been applied to the study of ohnolog retention in order to discriminate the two alternative hypotheses currently proposed to explain it, namely dosage balance constraints and the susceptibility to dominant deleterious mutations [40]. In particular, the total, direct and indirect effects of deleterious mutations and dosage balance constraints on the biased retention of human ohnologs have been

quantified, allowing to show that the retention of many ohnologs suspected to be dosage balanced is in fact indirectly mediated by their susceptibility to deleterious mutations [40].

However, a number of studies have shown that many genomic properties, such as gene essentiality, duplicability, functional ontology, network connectivity, expression level, mutational robustness, divergence rates, etc., all appear to be correlated to some extent, suggesting direct statistical associations [41, 103, 104]. Yet, statistically significant correlations may result from indirect rather than direct associations. Therefore, in order to further analyze ohnolog retention from a broader perspective, we need to include in the Mediation Analysis framework all other genomic properties possibly correlated to ohnolog retention. For this purpose, we aim at extending the Mediation framework defined for three properties (X, Y, Z), proposing an empirical method of general applicability to quantify direct and indirect causes in a network of multiple causally related variables.

In this chapter, we will first introduce Pearl's Mediation Analysis approach, defining the decomposition of the total effect in its direct and indirect components and showing their easy estimation from the data in the simple binary case. Then, we will briefly summarize how this framework has allowed to demonstrate that the retention of ohnologs in the human genome is more directly caused by their susceptibility to deleterious mutations than their interactions within multi-protein complexes, as shown in [40]. Finally, we will present two approaches to generalize the causal Mediation Analysis method in order to include more than three variables at once. This will enable to simultaneously consider many genomic properties that could a priori affect ohnolog retention and quantify direct and indirect causal effects among them, assessing the sign and the strength of the causal interactions. Moreover, we will compare our method with the approach developed by Maathuis and coworkers [128], highlighting the deeper understanding of the causal network of relationships that our method allows to obtain.

4.1 Pearl's Causal Mediation Analysis

The analysis of the causal relationships among variables is an important focus of interest in many empirical studies in the social, behavioral, and health sciences. The target of investigation has been usually represented by the causal effect among these variables, in particular the *total causal effect* of a manipulated variable (or a set of variables) X on a response variable Y. However, the total effect has often been depicted as not adequate to gain a deep understanding of the causal network of relationships among variables. Therefore, the introduction of the concept of *direct causal effect* has arisen new interest in the study of causal relationships. Such direct effect quantifies the sensitivity of the response variable Y to changes in the variable X while all other variables in the analysis are held fixed. Namely, this corresponds to prevent all causal paths from X to Y that are intercepted by intermediate variables, permitting only the direct link $X \to Y$ [124–127, 129], see Fig. 4.1.

This Mediation Analysis framework has been typically used in social sciences researches

as, for instance, in the context of legal disputes over race or sex discrimination in hiring [130]. In such cases, the problem is to establish whether gender or race (X variable) has directly influenced hiring (response variable Y) and not simply indirectly through differences in qualification or experience (intermediate variable, so-called *mediator*, Z). The Mediation Analysis has also been largely used in epidemiology, for instance, to quantify the direct effect of smoking (X) on the incidence of cardiovascular diseases (Y), while taking into account the indirect effect of other aggravating factors, such as hyperlipidemia (Z) [131].

Given the empirical importance of the decomposition of causal effects into their direct and indirect components, a substantial literature on the quantification of direct, indirect and total effects has been published. For the past few decades, structural equation models and linear regression paradigms have been largely used. However, methodologies in the empirical sciences have been restricted to linear analysis [126, 127, 129, 130, 132]. Only recently Judea Pearl, starting from basic principles, redefined causal effects and developed a new method that enables to ride out the limitation to the linear analysis and broaden the application of the decomposition of causal effects to a huge variety of new empirical problems, including the case of categorical data and highly nonlinear processes. However, the limitating assumption of error independence among measured variables remains valid also in this framework. Despite that, the importance of this method lies on the derivation of a general and easy-to-use formula, the so-called "Causal Mediation Formula", for evaluating the extent to which the effect of one variable on another is mediated by a third. Due to the general applicability of Pearl's approach, the Causal Mediation Formula is suitable to deal with nonlinear models, involving both discrete and continuous variables [124–127].

In this section, we will introduce the total, direct and indirect causal effects following Pearl's Mediation Analysis approach, through conceptual and mathematical definitions. Furthermore, we will illustrate the general applicability of the Mediation Analysis formulae in the simple binary case, since its low dimensionality permits to clearly show how causal effects can be directly estimated from the data.

4.1.1 Total, direct and indirect effects

Consider the nonlinear mediation model depicted in Fig.4.1, where X, Y, Z are discrete or continuous random variables, F_1, F_2, and F_3 are arbitrary functions that define the corresponding structural equations for the random variables (*i.e.* $x = F_1(\epsilon_1)$, $z = F_2(x, \epsilon_2)$, $y = F_3(x, z, \epsilon_3)$), and ϵ_1, ϵ_2, ϵ_3 represent error terms, such as noise terms, which are assumed to be mutually independent yet arbitrarily distributed. It is also assumed that the direction of causal influences is known. Starting from this simple mediation model, we will define total, direct and indirect effects through Pearl's Mediation Analysis approach [124–127]. Then, we will show how it is possible to express these causal effects in terms of the available data, assumed to be in the form of random samples (x, y, z) drawn from a joint probability distribution $P(x, y, z)$ [126, 127].

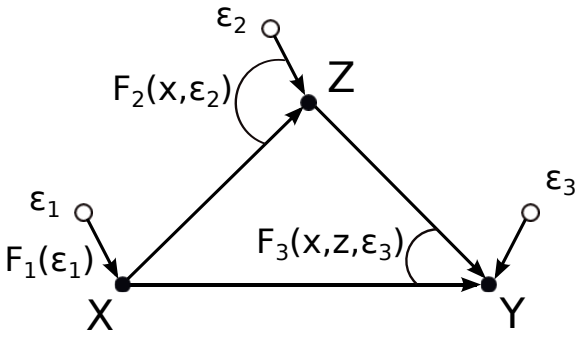

Figure 4.1: A generic mediation model. Generic model depicting mediation from a variable X on a response variable Y through the mediator Z with no confounders (unknown common causes), *i.e.* the error terms are mutually independent. Fig. adapted from [126].

The total effect

The *total effect* is the simplest effect to define and estimate. It is represented by the symbol $TE_{x,x'}$ and measures the change in Y produced by a change in X from $X = x$ to $X = x'$, where x and x' are any two levels of X. It is not necessary to specify the level of Z, since Z is allowed to track the changes in X. So, in the framework of Bayesian statistics [125], it is simply given by the difference in expected values of Y when X is changed from x to x',

$$TE_{x,x'} = E(Y|X = x') - E(Y|X = x). \tag{4.1}$$

The total effect can also be expressed in terms of the $do(x)$ operator – the mathematical operator introduced by Pearl in his *do*-calculus theory (see Appendix B), simulating physical interventions replacing X by a constant $X = x$ while keeping the rest of the model unchanged [124, 133] – as

$$TE_{x,x'} = E(Y|do(x')) - E(Y|do(x)). \tag{4.2}$$

Note that, in nonlinear systems, both the reference level $X = x$ and the final level $X = x'$ may play a role in affecting the change of Y.

The direct effect

The definition of the *direct effect* involves the central concept of "holding the intermediate variables fixed", which has been difficult to express in a proper formalism and has limited the advance in the Mediation Analysis theory. The interpretation of this concept corresponds to (hypothetically) setting the intermediate variables to constants by physical intervention [125, 127]. The main issue emerges when nonlinear systems are considered. Indeed, while in linear systems the direct effect is independent of the levels at which Z is held, in nonlinear systems those values would usually modify the estimation of the direct

effect of X on Y, and therefore they should be chosen carefully to represent the reasons for interest in the analysis. Thus, in the latter case, it becomes more reasonable to consider the direct effect relative to some 'natural' reference level of Z, that may vary depending on the specific variable Z and represents its level just before the change in X [125–127, 131].

Conceptually, the natural *direct effect*, $DE_{x,x'}(Y)$, can be defined as the expected change in Y induced by changing X from x to x' while keeping all mediating factors Z constant at whatever value they *would have obtained* before the transition from x to x' [124, 125, 127, 131],

$$DE_{x,x'}(Y) = E(Y(x', Z(x))) - E(Y(x)). \tag{4.3}$$

This hypothetical change is the real issue in race or sex discrimination cases: "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (In *Carson versus Bethlehem Steel Corp.*, 70 FEP Cases 921, 7th Cir. (1996)) [127]. Specifically, this hypothetical change represents a *counterfactual*, a "what if" question: it begins with an evidence about an existing observed situation and implies a question about an alternative hypothetical world where the past is modified in some way [134]. In general, if x and x' are incompatible, then $Y(x)$ and $Y(x')$ cannot be measured simultaneously and the natural direct effect cannot be empirically estimated. Indeed, it is not possible to rerun history and measure variables response under conditions they have not actually experienced [126].

However, Pearl contributed also to derive conditions under which the natural direct effect can be expressed in terms of the $do(x)$ operator and thus estimated in controlled experiments [124, 125], see Appendix B.1. In particular, it results that in Markovian models (*i.e.*, acyclic models with no unobserved confounders) each *do*-expression can be reduced to a "*do*-free" expression, and thus the natural direct effect is identifiable [124]. For example, for the confounding-free model of Fig. 4.1, it turns out to be

$$DE_{x,x'}(Y) = \sum_z [E(Y|x', z) - E(Y|x, z)]P(z|x). \tag{4.4}$$

Namely, the direct effect corresponds to the average obtained using the pre-transition distribution $P(z|x)$ as a weighting function.

The indirect effect

The concept of *indirect effect* is problematic, since it is impossible to specifically deactivate the direct link from X to Y and let X affect Y exclusively through indirect paths, even controlling any variable [124]. However, a definition of the indirect effect, applicable to empirical data, can be interestingly obtained from the definition of the direct effect.

The *indirect effect*, $IE_{x,x'}(Y)$, of the transition from x to x' is defined as the expected change in Y affected by holding X constant at the reference level $X = x$, and changing

Z to whatever value it *would have attained* had X been set to the final level $X = x'$. Formally, this definition involving counterfactuals reads [125–127],

$$IE_{x,x'}(Y) = E(Y(x, Z(x'))) - E(Y(x)), \tag{4.5}$$

which resembles Eq.(4.3) for the direct effect, but x and x' are exchanged in the first term.

Moreover, it is possible to derive a formula for the indirect effect, following a reasoning analogous to the one that that led to the experimental identification of the direct effect. In particular, for the confounding-free model of Fig. 4.1, we finally obtain,

$$IE_{x,x'}(Y) = \sum_z E(Y|x,z)[P(z|x') - P(z|x)]. \tag{4.6}$$

This is a very general formula to estimate indirect effects, applicable to any nonlinear system and any type of variables. Indeed, due to its generality, Pearl has referred to it as the "Mediation Formula" [124].

Remark. In general, the relationship between the total, direct and indirect effects is non-additive, *i.e.* $TE_{x,x'}(Y) \neq DE_{x,x'}(Y) - IE_{x',x}(Y)$, due to the nonlinear coupling between direct and indirect effects. However, combining Eq.(4.3) and Eq.(4.5), the total effect of a transition from x to x' can be expresses as the *difference* between the direct effect and the indirect effect of the reverse transition obtained exchanging x' and x [124],

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) - IE_{x',x}(Y).$$

Note that, in linear systems, the reverse transition results in a change of the sign of the causal effects, and thus the standard additive formula is recovered.

4.1.2 An example: the simple binary case

As an example of the power of the Mediation Analysis, Pearl has shown how the Mediation Formula of Eq.(4.5) can be applied to categorical variables in nonlinear models [126, 127]. We consider the model of Fig.4.1, in which all error terms are mutually independent. Moreover, we assume that the observed data is given by Table 1 in [127], where n_{xzy} represents the number of observations of a given combination of values of the variables. The expectation values of Y, $E(Y|x,z) = g_{x,z}$, and Z, $E(Z|x) = h_x$, can be easily estimated from the data (*i.e.* the counts n_i), as shown in the two right-most columns, and lead to

$$\begin{aligned} DE &= (g_{10} - g_{00})(1 - h_0) + (g_{11} - g_{01})h_0 \\ IE &= (h_1 - h_0)(g_{01} - g_{00}) \\ TE &= f_1 - f_0 = g_{11}h_1 + g_{10}(1 - h_1) - \big[g_{01}h_0 + g_{00}(1 - h_0)\big] \end{aligned} \tag{4.7}$$

Remark. When the outcome Y is binary, the direct and indirect effects can be interpreted in terms of proportions of the total effect, as necessary and sufficient contributions from the direct and indirect causal pathways. Based on Pearl's definition, a response is "owed" to a path if it would not have occurred were it not for the mechanism represented by that path [126,127]. Therefore, the ratio $(1 - IE/TE)$ represents the contribution that is owed to direct path, while $(1 - DE/TE)$ represents the contribution owed to mediated paths. In particular, these two quantities are not necessarily mutually exclusive. In other words, the direct effect is sufficient as sole cause to account for a proportion DE/TE of the total effect, while, the indirect effect is necessary as complementary cause to account for a proportion $1 - DE/TE$. Vice versa, the indirect effect is sufficient as sole cause to account for a proportion IE/TE, while, the direct effect is necessary as complementary cause to account for the proportion $1 - IE/TE$ of the total effect [126, 127].

4.2 Application of the Mediation Analysis to genomic data

A first application of Pearl's Mediation Analysis binary approach to the study of the causal relationships among genomic properties has been performed in [40]. Singh et al. applied the Mediation Analysis to assess the relative importance, in terms of causal effects, of various genomic properties on the retention of ohnologs. In particular, they aimed at quantifying, through the direct and indirect effects, the critical role of the susceptibility to deleterious mutations *vs* the sensitiveness to dosage balance constraints on the biased retention of human ohnologs. Therefore, they considered three genomic properties: the sensitiveness to dosage balance constraints (Dosage Bal.), the susceptibility to deleterious mutations (Delet. Mut.) and the gene property of being ohnolog (Ohnolog). Gene classes susceptible to deleterious mutations include cancer, Mendelian disease, dominant negative and autoinhibitory genes, while protein complexes and haploinsufficient genes constitute the dosage balanced genes.

For the Mediation Analysis two scenarios are considered, in which each property is interpreted as a binary variable. First, the susceptibility to deleterious mutations is treated as the cause X and its direct and indirect (through the sensitiveness to dosage balance constraints) effects on ohnolog retention are the target of interest. Thus, $X = 1$ implies that the gene is susceptible to deleterious mutations; and $X = 0$, that it is not the case; and likewise for the mediator Z. $Y = 1$ reads that the gene is ohnolog and $Y = 0$ that it is not an ohnolog. Hence, the expected values g_{xz} and h_x of Y and Z are computed from the corresponding binary table analogous to Table 1 in [126] in terms of the number of occurrence of the combinations of the property values among the gene dataset, and finally these quantities are substituted in Eqs. (4.7). Then, in the second scenario the role of X and Z is exchanged and the direct and indirect (through the susceptibility to deleterious mutations) effects of the sensitiveness to dosage balance constraints on ohnolog retention are computed.

Their results are summarized in Fig. 12.1A in [96] and demonstrate that the reten-

tion of ohnologs in the human genome is more directly caused by their susceptibility to deleterious mutations than their interactions within multi-protein complexes.

Our aim is to go further and reach a broader perspective, including in the Causal Mediation Analysis framework the possibility to analyze more than three variables at once. This will enable to consider many genomic properties that could be a priori causally related to ohnolog retention and quantify their causal effects.

4.3 Extensions of the Mediation Analysis to more than three variables

Singh et al. [40], as discussed in the previous section, tried to discriminate, through Pearl's Mediation Analysis of causal effects, the two alternative hypotheses currently proposed to explain the observed biased retention of ohnologs. Their results prompt us to further analyze ohnolog retention in order to gain a more general perspective, guided by the two following objectives. Our first aim is to deduce the causal scenario directly from the data, through a proper causal inference method, instead of testing all different causal hypotheses for ohnolog retention. Indeed, the Mediation Analysis is based on the assumption that the direction of causal relationships is known. Our second aim is to generalize the causal Mediation Analysis method in order to include many variables at once. This will enable to simultaneously consider many genomic properties that could a priori affect ohnolog retention and quantify the causal effects among them.

In order to accomplish our first objective and obtain the causal scenario directly from the data, we need to rely on a causal inference reconstruction approach. In this context, different methods are available and allow to achieve as result the causal graph describing the causal relationships among the variables. Among these methods, the most used are the Bayesian approach (based on a score to compare causal graphs but not suitable for large networks involving many variables) and the constraint-based methods (based on independencies and suitable for large networks, but very sensitive to noise – so common in real data). In our group, a novel inference method to reconstruct causal networks from large scale datasets has been recently developed. This information-theoretic approach combines constraint-based and Bayesian inference methods to reliably infer large causal graphs, despite the presence of inherent sampling noise in finite datasets. In particular, it ascertains structural independencies in causal graphs based on a Bayesian ranking of their most contributing nodes. This is in contrast to classical constraint-based approaches, such as the efficient PC algorithm, which assess structural independencies in arbitrary order of the intervening variables, rendering them prone to spurious conditional independencies. This new method has been proved to be much more robust to sampling noise than classical inference methods [135]. Therefore, in this study we will rely on the causal graph resulting from this novel hybrid approach.

Once the underlying causal graph is known, we can apply the Mediation Analysis

and quantify direct and indirect effects, assessing the sign and the strength of the causal interactions. In this section, we will therefore focus on our second aim, trying to extend the Mediation Analysis framework to the simultaneous presence of more than three variables. Two approaches will be presented in the following. The first one is based on the distinction of the intermediate variables depending on their contribution to the causal effects of X on Y, but has a limited applicability. The second approach will reveal its usefulness when dealing with the general case of partially directed causal graphs resulting from the inference process.

4.3.1 First approach: the distinction between mediators and covariates

If more than one variable belongs to the indirect causal pathways between X and Y, we need to distinguish between different kinds of intermediate variables depending on their contribution to the causal effects of X on Y. The so-called *mediators* are the intermediate variables that, like Z in Fig.4.1, depend on X and thus belong to the indirect causal pathways directed from X to Y. By contrast, *covariates* are common causes of both X and Y and do not depend on X, therefore they contribute in a different way to the causal effects of X on Y. Moreover, attention should be paid to the selection of the intermediate variables contributing to the causal effects of X on Y, due to the possible presence of *colliders*. Colliders are downstream nodes (common effects) of both X and Y and should not be included in the analysis in order to avoid the creation of artificial correlations. Indeed, the knowledge of some information about their values allows to obtain also some information about X and Y, since X and Y are their common causes, and in this way artificial correlations between X and Y might be introduced. For this reason, colliders should not be included in the Mediation Analysis as intermediate variables contributing to the causal effects of X on Y. These conceptual distinctions, guided by the application of Pearl's theory of the *do*-calculus to the Mediation Analysis [124, 127] (see Appendix B.1), are summarized in Fig. 4.2.

Following this approach, the Mediation Analysis formulae for a general free-confounding mediation model become

$$\begin{aligned} TE &= \sum_{\{w\}} \left[E(Y|x',w) - E(Y|x,w)\right] P(w) \\ DE &= \sum_{\{z\},\{w\}} \left[E(Y|x',z,w) - E(Y|x,z,w)\right] P(z|x,w)\, P(w) \end{aligned} \tag{4.8}$$

where $\{w\}$ stands for the values of the set of covariates W and $\{z\}$ for the values of the set of mediators Z.

Although a subset of all the intermediate variables could be selected as a sufficient set for estimating the causal effects of X on Y given the topology of the causal graph (through the back-door criterion, see Appendix B.1), the inclusion of all mediators and

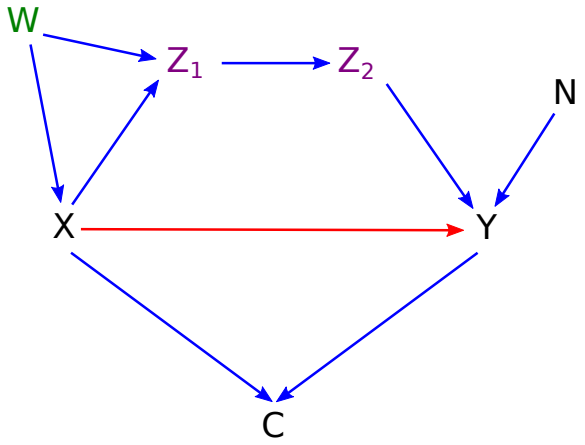

Figure 4.2: Conceptual distinction of the intermediate variables between X and Y. Variables that contribute in a different way to the causal effects of X on Y should be differently treated in the Mediation Analysis. In the figure, Z_1 and Z_2 are *mediators*, intermediate variables that depend on X, while W is a *covariate*, a common cause of X and Y that does not depend on X. By contrast, N and C (a *collider*, a common effect of both X and Y) do not contribute to the causal effect of X on Y and should not be included in the Mediation Analysis.

covariates guarantees the proper evaluation of causal effects in the Mediation Analysis framework. Therefore, in order to efficiently recover the sets of mediators and covariates for each pair of properties in a complex graph, we developed a recursive method based on the idea to iteratively collect the mediators and covariates of each edge of the graph. This iterative algorithm was inspired by the Dijkstra shortest path algorithm [136] and converges in polynomial time. As an example, the pseudocode for finding the set of mediators is detailed in Algorithm 1, and a similar algorithm allows to identify the set of covariates for each edge of the graph.

However, this approach is not suitable for partially directed graphs. Since conditional dependencies in the data only determine the *skeleton* (the undirected graph obtained removing all arrowheads) and the so-called *v-structures* (ordered triplets of vertices containing a collider, *e.g.* $X_i \rightarrow X_j \leftarrow X_k$) of a graph, a partially directed graph is typically obtained as a result of the causal inference process. It contains both directed and undirected edges, and represents an *equivalence class* of directed graphs, all corresponding to the same probability distribution. The absence of orientation for an edge means that it is not possible to infer the information about the direction of the causal relationships from the data through the inference reconstruction process, therefore the two alternative orientations are possible and should be taken into account. In this case, the approach described above could be extended to include undirected edges and then used, for the sake of simplicity, in the two limit cases in which all the undirected edges are assumed to represent either mediators or covariates, in order to limit the number of combinations in the analysis of the possible orientations. Yet, some issues emerge. First, the choice of the orientations of undirected edges is only locally adapted since it depends on which

Algorithm 1 Collect the mediators $med(x, y)$ for each oriented edge (x, y) in a graph G

INITIALIZATION

for all nodes y in G **do**
    $P_y = \emptyset$ {P_y represents the set of parents of y}
    for all nodes $z \neq y$ in G **do**
        $med(z, y) = \emptyset$
        if $z \to y$ **then**
            $P_y = P_y \cup z$
            $med(z, y) = \{z\}$ {$med(z, y)$ is initialized with the parents of y}

ITERATION

flag = **true**
while (flag = **true**) **do**
    flag = **false**
    for all ordered pairs (x, y) in G^2 with $P_y \neq \emptyset$ **do**
        if $x \to y$ **then**
            for all $z \in P_y$ with $z \neq x$ **do**
                if $med(x, z) \neq \emptyset$ **then**
                    if $med(x, z) \not\subseteq med(x, y)$ **then**
                        $med(x, y) = med(x, y) \cup med(x, z)$
                        flag = **true**
                    if $med(z, y) \not\subseteq med(x, y)$ **then**
                        $med(x, y) = med(x, y) \cup med(z, y)$
                        flag = **true**

pair of properties (X,Y) is currently analyzed (*e.g.* the same undirected edge can have an orientation to make the associated variable appear a mediator for given a pair of properties and the opposite orientation to make the same variable appear a mediator as well but for a different pair of properties). Then, only paths with one undirected edge can be identified, since the parents of the response variable Y represent a necessary requirement on the orientations during the iterative collection of mediators and covariates. Moreover, no clear relationship can be found between the direct effect for which the intermediate variables associated to undirected edges are considered as mediators compared to direct effect for which the intermediate variables are considered as covariates. Finally, the potential presence of loops in the graph affects the selection of mediators and covariates, since in this case an intermediate variable could be classified as both mediator and covariate. Therefore, a different approach with a more general applicability will be presented in the next section.

4.3.2 A general approach: the parents of X and Y

In order to reach a more general applicability and analyze through the Mediation Analysis framework also partially directed graphs, a different approach can be followed. Assuming that there are no confounders (unobserved variables creating spurious correlations), we can reduce the set of all intermediate variables to a subset sufficient to properly estimate the causal effects of X on Y, namely the parents of X ($\{pa_X\}$) and the parents of Y excluding X ($\{pa_{Y\setminus X}\}$) [124, 125], see Appendix B.1. Indeed, they directly affect X and Y and enable to directly assess their levels. Therefore, directly applying Pearl's general definition of causal effects [125] to these subsets gives

$$
\begin{aligned}
TE_{x,x'}(Y) &= \sum_{pa_X} [E(Y|x', pa_X) - E(Y|x, pa_X)]\ P(pa_X) \qquad (4.9)\\
DE_{x,x'}(Y) &= \sum_{\{pa_{Y\setminus X}\ \cup\ pa_X\}} [E(Y|x', pa_{Y\setminus X}) - E(Y|x, pa_{Y\setminus X})]\ P(pa_{Y\setminus X}|x, pa_X)\ P(pa_X)\\
IE_{x,x'}(Y) &= \sum_{\{pa_{Y\setminus X}\ \cup\ pa_X\}} E(Y|x, pa_{Y\setminus X})\ [P(pa_{Y\setminus X}|x', pa_X) - P(pa_{Y\setminus X}|x, pa_X)]\ P(pa_X)
\end{aligned}
$$

and it can be verified that $TE_{x,x'}(Y) = DE_{x,x'}(Y) - IE_{x',x}(Y)$.

In case of partially directed graphs, undirected edges may affect the selection of the parents of X and Y, but it is not necessary to compute all the directed graphs in the equivalence class – that becomes unfeasible for large graphs. One only needs to consider all the combinations of orientations of the possible parents of X and Y. The number of these combinations is $2^{n_{pa_{Y\setminus X}} + n_{pa_X}}$, where $n_{pa_{Y\setminus X}}$ and n_{pa_X} correspond to the number of undirected edges involving $Y \setminus X$ or X respectively, and remains tractable for partially directed sparse graphs. Then, propagation rules could be further applied to check for possible incompatibilities (*i.e.* new v-structures) with the equivalence class of the original graph.

Since the graph resulting from the causal inference process is usually partially directed, this general approach will be used to analyze the causal relationships among genomic properties, in order to reveal the direct causes of the reported ohnolog retention during vertebrate evolution. In this case, since genomic properties can assume more than two levels in the available human genes datasets, the formulae for the causal effects will be interpreted for categorical variables, instead of the simple situation of binary variables.

Remark. In this general approach, the possible presence of loops in the graph is not taken into account in the estimation of causal effects, since attention is focused on the direct parents of X and Y, ignoring the full intermediate paths connecting X and Y that could involve loops. Moreover, note that the choice of the orientation for undirected edge is local, since it depends on the specific pair (X, Y) under consideration.

4.4 Relationship to Maathuis's approach

In this section, we will compare our method with the approach developed by Maathuis and coworkers [128], highlighting the deeper understanding of the causal network of relationships that our method allows to obtain.

In [128], Maathuis et al. assumed to have observational data generated from an unknown underlying directed acyclic graph (DAG) model, and combined two processes. First, they aimed at estimating the equivalence class of DAGs from the data using the PC algorithm, since a DAG is typically not identifiable. Then, they estimated causal effects using Pearl's intervention *do*-calculus [124, 133, 137] (see Appendix B). The usefulness of their method rests in its practical applicability to biological problems, since it can score the variables according to their potential effect on the response, allowing the identification of variables that can be tested afterwards in biological experiments. In particular, they computed the total causal effect of each variable X_i on the response variable Y for each DAG in the equivalence class. Among the causal effects obtained for each variable X_i, they selected the minimum absolute value as a lower bound on the size of the causal effect of X_i on Y and used this bound to determine variable importance [128]. The interesting aspect of their approach is that the algorithm they developed uses only *local* information of the estimated CPDAG (completed partially directed acyclic graph). In particular, in order to avoid to compute all DAGs in the equivalence class, they restricted the set of variables used to compute the total effect of X_i on Y to the possible parental sets of X_i, checking that no additional v-structure with X_i as collider is created. This methods allows for efficient computation in large graphs.

The combination of the CPDAG estimation from observed data with causal inference methods originates from the same practical need that motivated our work, that is the problem to estimate causal effects when the graph structure among the (multiple) variables of interest is unknown. As in our case, they assumed no unmeasured confounders for both

the CPDAG and the total effect estimation. In their application to a biological issue, it is reasonable since they observe the expression levels from essentially all genes (although there could be other unobserved aspects of the genome). Our approach for the estimation of causal effects is local, too. However, it goes further their simple expression for the total effect involving only the parents of X, and estimates bounds also for the direct effects. Therefore, it enables to gain a deeper understanding of the causal network of relationships, obtaining information about the decomposition of causal effects that represents the real target of many empirical studies.

Part III

Results

Chapter 5

Population genetics results

The counterintuitive expansion of gene families prone to dominant deleterious mutations, outlined in Sec.2.3.1, prompted us to develop a consistent population genetics model in order to explain this evolutionary oddity (see Ch. 3). We first considered a simple haploid deterministic model in order to explicitly take into account the mode of duplication (WGD *vs* SSD) and the gene propensity to acquire dominant deleterious mutations, see Sec.3.2.1. In this chapter, we will derive the corresponding deterministic solutions, allowing to capture the main evolutionary process responsible for the different retention of SSD *vs* WGD duplicates susceptible to dominant deleterious mutations. Furthermore, the results of the stochastic simulations, introduced in Sec.3.3.2, will enable to investigate the effect of finite population sizes on duplicates retention. Finally, we will compare the theoretical results to the retention biases of SSD *vs* WGD duplicates reported for gene families with oncogenic properties and responsible for a broad range of primary tumors in human.

All in all, the predictions of this population genetics model will support the idea that the enhanced retention of WGD duplicates prone to dominant deleterious mutations is an indirect consequence of the initial speciation event triggered by WGD and the ensuing purifying selection in post-WGD species.

5.1 Deterministic solutions of the haploid model for neutral subfunctionalization

In section 3.2.1 we introduced a simple deterministic haploid model to study the fixation of SSD *vs* WGD duplicates. The advantage of this model relies on the fact that it can be analytically solved and, although simplified, it will give interesting insights into the biased retention of gene duplicates susceptible to dominant deleterious mutations. In particular, we focus on the case of neutral fixable genotypes through subfunctionalization for the duplicated loci, that is the system corresponding to Eqs.(3.2). Indeed, we are assuming that the two fixable genotypes, AA_{-} and $AA_{\star}$, have a neutral fitness, $w_{-} = 1$

and $w_\star = 1$, respectively (*i.e.* AA_+ with $w_+ < 1$ cannot be fixed in the limit of large population, $N \to \infty$). In Eqs.(3.2), $\phi_\circ(t)$, $\phi_+(t)$, $\phi_-(t)$ and $\phi_\star(t)$ represent the fractions of individuals in the population with the corresponding genotypes for the duplicated loci, $AA_\circ$, AA_+, AA_- and $AA_\star$. The solutions for $\phi_\circ(t)$, $\phi_-(t)$ and $\phi_\star(t)$ can be expressed in terms of a time integral $\Phi_+(t)$ of the fraction $\phi_+(t)$ of the population with deleterious gain-of-function mutations at the duplicated loci, as

$$\begin{aligned}
\phi_\circ(t) &= \epsilon e^{-\nu_f t}\Phi_+(t) \\
\phi_-(t) &= \left(\frac{\epsilon\nu_-}{\nu_f}\left(1-e^{-\nu_f t}\right)+1-\epsilon\right)\Phi_+(t) \\
\phi_\star(t) &= \frac{\epsilon\nu_\star}{\nu_f}\left(1-e^{-\nu_f t}\right)\Phi_+(t) \\
\Phi_+(t) &= \exp\left(\int_0^t s_+\phi_+(t')dt'\right)
\end{aligned}$$

where $\nu_f = \nu_+ + \nu_- + \nu_\star$ is the total rate of mutations with functional effect (*i.e.,* gain- or loss-of-function or subfunctionalization) and $\epsilon = \phi_\circ(0)$ is the initial fraction of individuals in the population with duplicated loci, $AA_\circ$. The remaining individuals present only a single functional locus, A, which is assumed to be equivalent to the loss-of-function mutation, AA_-, at either duplicated locus, *i.e.* $\phi_-(0) = 1 - \phi_\circ(0) = 1 - \epsilon$, whereas $\phi_+(0) = 0$ and $\phi_\star(0) = 0$. As a WGD event leads to a concomitant speciation event due to the ploidy incompatibility with pre-WGD individuals, it implies that all individuals of the post-WGD population have a duplicated genome, corresponding to the case $\epsilon = 1$. By contrast, a SSD event does not typically lead to a speciation, leaving a single (or a few) individual(s) with one duplicated gene in the post-SSD population corresponding to $\epsilon \simeq 1/N \ll 1$. Note that ϵ is also the expected fixation rate in absence of mutation, if all fixable genotypes are neutral, $\Pi_e = \epsilon$. Hence, using the asymptotic condition, $\phi_-(\infty) + \phi_\star(\infty) = 1$, for the fractions of individuals with the only fixable genotypes in the large population size limit — corresponding to the loss of one duplicate (AA_-) or the retention of both duplicates through subfunctionalization ($AA_\star$) — we obtain

$$\Phi_+(\infty) = \frac{\nu_f}{\nu_f - \epsilon\nu_+}$$

and thus the asymptotic fraction of subfunctionalized duplicated loci becomes

$$\phi_\star(\infty) = \frac{\epsilon\nu_\star}{\nu_f - \epsilon\nu_+}$$

Note, that the same result is obtained if the fitness parameters are rescaled by the average fitness, $w_i \to w_i/\bar{w}$, which only affects transient regimes but not asymptotic distributions.

For neutral fixable genotypes, AA_- and $AA_\star$ ($w_- = w_\star = 1$), $\phi_\star(\infty)$ corresponds to the expected fixation rate of $AA_\star$ in the population by coalescence, $\Pi_\star = \phi_\star(\infty)$. Thus, we obtain the following expressions for SSD duplicates with $\Pi_e^{\text{SSD}} = \epsilon = 1/N \ll 1$ and

WGD duplicates with $\Pi_e^{\text{WGD}} = \epsilon = 1$,

$$\begin{aligned} \Pi_\star^{\text{SSD}} &\simeq \frac{\nu_\star}{\nu_f}\Pi_e^{\text{SSD}} = \frac{\nu_\star}{\nu_+ + \nu_- + \nu_\star}\Pi_e^{\text{SSD}} \\ \Pi_\star^{\text{WGD}} &= \frac{\nu_\star}{\nu_f - \nu_+}\Pi_e^{\text{WGD}} = \frac{\nu_\star}{\nu_- + \nu_\star}\Pi_e^{\text{WGD}} \end{aligned} \tag{5.1}$$

Hence, the mutation rate ν_+ leading to deleterious phenotypes with decreasing fitness ($w_+ < 1$) favors the elimination of gene duplicates susceptible to dominant deleterious mutations after SSD events, as expected and illustrated in Fig. 4 in [40]. Instead, the same mutation rate ν_+ leading to deleterious phenotypes does *not* appear in the fixation rate of gene duplicates following a WGD-induced speciation event. It implies that the mechanism of purifying selection does *not* contribute to the elimination of gene duplicates susceptible to dominant deleterious mutations in post-WGD populations following a WGD-induced speciation event ($\epsilon = 1$), see Fig. 4 in [40]. This peculiar result is the opposite of what happens in post-SSD populations without speciation ($\epsilon \ll 1$), highlighting the key role of speciation triggered by WGD events during evolution.

From Eq.(5.1), we can directly compare the fixation rate through subfunctionalization of WGD with SSD duplicates,

$$\frac{\Pi_\star^{\text{WGD}}/\Pi_e^{\text{WGD}}}{\Pi_\star^{\text{SSD}}/\Pi_e^{\text{SSD}}} \simeq \frac{\nu_f}{\nu_f - \nu_+} \tag{5.2}$$

obtaining a different fixation of duplicates through WGD and SSD events, which favors the retention of WGD duplicates prone to dominant deleterious (*e.g.* gain-of-function) mutations. Indeed, for genes prone to deleterious gain-of-function mutations ($\nu_+ \gtrsim \nu_- + \nu_\star$, *i.e.* $\nu_f > \nu_f - \nu+$) we find a significantly enhanced retention of duplicates through WGD as compared to SSD events ($\Pi_\star^{\text{WGD}}/\Pi_e^{\text{WGD}} > \Pi_\star^{\text{SSD}}/\Pi_e^{\text{SSD}}$). By contrast, for most genes which lack gain-of-function mutations ($\nu_+ \ll \nu_f$), we find a comparable retention of neutral duplicates through WGD and SSD events ($\Pi_\star^{\text{WGD}}/\Pi_e^{\text{WGD}} \simeq \Pi_\star^{\text{SSD}}/\Pi_e^{\text{SSD}}$). This effect of the WGD-induced speciation on the retention of gene duplicates prone to dominant deleterious mutations, Eq.(5.2), is the main analytical result of this study, which rationalizes from a population genetics perspective the counterintuitive WGD expansion of gene families prone to dominant deleterious mutations during vertebrate evolution.

Remark. If we now assume that the dominance coefficient h, introduced in section 3.2.2 to deal with the more general situation of a diploid population, can be approximated as the average fraction of dominant deleterious mutations (*i.e.* $h \simeq \nu_+/\nu_f$), then Eq.(5.2) leads to the following retention rate of WGD duplicates with dominance coefficient h in diploid genomes,

$$\Pi_\star^{\text{WGD}}(h) \simeq \frac{\Pi_\star^{\text{WGD}}(0)}{1-h} \tag{5.3}$$

This result will be particularly useful for a quantitative comparison of the theoretical predictions of the population genetics model with available experimental data on WGD duplicates of human oncogenes, see Sec. 5.3.

5.2 Analysis of gene duplicates fixation through stochastic simulations

When the population size is small, stochastic effects emerge and it becomes important to analyze how they affect the retention of SSD and WGD duplicates. For this purpose, we introduced in section 3.3.2 stochastic simulations of the birth, death and mutation processes for the population genetics models corresponding to the one-step process master equation formalism detailed in section 3.3.1. In this section, we will present the results of these simulations, focusing first on the fixation rates for neutral subfunctionalization and the stochastic effects emerging especially in the SSD scenario, and then investigating the case of adaptive subfunctionalization, relevant for the fixation of SSD duplicates.

5.2.1 Fixation rates for neutral subfunctionalization

We first performed stochastic simulations to compute the fixation rate of gene duplicates through neutral subfunctionalization ($w_\star = 1$), in order to study the different retention of SSD *vs* WGD duplicates in the cases that, in the large population size limit, correspond to the deterministic system, Eq. (3.2). We analyze the probability of fixation for the allele $AA_\star$ as a function of the ratio ν_+/ν_f, which measures the "dangerousness" of the gene duplicates, that is their susceptibility to dominant deleterious mutations. Fig. 2 in Malaguti et al. [106] shows the results comparing the SSD and WGD scenarios. The simulations are performed for the uncoupled mutation/selection model (since simulations for the coupled dynamics give consistent results but are more time-consuming) and a population size ranging from $N = 10^3$ (violet) to 10^5 (red). For a given ratio ν_+/ν_f, the simulated fixation rate is averaged over 10^2 to 10^4 (WGD) or 10^6 to 10^7 (SSD) fixation trajectories and the standard deviations are shown as error bars. For $Ns_+ \gg 1$ (*i.e.* $N \gg 20$, see next section 5.2.2 on finite size effects), the strong fitness disadvantage s_+ prevents the fixation of the allele AA_+, and leads to an eventual competition between the two neutral alleles AA_- and $AA_\star$. In this situation, the fixation rate $\Pi_\star$ corresponds to the allele frequency in the asymptotic limit, $\phi_\star(\infty)$, as given by Eq. (5.1),

$$\begin{aligned} \frac{\Pi_\star^{\mathrm{WGD}}}{\epsilon} &= \frac{\nu_\star}{\nu_f}\,\frac{1}{1-\nu_+/\nu_f} \\ \frac{\Pi_\star^{\mathrm{SSD}}}{\epsilon} &= \frac{\nu_\star}{\nu_f} \end{aligned} \tag{5.4}$$

As the ratio $\nu_\star/\nu_f$ is kept fixed at the value 0.1, the two theoretical curves for $\Pi_\star^{\mathrm{WGD}}/\epsilon$ and $\Pi_\star^{\mathrm{SSD}}/\epsilon$ become simple functions of the ratio ν_+/ν_f and are plotted as continuous red

lines in Fig. 2 in Malaguti et al. [106].

The comparison between the WGD and SSD scenarios for large population size ($N \geqslant 10^5$) gives interesting insights into the retention of deleterious duplicates prone to gain-of-function mutations. First, the retention of neutral duplicates $AA_\star$ is associated to a low fixation rate for both SSD and WGD duplicates *lacking* dominant deleterious mutations (corresponding to the region $\nu_+/\nu_f \to 0$). Conversely, for gene duplicates *prone* to dominant deleterious mutations (corresponding to $\nu_+/\nu_f \to 1 - \nu_\star/\nu_f$), the retention of neutral duplicates $AA_\star$ is clearly enhanced after WGD events for all population sizes ($N \geqslant 10^3$), while the retention of such deleterious duplicates after SSD events becomes lower than their WGD counterparts for $N \geqslant 10^4$ and reaches a limit independent of ν_+/ν_f for $N \geqslant 10^5$. These results are in agreement with the predictions of the initial simplified deterministic model for large population, Eq. (5.1), and support the idea that WGD events have effectively favored the expansion of gene families prone to dominant deleterious mutations.

Note, however, that the agreement of the asymptotic allele frequency with the fixation rate only holds for large enough population size in the SSD scenario. The discrepancy at lower population sizes is due to finite size effects that allow the initial unstable duplicate $AA_\circ$ to reach fixation by drift before the mutations actually occur, hence, making the duplicates' fixation rate converge towards the WGD scenario. These finite size effects, which are the hallmark of population genetics, are analyzed in more details in the next section.

5.2.2 Finite size effects on the fixation of gene duplicates

The emergence of finite size effects in the fixation rate of SSD duplicates is clearly visible on Fig. 2 in Malaguti et al. [106]. Their interpretation requires, however, a detailed analysis of the consequences of stochastic noise on the evolutionary dynamics of a population of finite size N. We consider separately the WGD and SSD scenarios, below, illustrating the average fixation trajectories in Figs. 3 and 4 in Malaguti et al. [106] for duplicates with a very high susceptibility to dominant deleterious mutations, $\nu_+/\nu_f = 0.825$, to emphasize the different evolutionary scenarios of the proposed population genetics model.

In the WGD case, the effect of stochastic sampling is only visible for a very small population size ($N = 10^2$, Fig. 3 in Malaguti et al. [106]), for which drift can outcompete purifying selection ($Ns_+ \simeq 1$) and results in a non-negligible fixation of the deleterious allele AA_+ (green dotted line) and a simultaneous reduction of the frequencies of the other fixable alleles $AA_\star$ (red dotted line) and AA_- (black dotted line). Then, as the population size increases above $N = 10^3$, the condition $Ns_+ \gg 1$ is always satisfied, leading to the expected fixation rates of the deterministic limit, Fig. 2A in Malaguti et al. [106] (*i.e.* for a negligible fixation of the deleterious allele AA_+). Yet, we expect some additional stochastic effects on the population dynamics due to the discretization of frequencies if $\delta\phi_\pm s_+ \equiv s_+/N \gtrsim \nu_\pm$, that is $s_+ \gtrsim N\nu_\pm$. In practice, this condition delays the transient

dynamics of the simulated trajectories with respect to the deterministic solution (Fig. 3 in Malaguti et al. [106]). However, as N increases from 10^3 to 10^6, the large population condition is more and more verified, $s_+ \ll N\nu_\pm$, leading to average stochastic trajectories (red, blue and black lines) that eventually converge to their deterministic solutions (orange lines), for $N = 10^6$.

In the SSD case, stochastic noise affects not only the transient dynamics but also the fixation rate of SSD duplicates for a wider range of population sizes, Fig. 2B in Malaguti et al. [106]. A detailed analysis based on the comparison with the WGD case is shown in Fig. 4 in Malaguti et al. [106]. In the WGD scenario with $N \geqslant 10^3$, finite size effects affect only the transient dynamics, as discussed above. By contrast, in the SSD scenario, drift caused by stochastic sampling in small population results in the spreading of the initial $AA_\circ$ duplicates to the whole population before they have the chance to mutate into other alleles, leading to a population dynamics after SSD that resembles the WGD scenario, Fig. 4 (top) in Malaguti et al. [106]. This effect is evident and strong for population size $N = 10^3$, where the average simulated trajectories for SSD essentially reduce to the corresponding trajectories for WGD, after proper rescaling by $\epsilon = 1/N$. For increasing population size, this effect weakens and the fixation rates of SSD duplicates become lower than for WGD duplicates for $N \geqslant 10^4$ and eventually reach their asymptotic limit at $N \geqslant 10^5$ for SSD duplicates prone to dominant deleterious mutations, Fig. 2B and Fig. 4 (bottom) in Malaguti et al. [106].

This detailed analysis allows to estimate the conditions for which finite population size becomes relevant for the fixation of duplicates prone to dominant deleterious mutations. In summary, while the transient dynamics of the stochastic trajectories is simply delayed with respect to the deterministic solutions in the WGD scenario, stochastic noise affects also the fixation rate of SSD duplicates for a wide range of population sizes. In particular, it results in decreasing the great gap in the fixation rates of duplicates prone to dominant deleterious mutations between the WGD and SSD scenarios.

5.2.3 Extension to adaptive subfunctionalization for SSD duplicates

The previous sections 5.2.1 and 5.2.2 demonstrate that the fixation of neutral SSD duplicates by drift is at most equal to the initial fraction of SSD duplicates in the population, that is $\Pi_\star^{\text{SSD}} \leqslant \epsilon \simeq 1/N$, which is further reduced to $\Pi_\star^{\text{SSD}} \simeq \nu_\star/(\nu_f N)$ for large population, as the initial $AA_\circ$ duplicates can be lost through loss-of-function or gain-function mutations before they become fixed as $AA_\star$ through subfunctionalization, Fig. 2B in Malaguti et al. [106].

Hence, the fixation of SSD duplicates by drift is clearly inefficient and should be quite rare in large populations [138, 139]. However, beneficial mutations are likely to be particularly important for adaptation [109, 140, 141]. Indeed, it is easy to see that the fixation of SSD duplicates increases rapidly if their retention is associated even to a small fitness benefice ($s_\star > 0$) as shown on Fig. 5 in Malaguti et al. [106]. A sharp rise in the average

fixation trajectories is obtained for increasing values of the fitness parameter from $s_\star = 0$ (black), 10^{-4} (magenta), 10^{-3} (red) to 10^{-2} (blue). This demonstrates that the fixation of SSD duplicates is strongly enhanced under positive selection compared to the low fixation rates of neutral SSD duplicates by drift in large population. Note, in particular, that the fixation rate $\Pi_\star^{\text{SSD}}$ approaches the asymptotic value of the classical two-allele models ($\Pi_\star^{\text{SSD}} = s_\star$, A.3) times the fraction of mutation rates leading to subfunctionalization, *i.e.* $\Pi_\star^{\text{SSD}} \simeq s_\star \times \nu_\star/\nu_f = s_\star/10$. This takes into account the fact that the subfunctionalized duplicates $AA_\star$ arise from the initial redundant duplicates $AA_\circ$ through a mutation rate $\nu_\star$ ten times smaller than ν_f. The slight discrepancy (increasing with increasing $s_\star$) of this estimate from the simulated $AA_\star$ fixation rate (Fig. 5 in Malaguti et al. [106]) is related to the fixation time of a new beneficial mutant, $t_{\text{fix}} \simeq 1/s_\star$. Indeed for increasing $s_\star$, t_{fix} becomes shorter and shorter such that no other $AA_\circ$ individuals, if present, can significantly affect the dynamics of the fixation trajectory, as they are unlikely to experience themselves subfunctionalization mutations before the first $AA_\star$ mutant spreads through the entire population by positive selection. This reduces, in practice, the apparent initial fraction of $AA_\circ$ alleles that effectively contribute to the fixation rate of $AA_\star$ through positive selection. Alternatively, positive selection might also favor the enhanced expression levels of initial SSD duplicates prior to mutations [138,139], leading to the classical result, $\Pi_\star^{\text{SSD}} = s_\star$ (A.3).

These results demonstrate that the fixation of SSD duplicates typically requires positive selection in large populations, while a different mechanism based on purifying selection governs the fixation of deleterious WGD duplicates prone to dominant deleterious mutations following WGD-induced speciation. Besides, as noted earlier [40], we expect that the population bottleneck associated with WGD-induced speciation limits the efficacy of the retention of beneficial WGD duplicates through positive selection.

5.3 Application to the prevalence of human oncogenes with WGD *vs* SSD duplicates

The results obtained from this population genetics model for the fixation rates of gene duplicates can be applied to interpret the reported retention biases of SSD *vs* WGD duplicates prone to dominant deleterious mutations [40, 41, 90], see Sec. 2.3.1. Data on human oncogenes have recently become increasingly available thanks to the numerous cancer genome sequencing projects covering a broad range of primary tumors. Therefore, here, we illustrate the biased retention of WGD duplicates for gene families with oncogenic properties and responsible for a broad range of primary tumors in human.

The datasets of human oncogenes and ohnologs are defined in [106]. In particular, two datasets of oncogenes are obtained from available databases, depending on the criteria to identify the dominance of mutations. As a result, the restricted dataset (Table 1 in Malaguti et al. [106]) and the extended dataset (Table 2 in Malaguti et al. [106]) include

a total of 1,883 and 5,956 oncogene candidates, respectively. Moreover, it results that ohnologs can be classified as either non-SSD ohnolog (5653) or duplicated by SSD (1,422), and non-ohnolog genes either as duplicated by SSD (8,494) or non-SSD genes (4,846).

Both the restricted dataset (Table 1 in Malaguti et al. [106]) and the extended dataset (Table 2 in Malaguti et al. [106]) show that human oncogenes mutated in different primary tumors have indeed retained an excess of ohnologs dating back from the onset of jawed vertebrates. These enhanced ohnolog retentions are highly significant for both datasets as compared to the average retention of ohnologs in the whole human genome, *i.e.* 58.3 % *vs* 34.7 % for the restricted dataset ($P = 3.39 \times 10^{-103}$, χ^2 test, Table 1 in Malaguti et al. [106]) and 48.3 % *vs* 34.7 % for the extended dataset ($P = 4.41 \times 10^{-109}$, χ^2 test, Table 2 in Malaguti et al. [106]). Interestingly, mutated oncogenes from most primary tumors have even higher ohnolog retention biases than the average over all primary tumors, as some ohnolog oncogenes tend to exhibit driver mutations in multiple primary tumors. By contrast, human oncogenes are only slightly depleted in SSD duplicates, as compared to the average SSD retention in the whole human genome, for the restricted dataset, *i.e.* 43.9 % *vs* 48.6 % ($P = 5.35 \times 10^{-5}$, χ^2 test, Table 1 in Malaguti et al. [106]). No significant SSD bias is even observed on the extended dataset, *i.e.* 49.3 % *vs* 48.6 % ($P = 0.29$, χ^2 test, Table 2 in Malaguti et al. [106]).

These results are consistent with the predictions of our evolutionary model, outlined in the present chapter. Indeed, the retention of WGD duplicates should be enhanced for genes prone to dominant deleterious mutations as for the human oncogenes considered in the above datasets, Tables 1 & 2 in Malaguti et al. [106]. Instead, the retention of SSD is predicted to be largely independent of dominant deleterious mutations, requiring instead positive selection of higher expression levels or advantageous mutations, as outlined in the previous section 5.2.3 as well as in earlier studies [138, 139].

In order to be more quantitative in comparing the available experimental data on WGD duplicates of human oncogenes with the predictions of our model, it is necessary to translate the observed fraction of ohnologs, f_s, for a gene class s into an average ohnolog retention rate, p_s, over the two rounds of WGD that occurred at the onset of jawed vertebrates. This can be done through a simple mean field approximation, leading to the following expression, $p_s = 2/f_s - 1 - \sqrt{(2/f_s - 1)^2 - 1}$ [40]. Hence, the observed fraction of ohnologs for human oncogenes, $f_{\text{onc}} = 58.3\%$, corresponds to an average ohnolog retention rate of $p_{\text{onc}} = 21.5\%$ at each round of WGD for the restricted dataset, while the fraction of ohnologs for the extended dataset, $f'_{\text{onc}} = 48.3\%$, corresponds to an average ohnolog retention rate, $p'_{\text{onc}} = 16.3\%$, at each round of WGD. Similarly, the reference over the whole human genome, which corresponds to the observed fraction of ohnologs, $f_{\text{ref}} = 34.7\%$, leads to an average ohnolog retention rate, $p_{\text{ref}} = 10.6\%$, at each round of WGD. This fixation rate of typical duplicates after WGD is in agreement with the solution of the deterministic system of Eqs. (3.2) (see Sec. 5.1, Eq. (5.1) with $\Pi_e^{\text{WGD}} = \epsilon = 1$ and $\nu_f \gg \nu_+$), as assumed in the Model section 3.3.2. Thus, it implies that the observed ohnolog retention bias of human oncogenes (*i.e.* $16.3 - 21.5\%$ *vs* 10.6%) is consistent with an average degree

of dominance, $h \simeq 0.35 - 0.5$, according to Eq. (5.3), *i.e.*, $0.163 - 0.215 \simeq 0.106/(1-h)$.

Chapter 6

Mediation Analysis results

In the previous chapter, the evolutionary mechanism proposed in chapter 3 as responsible for the retention of ohnologs has been analyzed from a population genetics theoretical perspective, showing that the susceptibility to dominant deleterious mutations is a crucial factor for ohnolog retention. In particular, the results predict that the retention of ohnologs is significantly enhanced for genes prone to dominant deleterious mutations, in agreement with the reported bias of WGD *vs* SSD duplicates in human oncogenes (see Chp. 5). Therefore, these population genetics models suggest that the susceptibility of individual genes to deleterious mutations directly underlies their retention after WGD, proposing a consistent explanation to the great expansion of dominant deleterious gene families in vertebrates, outlined in section 2.3.1.

However, an alternative hypothesis has been proposed to explain the differential retention of gene duplicates after SSD and WGD events, the so-called dosage balance hypothesis [41], stating that interacting protein partners participating to the same complex tend to maintain balanced expression levels in the course of evolution. Since duplication of only one interacting partner by SSD leads to dosage imbalance whereas WGD initially preserves correct relative dosage as all genes are duplicated simultaneously, dosage balance constraints have been frequently invoked to explain the biased retention of SSD and WGD genes (see Sec. 2.2.4). Moreover, many genomic properties, such as gene essentiality, duplicability, functional ontology, network connectivity, expression level, mutational robustness, divergence rates, etc., all appear to be correlated to some extent.

Therefore, we aim to further analyze ohnolog retention from a broader perspective and assess the relative contribution of each of these properties to ohnolog retention, quantifying direct causal effects beyond statistical correlations. To this end, we resorted to the Causal Mediation Analysis following the approach of Judea Pearl, introduced in section 4.1, which provides a framework where the relative causal effect of a property can be quantitatively evaluated [124–127]. However, in order to analyze causal relationships in the simultaneous presence of many properties, we needed to extend the Causal Mediation Analysis framework initially developed by Pearl for three properties, as discussed in

section 4.3, proposing a method of general empirical applicability to quantify direct and indirect causes in a network of multiple causally related variables.

In this chapter, we will show the application of our extended Causal Mediation Analysis framework to disentangle the direct from the indirect causal effects of multiple genomic properties on the retention of ohnologs in the human genome. Initially, we will present the available datasets on human genes, discussing genomic properties relevant for ohnolog retention. Then, we will show the causal graph describing the causal relationships among these genomic properties. Finally, the extended Mediation Analysis will be specifically applied to this problem, enabling to quantify direct and indirect effects and ultimately identify the direct causes of ohnolog retention.

6.1 The extended Mediation Analysis on genomic data

In order to identify the direct causes of ohnolog retention during vertebrate evolution, we decided to specifically focus on human. Indeed, genetic data on the human genome have recently become increasingly available thanks to numerous genome sequencing projects, enabling to obtain information about many genomic properties, in particular about abnormal genetic variants associated to diseases. In this section, the selection of genomic properties for the Mediation Analysis will be first explained in the light of their relevance for ohnolog retention. Then, the causal graph describing the causal relationships among these properties, result of the novel inference method developed in our group [135] (see Sec. 4.3), will be presented. Finally, the conditions for the application of the extended Mediation Analysis framework to the causal graph for ohnolog retention will be discussed, detailing how quantifying causal effects from the data.

6.1.1 Genomic properties related to ohnolog retention

To reach a broad perspective in the analysis of the causes of ohnolog retention, we need to consider many properties that could a priori affect, directly or indirectly, the retention of WGD duplicates. Therefore, based on previous studies [41,103,104], we selected nine properties for our analysis. These properties include classes of genes susceptible to deleterious mutations – *i.e.* cancer genes ("Cancer"), genes mutated in other genetic disorders ("Disease"), dominant negative genes for which a mutated allele adversely interferes with the functional allele ("Dominant Negative"), genes with autoinhibitory protein folds ("Autoinhibitory") – , human orthologs of mouse essential genes ("Essential"), genes participating to multiprotein complexes ("Complex"), haploinsufficient genes ("Haploinsufficient"), and gene duplicates coming from either recent SSD ("Young SSD") or WGD ("Ohnolog") events.

The four classes of genes prone to deleterious mutations represent genomic properties that we expect, based on our evolutionary hypothesis, to directly influence the retention of ohnologs. By contrast, genes participating to multiprotein complexes and haploinsufficient

genes – genes for which a single functional copy does not produce sufficient gene product, leading to a disorder, and that are known for their sensitivity to dosage balance constraints [142] – are associated to genomic properties that, according to the dosage-balance hypothesis, should directly drive the retention of ohnolgs.

Moreover, we decided to consider also SSD duplicates originated after the two rounds of WGD at the origin of vertebrates. In this way we avoid the inclusion of SSD duplicates whose retention could be affected by finite size effects due to the population bottlenecks associated to WGD events and resulting in similar fixation rates of SSD and WGD duplicates (see Sec.5.2.2). This will allow to check whether in our analysis the observed antagonist retention pattern of the SSD *vs* WGD mode of duplications could be retrieved.

Finally, essential genes are taken into account to discuss the relationship between essentiality and duplicability. Indeed, Makino et al. [103] observed that essential genes in mouse are only enriched in ohnologs but not SSDs and attributed this differential retention to dosage balance constraints. However, Singh et al. [40] found that, if genes susceptible to deleterious mutations are removed from essential genes, this enrichment vanishes. Moreover, they showed, applying Pearl's Causal Mediation Analysis for one intermediate variable (see Sec.4.1), that the effect of essentiality is mainly indirectly mediated by the susceptibility to deleterious mutations. Therefore, it becomes interesting to evaluate the role of essentiality in the global context of multiple causally-related genomic properties.

In our group, human data about the genomic properties outlined above have been recently obtained [40, 96, 143]. In particular, a comprehensive set of human genes having certain functional properties has been collected through the analysis of multiple databases combined with a direct genome comparison approach to identify gene duplicates. Databases that are regularly updated and containing experimentally verified or manually curated mutated genes have been preferentially used. As a result, 7350 ohnologs, 4464 recent SSDs, 8897 cancer genes, 5171 disease genes, 381 haploinsufficient genes, 565 dominant negative genes, 2937 essential genes, 460 autoinhibitory genes and 6118 genes participating to multiprotein complexes have been obtained.

6.1.2 Inferred causal graph for ohnolog retention

The novel hybrid inference method, recently developed in our group to reconstruct causal networks from large scale datasets (see Sec.4.3), has been applied to the human gene datasets outlined above. The resulting graph, shown in Fig.6.1, describes the causal relationships among the nine genomic properties introduced in the previous section. Arrow's heads indicate the direction of the causal links, while undirected edges represent links for which the direction of the causal relationship could not be deduced from the data. It is worth noting that there is not any causal edge linking gene classes representing the dosage balance hypothesis ("Complex" and "Haploinsufficient") to the retention of ohnologs ("Ohnolog"), suggesting the absence of a direct causal effect of dosage balance constraints on the retention of WGD duplicates.

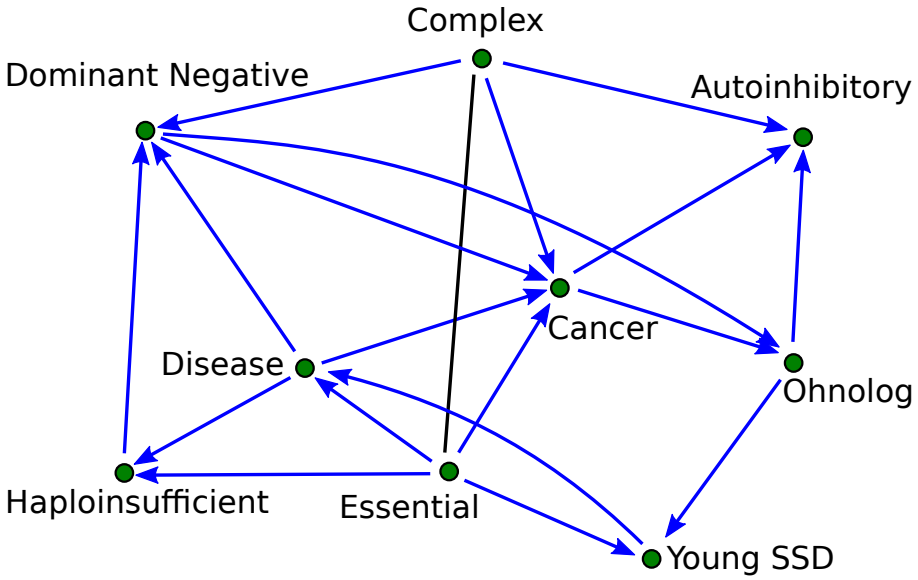

Figure 6.1: Inferred causal graph for ohnolog retention. Causal relationships among nine genomic properties, that have been introduced in section 6.1.1 to study the possible causes of ohnolog retention, are obtained through a novel inference approach (see Sec. 4.3). Arrows' heads indicate the direction of causal relationships, while for undirected edges the direction could not be deduced from the data.

This causal graph represents the starting point for the application of the extended Causal Mediation Analysis framework, that will enable to quantify the magnitude of the total effects obtained through the inference process, their sign (positive or negative, meaning that a property either does or does not favor the occurrence of another property) and the proportion of the direct and indirect causal components.

6.1.3 Application of the extended Mediation Analysis to genomic properties

The extended Mediation Analysis method, proposed in section 4.3.2 to deal with the general case of partially directed graph, can now be applied to analyze the causal effects in the graph of Fig. 6.1. Indeed, the only assumptions are represented by the knowledge of the causal relationships among the variables and the requirement of error independencies. The first assumption is directly encoded in the causal graph. Instead, the latter is a very strong assumption in general. In our case, this assumption may be reasonable since we try to include in the causal graph the genomic properties suspected to be related to ohnolog retention, such that the observed causal effects are the result of underlying causal relationships and not induced by correlations with unobserved variables. The nine genomic properties introduced in Sec. 6.1.1 fulfill this attempt and, although it is not excluded that other properties could eventually affect their network of relationships, the causal graph of Fig. 6.1 can be reasonably considered as Markovian.

For the computation of total, direct and indirect causal effects, each genomic property

is then interpreted as a categorical variable. For most genomic properties, the possible levels are two, corresponding to the presence/absence (Y/N, respectively) of the given property in the observed gene. "Essentiality" has been associated to three levels (Y, NA, N), where the intermediate level represents the absence of information for the observed gene. Similarly, "Cancer"property is classified as N (absence of the property), Putative (candidate cancer gene, since false positives maybe included in the dataset) and Core (cancer gene either experimentally known to be implicated in cancer or are highly mutated from tumor sequencing projects). The level y of the response value Y, measuring the sensitiveness of Y and used to compute the expectation values of Y, $E(y|x, pa_y)$, can be arbitrarily chosen among the levels of Y. In our analysis, we considered y=Y for the binary properties and "Essentiality" while y = "Core" for "Cancer", in order to estimate causal effects leading to the occurrence of the genomic property Y. Similarly, the levels x and x' of the treatment variable X, representing the amount of the change in X that finally affects Y, can be arbitrarily chosen among the levels of X. In our analysis, we considered x=N for all properties and x'=Y for the binary properties and "Essentiality", while x'= "Core" for "Cancer", in order to study causal effects generated by the occurrence of the genomic property X.

Then, in the causal graph, for each pair (X, Y) of properties related by a causal effect, $X \to Y$, the parents of X and Y are selected, eventually considering all the possible combinations of potential parents deriving from undirected edges. For each combination of property levels corresponding to the selected properties, the associated number of genes is computed. The role of this count is equivalent to the counts in Table 1 in [127] for binary variables, and it is finally used to estimate the expectation values in the expressions for the causal effects, Eqs. (4.9).

6.2 Direct causes of ohnolog retention

The results of the application of the generalized Mediation Analysis to the causal graph of Fig.6.1, involving the genomic properties discussed in section 6.1.1, are depicted in Fig.6.2 and detailed in Table 6.1.

In Fig.6.2, the magnitude of the most relevant direct effects ($DE > 0.1$) and the relative proportion of direct effect (DE/TE) are reported next to the corresponding edge. Moreover, different arrow ends are associated with the sign of the total causal effect. In particular, the positive [resp. negative] sign of the total effect means that for the pair of variables $X \to Y$ the variable X does favor [resp. disfavor] the occurrence of the variable Y. In Table 6.1, the amount of the total effect (TE), the direct effect (DE) and the proportion of the relative direct effect (DE/TE) are reported for each edge in the causal graph of Fig. 6.1. When different values of the total effect are obtained for the same pair of properties due to the inclusion of potential parents of X and Y, only the minimum and maximum values of the total effect and the corresponding proportion of direct effect are shown.

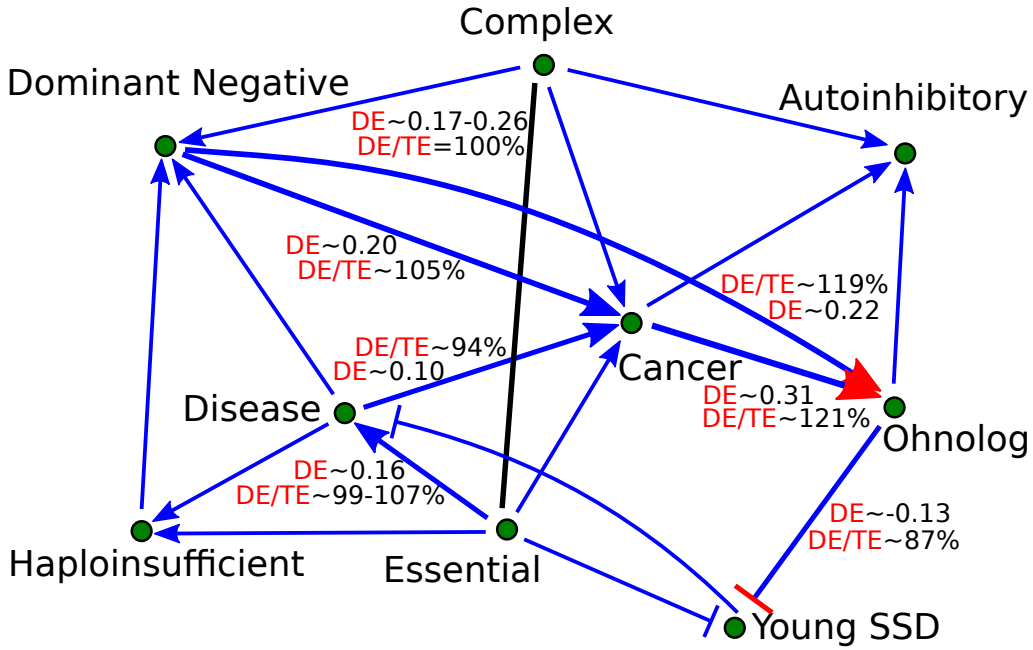

Figure 6.2: Direct causal effects for ohnolog retention. The magnitude of the most relevant direct causal effect ($DE > 0.1$, highlighted by a larger size of the edge) and the relative proportion of direct effect (DE/TE) are reported next to the corresponding edge. Arrow's heads [resp. arrow ending bars] represent positive [resp. negative] causal effects, meaning that for the causal effect between X and Y the variable X does favor [resp. disfavor] the occurrence of the variable Y. The negative sign of the causal effect of the genomic property of "Ohnolog" on "Young SSD" is consistent with the reported antagonist pattern between SSD and WGD duplicates (see Sec. 2.1.3). Most importantly, the genomic properties of being implicated in "Cancer" and "Dominant Negative" mutations, representing the susceptibility to dominant deleterious mutations, have the strongest, mainly direct, (positive) causal effect on "Ohnolog" (red arrow's heads). By contrast, "Complex", the genomic property representing dosage balance constraints, does not show any direct causal effect on "Ohnolog" retention.

A first interesting observation is the negative sign of the causal effect of the genomic property of "Ohnolog" on "Young SSD". Indeed, this result of the generalized Mediation Analysis method is in agreement with the reported antagonist pattern between SSD and WGD duplicates (see Sec. 2.1.3). Moreover, also the negative sign between the genomic properties "Essential" and "Young SSD" can be easily explained, since genes that are critical for survival do not appear in the genome in multiple copies, that could otherwise act as backup for the original copy, rendering it no more essential. Another interesting result is the relative low magnitude of total effects involving "Complex", the genomic property representing dosage balance constraints, except for the strong association "Complex" – "Essential" (see Table 6.1). Hence, the causal effects on other genomic properties due to the participation to complexes are largely mediated by the essentiality of these genes, that is, by the importance of the corresponding complexes for the cellular functions. By contrast, it clearly results that the genomic properties of being implicated in "Cancer" and

Table 6.1: Total and direct causal effects among genomic properties. The amount of the total effect (TE), the direct effect (DE) and the proportion of the relative direct effect (DE/TE) are reported for each edge in the causal graph of Fig. 6.1. The pairs of properties $X \to Y$ are listed in order of decreasing absolute value of the direct effect. When different values of the total effect are obtained for the same pair of properties due to the inclusion of potential parents of X and Y, only the minimum and maximum values of the total effect and the corresponding proportion of direct effect are shown. In particular, the first values are associated with the exclusion of the potential parent, while the second values with its inclusion.

Pair $X \to Y$	**TE**	**DE**	**DE/TE**
Cancer → Ohnolog	0.2536	0.3061	121%
Essential → Complex	0.2601	0.2601	100%
Dominant Negative → Ohnolog	0.1838	0.2180	119%
Dominant Negative → Cancer	0.1934	0.2033	105%
Complex → Essential	0.1737	0.1737	100%
Essential → Disease	0.1564/0.1454	0.1557/0.1557	99/107%
Ohnolog → Young SSD	−0.1472	−0.1287	87%
Disease → Cancer	0.1079	0.1012	94%
Essential → Cancer	0.1153/0.0891	0.0680/0.0682	59/76%
Haploinsufficient → Dominant Negative	0.0506	0.0526	104%
Disease → Dominant Negative	0.0487	0.0499	102%
Essential → Young SSD	$-0.0490/-0.0473$	0.0473/0.0473	97/100%
Essential → Haploinsufficient	0.0660/0.0560	0.0470/0.0470	71/84%
Disease → Haploinsufficient	0.0426	0.0429	101%
Cancer → Autoinhibitory	0.0367	0.0374	102%
Complex → Dominant Negative	0.0456/0.0304	0.0292/0.0301	64/99%
Complex → Cancer	0.0800/0.0454	0.0274/0.0310	34/68%
Ohnolog → Autoinhibitory	0.0263	0.0262	99%
Complex → Autoinhibitory	0.0356/0.0262	0.0236/0.0243	66/93%
Young SSD → Disease	$-8 \cdot 10^{-5}$	0.0114	$> 100\%$

"Dominant Negative" mutations, representing the susceptibility to dominant deleterious mutations, have the strongest (positive) causal effect on "Ohnolog" ($DE \sim 0.30$ and $DE \sim 0.22$, respectively), and this effects are mainly direct ($DE/TE \sim 121\%$ and $DE/TE \sim 119\%$, respectively).

All in all, the results of the generalized Mediation Analysis method allow to identify the direct causes of ohnolog retention, clearly showing that the high gene susceptibility to dominant deleterious mutations, and not dosage balance constraints or essentiality, plays a direct critical role in the retention of WGD duplicates in the human genome. Therefore, these observations allow to discriminate the different hypotheses proposed to explain the antagonist retention of gene duplicates after SSD and WGD events in vertebrates, supporting our hypothesis suggesting that the gene susceptibility to deleterious mutations directly underlies the retention of ohnologs. In conclusion, these results furnish further consistent evidences for the evolutionary mechanism for gene duplicates retention based on the WGD-induced speciation and responsible for the great expansion of gene families prone to dominant deleterious mutations in vertebrates.

Part IV

Discussion & Perspectives

Chapter 7

Discussion and Perspectives

WGD events, despite the huge and immediate changes that entail for an organism experiencing such a genetic accident, occur more frequently than traditionally expected and have now been established in all major eukaryote kingdoms [4,29,34,35,39,75,144,145]. These rare but dramatic evolutionary transitions due to whole genome duplications must have had major consequences on the long time scale evolution of genomes, supplying raw genetic material and providing unique opportunities to evolve and adapt. Moreover, it has become increasingly clear that WGD and SSD events actually lead to the expansion of different functional gene classes in the course of evolution [40–43, 75, 84], highlighting the need to discriminate between different duplication mechanisms while analyzing the evolution by gene duplication.

Furthermore, the bias in the retention of WGD *vs* SSD duplicates is enhanced for genes implicated in genetic diseases and susceptible to deleterious mutations [40, 41, 90, 143]. In particular, Singh et al. [40] have shown that the two rounds of WGD dating back from the onset of vertebrates have effectively favored the expansion of gene families prone to dominant deleterious mutations in the human genome. They also argue that this observed biased retention is a consequence of WGD-induced speciation and subsequent purifying selection in the post WGD population. This interesting perspective becomes critical to explain the apparent counterintuitive expansion of gene families implicated in dominant diseases. Indeed, the expansion of certain gene families prone to acquire dominant deleterious mutations could not be a mere by-product of other presumed advantageous functions. In that case, only the overall benefit of gene family expansion should matter, irrespective of the mechanism of gene duplication, in contradiction with the antagonist retention patterns observed after WGD *vs* SSD duplications. Therefore, it seems likely that the gene susceptibility to dominant deleterious mutations have played a driving role in the striking expansion by WGD of gene families implicated in dominant diseases.

The evolutionary hypothesis proposed by Singh et al. [40] for the biased antagonist retention pattern of WGD *vs* SSD duplicates calls for an evolutionary model in order to rationalize the surprising expansion of gene families prone to dominant deleterious mutations

during vertebrate evolution. To this end, we proposed a consistent population genetics model to analyze the fixation of gene duplicates following either a SSD or a WGD event, taking into account the gene propensity to acquire dominant deleterious mutations. The results support the idea that the enhanced retention of WGD duplicates prone to dominant deleterious mutations is an indirect consequence of the initial speciation events triggered by WGD and the ensuing purifying selection in post-WGD species. Indeed, WGD also induces a speciation event due to the ploidy incompatibility of the post-WGD individuals with the rest of the pre-WGD population. Therefore, through this speciation event, all the genes are already fixed in the new population and during subsequent evolution the loss of WGD duplicates prone to dominant deleterious mutations would be difficult. On the contrary, there is typically no such speciation event coupled to SSD events, and SSD duplicates have to rise in frequency in the population to reach fixation.

Our model rationalizes the surprising expansion during evolution of gene families prone to dominant deleterious mutations, but also predicts that the retention of dominant disease ohnologs remains intrinsically stochastic by nature as many of them are expected to be eliminated. This mainly occurs through loss-of-function mutations before ohnolog pairs could diverge and become non-redundant genes. Indeed, a simple theoretical estimate derived from the long-term retention statistics by Singh et al. [40] shows that only $\sim 10\%$ of the initial ohnolog duplicates have been retained on average at each round of WGD. However, $\sim$ 20%–30% of the initial ohnologs prone to gain-of-function mutations have been retained on average at each WGD. This implies that, although most ohnologs are lost, genes susceptible to deleterious mutations are two to three times more likely to retain ohnologs on long evolutionary timescales. Moreover, this evolutionary trend is emphasized for genes combining several factors associated with the enhanced susceptibility to dominant deleterious mutations, since they have been found to be more than ten times likely to retain ohnologs than genes lacking gain-of-function mutations [40].

Also a small number of SSD duplicates susceptible to dominant deleterious mutations have been fixed in the human genome [40]. In our model, this observation could result from the divergence of the two gene copies through subfunctionalization before one of them is lost by loss-of-function mutations (see Sec. 5.1), from finite size effects emerging during the bottleneck concomitant to WGD-induced speciations (see Sec. 5.2.2) or from positive selection related to possible beneficial mutations or possibly beneficial subfunctionalization (see Sec. 5.2.3).

Moreover, in the quantitative comparison of our model with the observed ohnolog retention bias of human oncogenes, the only adjustable parameter we used to fit the data corresponds to the average degree of dominance of human oncogenes, which is estimated to be in the range of $h \simeq 0.35$–0.5. Although no large scale measurement of the degree of dominance of human oncogenes is currently available from the literature, the inferred estimate seems rather consistent with a number of independent reports on the average and variance of dominance coefficients in other organisms [146–152]. While the reported average degrees of dominance are relatively low (*e.g.*, $h \simeq 0.1$–0.2), their typical distri-

butions appear to be quite broad across gene classes, making our estimate for human oncogenes rather expected for gene classes prone to dominant deleterious mutations (*i.e.*, $h \simeq 0.35$–0.5).

The relationship between gene duplication and speciation events, which is a key component in our model for explaining the evolutionary fate of WGD duplicates, has long been addressed in the literature. In particular, it has been recognized that gene duplicates located at separate loci favor the emergence of new species [107, 108]. This results from a progressive incompatibility between mating partners undergoing reciprocal gene silencing of different duplicate copies, as outlined above in the Section 3.2.2 on the extension to diploid models. In particular, the efficiency of such speciation mechanism is expected to increase with the number of genes simultaneously duplicated in a genome and, therefore, to be most effective after WGD events in the course of evolution [107]. Specifically, such interspecific incompatibilities after WGD are likely responsible for the radiations of species that have been reported in plant genomes, such as in angiosperms at early Cretaceous some 140 MY ago [153], as well as in animal genomes, such as in early jawed vertebrates some 500 MY ago and subsequently in teleost fish some 300 MY ago [154].

By contrast, the reciprocal effect of speciation on the selection of specific gene duplicates, which is at the basis of our modeling approach, has been largely overlooked so far. This is because the fixation of a SSD duplicate is typically thought to be faster than the emergence of a new species, implying that the fixation of single gene duplicates in a population typically precedes speciation events. Yet, it is no longer the case when gene duplicates arise through WGD rather than SSD events [155]. This is because successful WGD are necessarily coupled to a concomitant speciation event, due to the ploidy mismatch between pre- and post-WGD relatives. The subsequent elimination of many WGD duplicates in post-WGD species then unfolds over tens to hundreds millions of years starting from post-WGD populations with already fixed ohnolog duplicates. In particular, this initial fixation of ohnologs is expected to enable the retention of gene duplicates that would have been normally eliminated through purifying selection following an SSD event in the genome of a single individual. This is especially the case for WGD duplicates prone to dominant deleterious mutations, that are expected to be preferentially retained in the genome. Conversely, SSD duplicates of genes prone to dominant deleterious mutations are expected to be eliminated by purifying selection, before they can be fixed in a population.

In particular, we expect that the initial retention bias of ohnologs prone to dominant deleterious mutations due to the WGD-induced speciation effectively promotes, on longer timescales, a prolonged genetic drift and, thus, a progressive functional divergence between ohnolog pairs. This eventually favors the subfunctionalization of ancestral functions [64,67] between ohnolog pairs, which ultimately warrants their long-term maintenance following WGD events. However, this subfunctionalization process might be affected by further deleterious mutations in one of the retained duplicates. On the other hand, a recent study on the vertebrate *Xenopus laevis* [156] suggests that genes retained after WGD are particularly subject to subfunctionalization, thus possibly limiting this effect. Therefore,

it becomes interesting to analyze this potential effect, and our model could be further improved to take into account possible subsequent mutations in the same retained ohnolog pair during evolution.

From a broader context, the selection of gene mutants with slightly deleterious mutations has a long history starting with the nearly neutral theory devised by Ohta [157]. According to the nearly neutral theory, slightly deleterious mutations inevitably accumulate by drift in small populations, thereby, reducing the average fitness and, potentially, the population size itself. This implies that more deleterious mutations might become fixed and, in extreme cases, lead to the extinction of the population through mutational meltdown, for species with less than a few hundreds remaining individuals [158]. However, beyond this accumulation of slightly deleterious mutations, we propose that the specific role of WGD-induced speciation should also be taken into account to interpret the enhanced retention of the WGD duplicates prone to strongly deleterious mutations with dominant phenotypes. This suggests that not only slightly deleterious but also strongly deleterious mutations have impacted the long-term evolution and organismal complexity of vertebrates following their early two rounds of WGD. For this reason, we have specifically included in our model strongly deleterious gain-of-function mutations, unlike classical models in population genetics, traditionally focusing on loss-of-function and beneficial mutations [6, 7, 53, 54, 64, 141].

Some previous studies have alluded to the origin of Mendelian disease genes [41, 90, 143] by WGD. In addition, Singh et al. [40] have also further analyzed cancer genes and oncogenes, showing that also dominant cancer genes in human present a very strong retention bias after WGD. To many, these findings might seem surprising, as the role of cancer mutations in evolution is usually considered to be irrelevant because these mutations are mostly somatic and are believed to typically occur later the reproductive periods of extant animals. However, even though most cancers indeed occur later in the life of an organism, cancer is also one of the leading cause of non-accidental death, even in young adults (see Table 6 in [159]). For example, cancer is the first leading cause of natural deaths in both males and females of age $1 - 19$ and $20 - 39$ in the Unites States of America, according to the cancer mortality statistics in 2009 [159]. Similarly, cancer is expected to be the second cause of natural death after infectious diseases in less developed countries. This underlies the non-negligible incidence of cancers also in juvenile and young organisms and its potential effects on their long-term evolution. Moreover, despite that the occurrence of cancer has certainly increased in the modern times due to a variety of factors, there is no doubt that cancer affected human also in the antiquity [160–163]. Likewise cancer is known to affect other multicellular organisms, including basal metazoan [164].

Applying our extension of Pearl's Causal Mediation Analysis, we have also found further evidences suggesting that, among many genomic properties, the high gene susceptibility to dominant deleterious mutations plays a direct role on the retention of ohnologs. This advanced inference approach becomes of critical importance in the study of ohnolog retention from a broader perspective. Indeed, a number of studies have shown that many

genomic properties appear to be correlated to some extent [41, 103, 104]. However, many of these statistically significant correlations are suspected to mainly result from indirect rather than direct associations [40]. For instance, Singh et al. [40] have shown that the retention of many ohnologs, supposed to be under dosage balance constraints [41], is in fact indirectly mediated by their susceptibility to deleterious mutations. In this general context, the extension of Pearl's Mediation Analysis reveals its usefulness in quantifying direct and indirect causal effects among multiple properties, enabling to discriminate different causal hypotheses. Moreover, this method not only allows to possibly confirm observed associations among genomic properties, but also predict the causal role of unexpected variables that could be eventually experimentally tested.

Our extension of the Causal Mediation Analysis is in full continuity with Pearl's attempt to go beyond traditional methods in the analysis of mediation, which has long been a complex issue in the empirical sciences and has been limited to linear models [129,130,165–167], and to cross the linear-to-nonlinear barrier [126,127]. Pearl's recent method, based on the same causal assumptions that support the standard linear analysis but redefining and deriving causal effects from first principles, avails mediation analysis to a large space of new applications, especially those involving categorical data and highly nonlinear processes [126, 127]. Indeed, the power of the Mediation Formula has been largely recognized [168–172], confirming it as a powerful tool for the assessment of causal pathways in many of the social, behavioral and health-related sciences.

Pearl's Mediation Analysis, however, has been initially formulated only for one mediator with a simple generalization to multiple causally *independent* mediators, since the case of multiple interconnected mediators, affected by an intricate network of observed and unobserved confounders, must be handled with some care [126, 127]. Nevertheless, this is a common scenario when dealing with experimental data involving many variables. Therefore, efforts have been made to extend the Mediation Analysis, starting from the simple case of several causally-*ordered* mediators [173]. Moreover, it is usually assumed that the data are generated by a directed acyclic graph, which codes causal relationships between variables and is known beforehand, and interest has been focused on the estimation path-specific effects, that is the effect of the treatment on the outcome variable through a selected set of paths with all other paths deactivated [174, 175].

However, we are specifically interested in disentangling the direct effect from all the indirect effects of the treatment on the outcome variable in the presence of several causally-related variables, rather than path-specific effects. This is motivated by the practical need to uncover direct causal relationships among multiple correlated variables in observational data. From this perspective, complications due to multiple causally-related intermediate variables can be handled, and simple formulae to quantify causal effects and their direct component can be obtained. Moreover, our method enables to obtain causal information also in the realistic scenario in which the direction of some causal relationships among variables is not known. Therefore, this approach, combined with the causal inference method recently developed in our group, can be applied to a wide range of problems

aiming at the identification of direct causes in experimental observational data, when the causal underlying graph is unknown.

In this context, also Maathuis et al. [128], motivated by the biological problem of identifying the genes that play a role in the riboflavin production in the bacterium *Bacillus subtilis*, aimed at estimating causal effects from observational data. Since in their case, as in many practical problems, the graph describing the causal relationships among the variables is unknown, they combined the graph estimation with the computation of causal effects, similarly to our procedure. However, they limited their analysis to the estimation of *total* causal effects through Pearl's *do*-calculus, the statistical framework at the basis of the Mediation Analysis. Conversely, our work belongs to the attempt of fully exploiting the power of Pearl's Mediation Analysis, oriented at the quantification of the finer components of causal effects.

Our extension of Pearl's Mediation Analysis is particularly useful to uncover direct causal relationships among multiple correlated variables in observational data, but some aspects could be further improved. First, our method is based on the assumption that the underlying causal graph is known, through causal inference methods. When the causal graph is only partially oriented, we consider all the possible combinations of orientation of the undirected edges pointing to the treatment X and the outcome Y. However, unlike Maathuis et al. [128], for the sake of simplicity, we do not check whether additional v-structures are created, leading to a situation incompatible with the equivalence class represented by the graph. This aspect could be taken into account by checking that no additional v-structures is generated with the treatment variable X as collider, then the orientations need to be propagated to the response variable Y, and the resulting constraints on Y have to be analyzed. This process of propagation of information is deterministic and should not alter the efficiency of the algorithm, mainly limited by the computation of all the possible combinations of orientations involving X or Y directly. Moreover, particular attention should be paid to the presence of loops in the causal graph, arising from the hybrid nature (not purely Bayesian, but also constraint-based) of the inference approach developed in our group. Therefore, the effect of loops on the estimation of causal effects need to be investigated. Finally, our extension of Pearl's Mediation Analysis still lacks a proper analysis of the variability of the estimated causal effects. This could be obtained through simple bootstrapping methods, as performed by Maathuis et al. [128], or relying on complex sensitivity analysis to quantify the robustness of the results in relation to the possible existence of unmeasured variables affecting the correlations among the measured variables, as proposed by Imai et al. [132].

In conclusion, we present a consistent population genetics model to rationalize, from an evolutionary perspective, the surprising accumulation of WGD and not SSD duplicates in gene families frequently implicated in genetic disorders and cancers. Thus, our analysis provides a theoretical rationale linking the mutational effects of gene classes prone to dominant deleterious mutations at vastly different time scales, from the effect of somatic mutations in tumor progression to the long-term evolution of vertebrate genomes through

germline mutations and purifying selection in post-WGD species since early vertebrates. In this framework, we stress the importance of considering also the effect of WGD-induced speciation on the selection of specific gene duplicates and not only the reciprocal long studied effect of gene duplicates located at separate loci for the emergence of new species. In particular, we highlight the critical role of purifying selection after WGD events, credited for the successfull radiation of vertebrate species, on the evolution of vertebrates and, beyond, exemplify the role of non-adaptive forces on the emergence of eukaryote complexity. Moreover, we present an extension of wide empirical applicability of Pearl's Mediation Analysis theory, in order to uncover direct causal relationships among multiple correlated variables in observational data. Based on the results of its application to the available data about the retention of ohnologs in the human genome, we further highlight the direct critical role of the gene susceptibility to dominant deleterious mutations on the retention of ohnologs, consistently with the evolutionary mechanism rationalized through our population genetics model.

Part V

Appendices

Appendix A

Population genetics: a general stochastic approach

In this chapter, we will detail the general master equation formalism that enables to properly analyze the discrete allele frequency states of a finite population, allowing to go beyond the simple deterministic model introduced in section 3.2.1 and consider the stochastic effects emerging from finite population size, see Sec. 3.3. In particular, it will allow to include either the coupled or uncoupled mutation/selection dynamics traditionally used in the literature. We will further apply this general approach to the four alleles introduced in section 3.2.1, showing that it leads to the same approximate deterministic equation system between the four alleles for both the coupled and uncoupled dynamics considered and, therefore, to very close deterministic solutions, in the limit of small fitness decrement caused by deleterious mutations, $s_+ \ll 1$. Finally, although the solutions of the master equations are not accessible analytically in the case of four alleles, the corresponding stochastic system with only two alleles can be solved exactly, and it is possible to retrieve classical results of the Wright-Fisher model [176] as approximations.

A.1 General stochastic models using a master equation

We present here a general approach to describe the stochastic dynamics of a population of fixed size N, based on a generic one-step process master equation for K alleles $A_1,\cdots,A_K$ ($K \geqslant 2$) governing the probability, $P(n_1,\cdots,n_K,t)$, of observing n_i individuals with allele A_i at time t (with $\sum_i n_i = N$), as

$$\frac{\partial P}{\partial t} = \sum_{i,j=1}^{K} (\mathbb{E}_i^{-1}\mathbb{E}_j^1 - 1) W_{ij}(\{n_k\}) P(\{n_k\},t) \tag{A.1}$$

where $\mathbb{E}^{\pm 1}$ is the "step operator" [117] such that $\mathbb{E}_i^{\pm 1} f(n_i) = f(n_i \pm 1)$ and $W_{ij}(\{n_k\})$ is the transition rate from allele j to allele i, which can be expressed in terms of the numbers

of individuals with the different alleles as,

$$W_{ij}(n_1,\cdots,n_K) = \frac{n_j}{N}\sum_k \beta_{ik}^{(j)} n_k \tag{A.2}$$

where n_j/N is the probability that one individual with allele j is randomly chosen to die, while $\beta_{ik}^{(j)} n_k$ is the rate at which one individual with allele k is chosen to reproduce and mutate into allele i, given that an individual with allele j has been chosen to die. In particular, this general expression enables to include either coupled or uncoupled mutation/selection dynamics depending on the definition of the transition rates $\beta_{ik}^{(j)}$, see below.

Following the van Kampen's expansion [117], we apply the following transformation $n_i = N\phi_i + N^{\frac{1}{2}}\xi_i$ to the master equation, where $\phi_i(t)$ correspond to the noiseless deterministic solutions of the dynamics in the large population size limit $N \gg 1$, while the new variables ξ_i, which will replace n_i in the master equation, correspond to the stochastic noise in n_i for a finite size population. Accordingly, the distribution $P(n_1,\cdots,n_K,t)$ is now written as a function of ξ_i as, $P(n_1,\cdots,n_K,t) = \Pi(\xi_1,\cdots,\xi_K,t)$. The one-step operator $\mathbb{E}_i^{\pm 1}$ changes n_i into $n_i \pm 1$ and therefore ξ_i into $\xi_i \pm N^{-1/2}$, so that

$$\mathbb{E}_i^{\pm 1} = 1 \pm N^{-1/2}\frac{\partial}{\partial \xi_i} + \frac{1}{2}N^{-1}\frac{\partial^2}{\partial^2 \xi_i} \pm \cdots$$

while the time derivative $\partial_t P(n_1,\cdots,n_K,t)$ is taken with constant n_i, leading to

$$\frac{\partial P}{\partial t} = \frac{\partial \Pi}{\partial t} - N^{\frac{1}{2}}\sum_i \frac{d\phi_i}{dt}\frac{\partial \Pi}{\partial \xi_i}$$

Hence the master equation in the new variables ξ_i takes the form of an expansion in $N^{-1/2}$,

$$\begin{aligned}
&\frac{\partial \Pi}{\partial t} - N^{\frac{1}{2}}\sum_i \frac{d\phi_i}{dt}\frac{\partial \Pi}{\partial \xi_i} \\
&\quad = \sum_{i,j=1}^{K}\Big[N^{-\frac{1}{2}}\Big(\frac{\partial}{\partial \xi_j} - \frac{\partial}{\partial \xi_i}\Big) + \frac{1}{2}N^{-1}\Big(\frac{\partial}{\partial \xi_j} - \frac{\partial}{\partial \xi_i}\Big)^2 + \cdots\Big] \\
&\qquad \Big[N\phi_j\sum_k \beta_{ik}^{(j)}\phi_k + N^{\frac{1}{2}}(\phi_j\sum_k \beta_{ik}^{(j)}\xi_k + \xi_j\sum_k \beta_{ik}^{(j)}\phi_k) \\
&\qquad + \xi_j\sum_k \beta_{ik}^{(j)}\xi_k\Big]\Pi
\end{aligned}$$

The largest terms of order $N^{\frac{1}{2}}$ cancel each other out if $\phi_i(t)$ are taken as the solutions of the deterministic equations,

$$\begin{aligned}
\frac{d\phi_i}{dt} &= \phi_i\sum_k(\beta_{ii}^{(k)} - \beta_{kk}^{(i)})\phi_k - \phi_i\sum_{l;k\neq l}\beta_{lk}^{(i)}\phi_k \\
&\quad + \sum_{j,k\neq i}\phi_j\beta_{ik}^{(j)}\phi_k
\end{aligned}$$

which leads to the following deterministic equations in the three main population genetics models described in the literature,

1. The first Moran model [116] of coupled mutation/selection processes with mutations occurring *before* selection, which is assumed to control death rates. This is a selection on the lifespan of adults rather than their reproductive success,

$$\begin{aligned}
\beta_{ik}^{(j)} &= \lambda_j \nu_{ik}, \qquad \text{for} \quad k \neq i, \\
\beta_{ii}^{(j)} &= \lambda_j (1 - \sum_{l \neq i} \nu_{li}), \\
\frac{d\phi_i}{dt} &= \phi_i \left(\sum_k \lambda_k \phi_k - \lambda_i \right) \\
&\quad - \phi_i \sum_{l \neq i; k} \nu_{li} \lambda_k \phi_k + \sum_{j,k \neq i} \phi_j \lambda_j \nu_{ik} \phi_k
\end{aligned}$$

   where ν_{ik} corresponds to the probability to experience a mutation from k to i at the time scale of death rate λ_j.

2. The second Moran model [116] of coupled mutation/selection processes with mutations occurring *after* selection, which is assumed to control birth rates. This is a gametic selection with a death independent rate $\beta_{ik}^{(j)} \equiv \beta_{ik}$,

$$\begin{aligned}
\beta_{ik}^{(j)} &\equiv \beta_{ik} = \nu_{ik} w_k, \qquad \text{for} \quad k \neq i \\
\beta_{ii}^{(j)} &\equiv \beta_{ii} = (1 - \sum_{l \neq i} \nu_{li}) w_i, \\
\frac{d\phi_i}{dt} &= \phi_i \left(w_i - \sum_k w_k \phi_k \right) \\
&\quad - \phi_i w_i \sum_{l \neq i} \nu_{li} + \sum_{k \neq i} \nu_{ik} w_k \phi_k
\end{aligned}$$

   where ν_{ik} corresponds to the probability to experience a mutation from k to i at the time scale of birth rate w_k.

3. The case of uncoupled mutation/selection, outlined in section 3.2.1, which amounts to use "average" mutation rates $\bar{\nu}_{ij}$ and "average" selection rates $\bar{w}_i$ and $\bar{w}^{(i)}$ as

model parameters, as frequently used in recent years [118–122],

$$\begin{aligned}
\bar{\nu}_{ij} &= \sum_{k\neq i} \beta_{ik}^{(j)} \phi_k, \qquad \text{for} \quad j \neq i \\
\bar{\nu}_{ii} &= 0, \\
\bar{w}_i &= \sum_j \beta_{ii}^{(j)} \phi_j, \\
\bar{w}^{(i)} &= \sum_k \beta_{kk}^{(i)} \phi_k \\
\frac{d\phi_i}{dt} &= \phi_i(\bar{w}_i - \bar{w}^{(i)}) - \phi_i \sum_l \bar{\nu}_{li} + \sum_j \bar{\nu}_{ij}\phi_j
\end{aligned}$$

Therefore, the stochastic approach based on the one-step master equation formalism described by Eqs. (A.1,A.2) enables to obtain a general expression including either coupled or uncoupled mutation/selection dynamics depending on the definition of the transition rates β_{ik}.

A.2 Four-allele deterministic models of SSD *vs* WGD duplicates retention

Here, we apply the three mutation/selection deterministic models defined in the previous section A.1 to study the fixation of gene duplicates following either a SSD or a WGD event. We consider the four different genotypes described in section 3.2.1: the initial (unstable) duplicates, $AA_\circ$, and the three alleles arising by mutation from $AA_\circ$, *i.e.* $AA_- \equiv A$, AA_+ and $AA_\star$. The mutations with functional effect are therefore occurring from allele $j = \circ$ to $i = +, -, \star$ with probabilities (or rates) $\nu_{ij} = \nu_{i\circ(i\neq\circ)} = \nu_i$ for $i = +, -, \star$. For the first Moran model where mutations occur before selection, the deterministic system of equations becomes

$$\begin{aligned}
d_t\phi_\circ &= \phi_\circ(\bar{\lambda} - \lambda_\circ) - \nu_f\bar{\lambda}\phi_\circ \\
d_t\phi_- &= \phi_-(\bar{\lambda} - \lambda_-) + \nu_-\bar{\lambda}\phi_\circ \\
d_t\phi_+ &= \phi_+(\bar{\lambda} - \lambda_+) + \nu_+\bar{\lambda}\phi_\circ \\
d_t\phi_\star &= \phi_\star(\bar{\lambda} - \lambda_\star) + \nu_\star\bar{\lambda}\phi_\circ,
\end{aligned}$$

where $\bar{\lambda} = \sum_k \lambda_k\phi_k$. For the second Moran model where mutations occur after selection,

$$\begin{aligned}
d_t\phi_\circ &= \phi_\circ(w_\circ - \bar{w}) - \nu_f w_\circ\phi_\circ \\
d_t\phi_- &= \phi_-(w_- - \bar{w}) + \nu_- w_\circ\phi_\circ \\
d_t\phi_+ &= \phi_+(w_+ - \bar{w}) + \nu_+ w_\circ\phi_\circ \\
d_t\phi_\star &= \phi_\star(w_\star - \bar{w}) + \nu_\star w_\circ\phi_\circ,
\end{aligned}$$

where $\bar{w} = \sum_k w_k \phi_k$. For the case of uncoupled selection/mutation we retrieve the initial uncoupled mutation/selection dynamics of Eq. (3.1),

$$\begin{aligned} d_t\phi_\circ &= \phi_\circ(w_\circ - \bar{w}) - \nu_f\phi_\circ \\ d_t\phi_- &= \phi_-(w_- - \bar{w}) + \nu_-\phi_\circ \\ d_t\phi_+ &= \phi_+(w_+ - \bar{w}) + \nu_+\phi_\circ \\ d_t\phi_\star &= \phi_\star(w_\star - \bar{w}) + \nu_\star\phi_\circ \end{aligned}$$

The structure of these equations is the same for all models and the differences come from the specific choices of the parameters in the transition rates. The two Moran models can be compared using $\lambda_k = w_k^{-1}$.

In the limit of small fitness decrement caused by deleterious mutations, $s_+ \ll 1$, all three models lead to the same approximate equation system between the four alleles and, therefore, to very close deterministic solutions. However, the stochastic effects encompassed in the full distribution, solution of the master equation, are not accessible analytically in the case of four alleles. Yet, the corresponding population genetics system with only two alleles can be solved exactly, as first shown in [116], and can bring insights on the competition between the two main fixable alleles of our four-allele system (*i.e.* AA_- and $AA_\star$) as shown in the next section.

A.3 Exact results for two-allele stochastic models

We consider the continuous time one-step process master equation for death-birth and mutation stochastic processes between only two alleles A and a in a haploid population of fixed size N, with n A-individuals and $(N-n)$ a-individuals. This equation does not include any approximation, unlike the diffusion equation that is valid in the limit of large populations and small selection, and allows to obtain an exact analytical solution in terms of hypergeometric functions for any values of the model parameters. Moreover, we will show below that it is possible to retrieve classical results of the Wright-Fisher model [176] as approximations.

The one-step transition rates correspond to the probability density for the system to change its number of individuals with allele A from n to $n+1$ or $n-1$ during an infinitesimal time dt,

$$\begin{aligned} W(n \to n+1) &= W^+(n) \\ W(n \to n-1) &= W^-(n) \end{aligned}$$

while $W(n \to n \pm k) = 0$ if $|k| > 1$. The master equation governing the probability, $P(n,t)$,

of observing n individuals with allele A at time t, is given by

$$\begin{aligned}\partial_t P(n,t) &= (\mathbb{E}^{-1}-1)W^+(n)P(n,t) + (\mathbb{E}^{1}-1)W^-(n)P(n,t)\\ &= W^+(n-1)P(n-1,t) - W^+(n)P(n,t)\\ &\quad + W^-(n+1)P(n+1,t) - W^-(n)P(n,t)\end{aligned}$$

where $\mathbb{E}^{\pm 1}$ is the one-step operator. Using the following transition rates, $W^{\pm}(n)$, for the three models outlined above, we obtain,

1. For the first Moran model [116] of coupled mutation/selection with mutations occurring before selection which controls death rates,

$$\begin{aligned}W^+(n) &= \mu\,\frac{(N-n)}{N}\Big[n(1-\nu_1)+(N-n)\nu_2\Big]\\ W^-(n) &= \mu\,\frac{n}{N}\frac{1}{(1+s)}\Big[(N-n)(1-\nu_2)+n\nu_1\Big]\end{aligned}$$

   where μ is the equal birth-death rate of each allele and ν_1 [resp. ν_2] the mutation probability from allele A to a [resp. from a to A]. The factor $1/(1+s)$ implies a reduced ($s>0$) or enhanced ($s<0$) death rate of allele A.

2. For the second Moran model [116] of coupled mutation/selection with mutations occurring after selection which controls birth rates,

$$\begin{aligned}W^+(n) &= \mu\,\frac{(N-n)}{N}\Big[n(1+s)(1-\nu_1)+(N-n)\nu_2\Big]\\ W^-(n) &= \mu\,\frac{n}{N}\Big[(N-n)(1-\nu_2)+n(1+s)\nu_1\Big]\end{aligned}$$

   where $1+s$ is the gain ($s>0$) or loss ($s<0$) of reproductive success of allele A.

3. For the uncoupled mutation/selection model with averaged transition parameters outlined above,

$$\begin{aligned}W^+(n) &= \mu\,(N-n)\frac{n}{N}(1+s)+(N-n)u_2\\ W^-(n) &= \mu\,n\frac{(N-n)}{N}+nu_1\end{aligned}$$

   where u_1 [resp. u_2] is the mutation rate from allele A to a [resp. from allele a to A].

Introducing the rescaled mutation rates $\nu_1 = u_1/\mu$ and $\nu_2 = u_2/\mu$ for the uncoupled mutation/selection model leads to a common form for all three models,

$$\begin{aligned}W^+(n) &= \mu\,\frac{(N-n)}{N}\Big(nA^+ + B^+\Big)\\ W^-(n) &= \mu\,\frac{n}{N}\Big(nA^- + B^-\Big)\end{aligned}$$

where $A^+ = 1-\nu_1-\nu_2$, $B^+ = N\nu_2$, $A^- = -(1-\nu_1-\nu_2)/(1+s)$, $B^- = N(1-\nu_2)/(1+s)$, for the first Moran model; $A^+ = (1-\nu_1)(1+s)-\nu_2$, $B^+ = N\nu_2$, $A^- = -(1-\nu_1(1+s)-\nu_2)$,

$B^- = N(1-\nu_2)$, for the second Moran model and $A^+ = 1+s$, $B^+ = N\nu_2$, $A^- = -1$, $B^- = N(1+\nu_1)$, for the uncoupled mutation/selection model.

The corresponding master equation can then be solved by introducing the generating function,

$$\phi(z,t) \;=\; \sum_n z^n P(n,t),$$

which leads to the following differential equation (using the boundary conditions $W^+(N) = W^-(0) = 0$),

$$\partial_t \phi(z,t) = (z-1)\sum_n W^+(n) z^n P(n,t) + (z^{-1}-1)\sum_n W^-(n) z^n P(n,t)$$

$$\begin{aligned}
\frac{N}{\mu}\partial_t\phi \;&= (z-1)\left(A^+ z\partial_z(N\phi - z\partial_z\phi) + B^+(N\phi - z\partial_z\phi)\right) \\
&\quad + (z^{-1}-1)\left(A^- z\partial_z(z\partial_z\phi) + B^- z\partial_z\phi\right) \\
&= (z-1)\Big[A^+\left((N-1)z\partial_z\phi - z^2\partial_z^2\phi\right) \\
&\quad + B^+(N\phi - z\partial_z\phi) - A^-\left(\partial_z\phi + z\partial_z^2\phi\right) - B^-\partial_z\phi\Big] \\
&= (z-1)\Big[\left(-z^2A^+ - zA^-\right)\partial_z^2\phi \\
&\quad + \left((A^+(N-1) - B^+)z - A^- - B^-\right)\partial_z\phi + B^+N\phi\Big]
\end{aligned}$$

The stationary solutions correspond to the following homogeneous second order ordinary differential equation

$$\left(-z^2A^+ - zA^-\right)\partial_z^2\phi + \Big[A^+(N-1) - B^+)z - A^- - B^-\Big]\partial_z\phi + B^+N\phi = 0$$

which can be transformed into the hypergeometric differential equation through the rescaling $z \to -zA^-/A^+$, see [177],

$$z(z-1)\partial_z^2\phi + \left((\alpha+\beta+1)z - \gamma\right)\partial_z\phi + \alpha\beta\phi = 0$$

where $\alpha = -N$, $\beta = B^+/A^+$, $\gamma = 1 + B^-/A^-$. The only acceptable solution is a polynomial of finite degree N corresponding to the following hypergeometric function (as $\alpha = -N$ is a negative integer),

$$\phi_s(z) \;=\; 1 + \sum_{n=1}^{N} \frac{(\alpha)_n(\beta)_n}{(\gamma)_n}\frac{(1+s)^n}{n!}z^n$$

where $(u)_n$ is the Pochhammer symbol, $(u)_n = u(u+1)\cdots(u+n-1) = \Gamma(u+n)/\Gamma(u)$.

These stationary solutions can be rewritten, using the Γ function, as,

$$\phi_s(z) = \sum_{n=0}^{N} \binom{N}{n} \frac{\Gamma(\delta_1 - n)\Gamma(\delta_2 + n)}{\Gamma(\delta_1)\Gamma(\delta_2)} (1+s)^n z^n$$

where

$$\begin{aligned} \delta_1 &= 1 - \gamma = -B^-/A^- = N(1+\nu_1) > N \\ \delta_2 &= \beta = B^+/A^+ = N\nu_2/(1+s) \end{aligned}$$

for the parameters of the uncoupled mutation/selection model above. This leads to the *exact* stationary distribution, $P_s(n)$, of individuals with allele A, for *all* $n = 0, \ldots, N$ and δ_1 and δ_2 expressions valid for *any* population size N, fitness increment s and mutation rates ν_1 and ν_2,

$$P_s(n) = \binom{N}{n} (1+s)^n \frac{\Gamma(\delta_1 - n)\Gamma(\delta_2 + n)}{\Gamma(\delta_1)\Gamma(\delta_2)} \tag{A.3}$$

This expression holds, however, only if the arguments of the Γ functions are different from zero (*i.e.* $\delta_1, \delta_2 \neq 0$, $\delta_1 - n \neq 0$, $\delta_2 + n \neq 0$). This means that the Moran model in absence of mutations cannot be derived as the limit of this exact stationary distribution for $\nu_1, \nu_2 \to 0$. To retrieve this case, the partial differential equation for the probability generating function has to be directly rewritten for the suitable transition rates $W^{\pm}(n, \nu_1 = 0, \nu_2 = 0)$, which are equivalent for all three models (up to a rescaling of time scale for the first Moran model). The stationary solution has the form, $\phi_s(z) = \Pi_N z^N + \Pi_0$, and can be solved, following [178], leading to $\Pi_N = (1 - \sigma^{n_0})/(1 - \sigma^N)$, $\Pi_0 = (\sigma^{n_0} - \sigma^N)/(1 - \sigma^N)$, where $\sigma = 1/(1+s)$ and n_0 is the initial number of A-individuals in the population. In particular, the well-known result for the fixation probability in an haploid populations is readily retrieved as an approximation for a small selection coefficient s, noting that $\sigma^N = e^{n \log \sigma} \simeq e^{-ns}$,

$$\Pi_N = \frac{1 - e^{-sn_0}}{1 - e^{-sN}}$$

In particular, the probability of fixation of a new arisen mutant ($n_0 = 1$) in a large population reduces to $\Pi_N \simeq s$.

Using the exact solution of Eq. (A.3) and the Stirling's approximation for large factorials ($\Gamma(z+1) = z! \simeq e^{z \ln z - z}$), in addition to low fitness gain $s \ll 1$ and mutation rates $\nu_1, \nu_2 \ll 1$, then leads to the approximation,

$$P_s(p) \simeq (1+s)^{Np} (1-p)^{N\nu_1 - 1}\, p^{N\nu_2 - 1},$$

where $p = n/N$. This allows to recover the well-known approximate solution of the Wright-

Fisher model [176] for a diploid population ($N = 2N_e$) with non-overlapping generation,

$$P_s(p) \quad \propto \quad \bar{w}^{2N_e} p^{4N_e\nu_2'-1}(1-p)^{4N_e\nu_1'-1}$$

with $\bar{w} = 1 + ps$ and $2\nu_i' = \nu_i$ due to the difference in the distribution of offspring between the overlapping and non-overlapping generation models. Note that this factor 2 can be readily recovered in the uncoupled model assigning μ as the equal birth-or-death rate of each allele per generation and thus, $\mu/2$, as the rate of a-death-A-birth process in $W^+(n)$ and, similarly, as the converse rate of A-death-a-birth process in $W^-(n)$.

Appendix B

Pearl's theory of the *do*-calculus

In this chapter, Pearl's theory of the *do*-calculus and its application to the Mediation Analysis [124, 133, 137] will be discussed. Moreover, we will show how our general approach, introduced in section 4.3 in order to extend the Mediation Analysis framework to Markovian models with more than one intermediate variable and resulting in a practical method to estimate direct and indirect effects, is a consistent extension of Pearl's theory.

B.1 Basic concepts of the *do*-calculus

Causal analysis aims to infer not only beliefs or probabilities of events under static conditions, but also their dynamics under *changing conditions*, for example, changes induced by treatments or external interventions. In this context, the *do*-calculus theory – first introduced and developed by Pearl [179] – facilitates the identification of causal effects in non-parametric models. Interventions and counterfactuals are defined through a mathematical operator called $do(x)$, which simulates physical interventions by deleting certain functions from the model, replacing them with a constant $X = x$, while keeping the rest of the model unchanged. However, this theory unveils its usefulness also in the field of Mediation Analysis, for the estimation of direct and indirect causal effects [124,125,133,137]. In the following, causal analysis in graphical models and the decomposition of causal effects will be discussed in the light of the *do* operator. In particular, we will begin with the simple case of *Markovian* models, in which all causal effects are identifiable, *i.e.* discernible unambiguously from the data. Non-markovian models, such as those involving correlated errors (resulting from unmeasured confounders), permit identification only under certain conditions, that can be determined from the graph structure [124].

B.1.1 Markovian models

Denoting endogenous variables as variables whose values is determined by the model, the following theorems hold in Markovian models.

Theorem 1 (**The Causal Markov Condition**). *Any distribution generated by a Markovian model M can be factorized as*

$$P(v_1, \ldots, v_n) = \prod_i P(v_i|pa_i) \tag{B.1}$$

where $V_1, V_2, \ldots, V_N$ are the endogenous variables in M and pa_i are values of the endogenous parents of V_i in the causal diagram associated with M [124, 133].

Corollary 1 (**Truncated factorization**). *For any Markovian model, the distribution generated by an intervention $do(X = x_0)$ on a set X of endogenous variables is given by the factorization*

$$P(v_1, \ldots, v_k|do(x_0)) = \prod_{i|V_i \notin X} P(v_i|pa_i)|_{x=x_0} \tag{B.2}$$

where $P(v_i|pa_i)$ are the pre-intervention conditional probabilities [124, 133].

This formula reflects the removal of the terms $P(v_i|pa_i)$ for $V_i \in X$ from the product, since pa_i no longer influences $V_i \in X$; graphically it is equivalent to removing the links between PA_i and $V_i \in X$ while keeping the rest of the network intact.

Therefore, in Markovian models it results that the parents of X are the only variables that need to be measured to estimate the causal effects of X on Y, as stated by the following theorem.

Theorem 2 (**Adjustment for direct causes**). *Let PA_i denote the set of direct causes of variable X_i and let Y be any set of variables disjoint of $\{X_i \cup PA_i\}$. The effect of the intervention $do(X_i = x')$ on Y is given by*

$$P(y|do(x')) = \sum_{pa_i} P(y|x'_i, pa_i)P(pa_i), \tag{B.3}$$

where $P(y|x'_i, pa_i)$ and $P(pa_i)$ represent pre-intervention probabilities.

It calls for conditioning $P(y|x')$ on the parents of X_i and then averaging the result, weighted by the prior probability of $PA_i = pa_i$. This operation is known as *adjustment for PA_i*. The conditioning operator is not introduced to suppress spurious correlations between the cause X_i and the effect Y; rather, it emerges formally from the deeper principle represented in Eq. (B.2), that of preserving all the invariant information that the pre-intervention distribution can provide [124].

B.1.2 General models

In semi-Markovian models, the situation becomes more complicated. Indeed, unmeasured variables, known as *confounders*, are usually involved in the estimation of the effect of X on Y. The problem that arises is whether measurements should be adjusted

for possible variations in these other variables. Therefore, it becomes necessary to decide which variables are appropriate for adjustment. The following criterion settles this problem by providing a graphical method of selecting admissible sets of factors for adjustment [124, 133].

Theorem 3 (**The back-door criterion**). *A set of variables W satisfies the back-door criterion relative to an order pair of variables (X_i, X_j) in a DAG (direct acyclic graph) G and is therefore admissible (or sufficient) for adjustment if:*

1. *no node in W is a descendant of X_i, and*
2. *W blocks* every path *between X_i and X_j that contains an arrow into X_i.*

In this criterion, "blocking" is interpreted according to the following definition [124, 133].

Definition 1 (**d-separation**). A set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S. If S blocks *all* paths from X to Y, it is said to "d-separate X and Y," and then, X and Y are independent given S, written $X \perp\!\!\!\perp Y|S$.

The intuition behind the back-door criterion is as follows. The back-door paths in the diagram carry spurious associations from X to Y, while the paths directed along the arrows from X to Y carry causative associations. Blocking the former paths (by conditioning on W) ensures that the measured association between X and Y is purely causative. Finding an admissible set W guarantees to remove all confounding bias relative to the causal effect of X on Y [124, 133].

The back-door criterion allows to *directly* obtain the causal effect of X on Y, by selecting a sufficient set W directly from the diagram, without manipulating the truncated factorization formula.

Theorem 4 (**The back-door adjustment**). *If a set of variables W satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula*

$$P(y|do(x)) = \sum_w P(y|x, w)P(w). \tag{B.4}$$

B.1.3 The *do*-calculus and the Mediation Analysis

Based on the previous observations, it results that in Markovian models – as we assume in our analysis – the total effect of X on Y can be equivalently obtained applying Eq. (B.4) or Eq. (B.3) to Pearl's definition of total causal effect in Eq. (4.1), leading to the formula for the total effect in Eqs. (4.8) or in Eqs. (4.9) for each of our two approaches, respectively.

Concerning the direct effect, under some conditions (involving the back-door criterion, see [125] for details) it can be expressed in terms of the *do* operator. For example, if a set

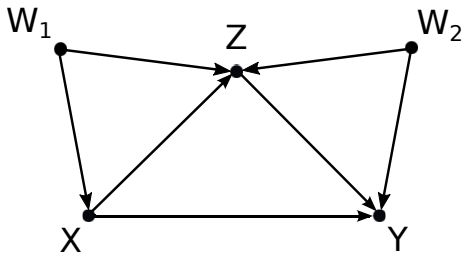

Figure B.1: Generic mediation model with two observed confounders. This graph shows a generic model depicting mediation from a variable X on a response variable Y through the mediator Z with two confounders W_1 and W_2. Figure adapted from [127].

W exists that satisfies the back-door criterion for the paths from Z to Y, the direct effect can be reduced to [124, 125, 127]

$$DE_{x,x'}(Y) = \sum_{z,w} [E(Y|do(x', z), w) - E(Y|do(x, z), w)] \; P(z|do(x), w) \; P(w).$$

In particular, in Markovian models each *do*-expression can be reduced to a *do*-free expression by covariate adjustment [124, 125, 133]. As an example, in the model in Fig. B.1 with two confounders W_1 and W_2, it is not sufficient to measure the association between X and Y by conditioning on Z, because we create spurious correlations between X and Y through W_2. Anyway, by covariate adjustment, the direct effect can be estimated as [127]

$$DE_{x,x'}(Y) = \sum_{z} \sum_{w_2} P(w_2)[E(Y|x', z, w_2) - E(Y|x, z, w_2)] \sum_{w_1} P(z|x, w_1, w_2)P(w_1).$$

Our first approach, based on the distinction of the intermediate variables between mediators Z and covariates W, directly derives from the former equation and, indeed, this last expression can be directly obtained from the formula for the direct effect in Eqs. (4.8). However, there is no need to actually hold *all* the intermediate variables Z between X and Y constant, since holding constant the direct parents of Y excluding X suffices [124, 125]. Thus, an equivalent definition of the direct effect can be obtained in terms of $pa_{Y \setminus X}$, that, combined with Eq. (B.3), leads to the formula for the direct effect in Eqs. (4.9).

Bibliography

[1] S. Ohno, U. Wolf, and N. B. Atkin, "Evolution from fish to mammals by gene duplication," Hereditas (1968).

[2] M. Nei, "Gene duplication and nucleotide substitution in evolution," Nature **221**, 40–42 (1969).

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

[4] K. H. Wolfe and D. C. Shields, "Molecular evidence for an ancient duplication of the entire yeast genome," Nature **387**, 708–712 (1997).

[5] C. Seoighe and K. H. Wolfe, "Updated map of duplicated regions in the yeast genome," Gene **238**, 253–61 (1999).

[6] M. Lynch and J. S. Conery, "The Evolutionary Fate and Consequences of Duplicate Genes," Science **290**, 1151–1155 (2000).

[7] M. Lynch and J. S. Conery, "The evolutionary demography of duplicate genes," J. Struct. Funct. Genomics **3**, 35–44 (2003).

[8] W.-H. Li, Z. Gu, A. R. O. Cavalcanti, and A. Nekrutenko, "Detection of gene duplications and block duplications in eukaryotic genomes," J. Struct. Funct. Genomics **3**, 27–34 (2003).

[9] R. Friedman and A. L. Hughes, "Gene duplication and the structure of eukaryotic genomes," Genome Res. **11**, 373–81 (2001).

[10] D. Sankoff, "Gene and genome duplication," Curr. Opin. Genet. Dev. **11**, 681–684 (2001).

[11] J. Zhang, "Evolution by gene duplication: an update," Trends Ecol. Evol. **18**, 292–298 (2003).

[12] K. Tartof, "Unequal crossing over then and now," Genetics **6**, 1–6 (1988).

[13] J. Taylor, P. Woods, and W. Hughes, "The organization and duplication of chromosomes as revealed by autoradiographic studies using tritium-labeled thymidinee," Proc. Natl. Acad. Sci. U. S. A. **43**, 122–128 (1957).

[14] J. A. Bailey, G. Liu, and E. E. Eichler, "An Alu transposition model for the origin and expansion of human segmental duplications." Am. J. Hum. Genet. **73**, 823–34 (2003).

[15] E. Kraus, W. Y. Leung, and J. E. Haber, "Break-induced replication: a review and an example in budding yeast," Proc. Natl. Acad. Sci. U. S. A. **98**, 8255–62 (2001).

[16] J.-M. Chen, D. N. Cooper, N. Chuzhanova, C. Férec, and G. P. Patrinos, "Gene conversion: mechanisms, evolution and human disease," Nat. Rev. Genet. **8**, 762–75 (2007).

[17] L. Duret and N. Galtier, "Biased gene conversion and the evolution of mammalian genomic landscapes," Annu. Rev. Genomics Hum. Genet. **10**, 285–311 (2009).

[18] P. Hastings and J. Lupski, "Mechanisms of change in gene copy number," Nat. Rev. Genet. **10**, 551–564 (2009).

[19] G. Levinson and G. Gutman, "Slipped-strand mispairing: a major mechanism for DNA sequence evolution," Mol. Biol. Evol. **4**, 203–221 (1987).

[20] R. Koszul, S. Caburet, B. Dujon, and G. Fischer, "Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments." EMBO J. **23**, 234–243 (2004).

[21] M. Syvanen, "Horizontal gene transfer: evidence and possible consequences," Annu. Rev. Genet. **28**, 237–61 (1994).

[22] J. P. Gogarten and J. P. Townsend, "Horizontal gene transfer, genome innovation and evolution," Nat. Rev. Microbiol. **3**, 679–87 (2005).

[23] T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, *et al.*, "A unified classification system for eukaryotic transposable elements," Nat. Rev. Genet. **8**, 973–82 (2007).

[24] J. V. Moran, R. J. DeBerardinis, and H. H. Kazazian, "Exon shuffling by L1 retrotransposition," Science **283**, 1530–1534 (1999).

[25] J. A. Eisen, "Horizontal gene transfer among microbial genomes: new insights from complete genome analysis," Curr. Opin. Genet. Dev. **10**, 606–11 (2000).

[26] H. Kaessmann, N. Vinckenbosch, and M. Long, "RNA-based gene duplication: mechanistic and evolutionary insights," Nat. Rev. Genet. **10**, 19–31 (2009).

[27] D. E. Soltis and P. S. Soltis, "Polyploidy: recurrent formation and genome evolution," Trends Ecol. Evol. **14**, 348–352 (1999).

[28] D. E. Soltis, P. S. Soltis, and J. A. Tate, "Advances in the study of polyploidy since Plant speciation," New Phytol. **161**, 173–191 (2003).

[29] S. Otto and J. Whitton, "Polyploid incidence and evolution," Annu. Rev. Genet. (2000).

[30] G. L. Stebbins, "Chromosomal evolution in higher plants," Chromosom. Evol. High. plants. (1971).

[31] B. Mable, "'Why polyploidy is rarer in animals than in plants': myths and mechanisms," Biol. J. Linn. Soc. pp. 453–466 (2004).

[32] H. J. Muller, "Why polyploidy is rarer in animals than in plants," Am. Nat. pp. 346–353 (1925).

[33] O. Jaillon, J.-M. Aury, B. Noel, A. Policriti, C. Clepet, *et al.*, "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla." Nature **449**, 463–7 (2007).

[34] A. McLysaght, K. Hokamp, and K. H. Wolfe, "Extensive genomic duplication during early chordate evolution," Nat. Genet. **31**, 200–4 (2002).

[35] P. Dehal and J. L. Boore, "Two rounds of whole genome duplication in the ancestral vertebrate," PLoS Biol. **3**, e314 (2005).

[36] L. Z. Holland, R. Albalat, K. Azumi, E. Benito-Gutiérrez, M. J. Blow, *et al.*, "The amphioxus genome illuminates vertebrate origins and cephalochordate biology," Genome Res. **18**, 1100–11 (2008).

[37] Y. Jiao, N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, *et al.*, "Ancestral polyploidy in seed plants and angiosperms," Nature **473**, 97–100 (2011).

[38] A. Amores, A. Force, Y. L. Yan, L. Joly, C. Amemiya, *et al.*, "Zebrafish hox clusters and vertebrate genome evolution," Science **282**, 1711–4 (1998).

[39] O. Jaillon, J.-M. Aury, F. Brunet, J.-L. Petit, N. Stange-Thomann, *et al.*, "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype," Nature **431**, 946–57 (2004).

[40] P. P. Singh, S. Affeldt, I. Cascone, R. Selimoglu, J. Camonis, and H. Isambert, "On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates," Cell Rep. **2**, 1387–98 (2012).

[41] T. Makino and A. McLysaght, "Ohnologs in the human genome are dosage balanced and frequently associated with disease," Proc. Natl. Acad. Sci. U. S. A. **107**, 9270–4 (2010).

[42] L. Huminiecki and C. H. Heldin, "2R and remodeling of vertebrate signal transduction engine," BMC Biol. **8**, 146 (2010).

[43] 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**, 5454–9 (2005).

[44] T. Blomme, K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, and Y. Van de Peer, "The gain and loss of genes during 600 million years of vertebrate evolution," Genome Biol. **7**, R43 (2006).

[45] M. Freeling and B. C. Thomas, "Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity," Genome Res. **16**, 805–14 (2006).

[46] G. C. Conant and K. H. Wolfe, "Turning a hobby into a job: how duplicated genes find new functions," Nat. Rev. Genet. **9**, 938–50 (2008).

[47] H. Wada, "Origin and genetic evolution of the vertebrate skeleton," Zoolog. Sci. **27**, 119–123 (2010).

[48] L. Hakes, J. W. Pinney, S. C. Lovell, S. G. Oliver, and D. L. Robertson, "All duplicates are not equal: the difference between small-scale and genome duplication," Genome Biol. **8**, R209 (2007).

[49] Y. Guan, M. J. Dunham, and O. G. Troyanskaya, "Functional analysis of gene duplications in Saccharomyces cerevisiae," Genetics **175**, 933–43 (2007).

[50] C. B. Bridges, "The Bar 'gene' a duplication," Science **83**, 210–211 (1936).

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

[52] K. Wolfe, "Robustness? It's not where you think it is," Nat. Genet. **25** (2000).

[53] 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**, 1531–45 (1999).

[54] M. Lynch, M. O'Hely, B. Walsh, and A. Force, "The probability of preservation of a newly arisen gene duplicate," Genetics **159**, 1789–1804 (2001).

[55] M. Lynch and B. Walsh, "Genetics and analysis of quantitative traits," (1998).

[56] J. B. S. Haldane, "The part played by recurrent mutation in evolution," Am. Nat. pp. 5–19 (1933).

[57] R. A. Fisher, "The sheltering of lethals," Am. Nat. **69**, 446–455 (1935).

[58] M. Nei and A. K. Roychoudhury, "Probability of fixation of nonfunctional genes at duplicate loci," Am. Nat. pp. 362–372 (1973).

[59] G. S. Bailey, R. T. Poulter, and P. A. Stockwell, "Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci," Proc. Natl. Acad. Sci. U. S. A. **75**, 5575–9 (1978).

[60] W. H. Li, "Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes," Genetics **95**, 237–58 (1980).

[61] L. G. Lundin, "Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse," Genomics **16**, 1–19 (1993).

[62] J. F. Y. Brookfield, "Genetic redundancy: screening for selection in yeast," Curr. Biol. **7**, R366–8 (1997).

[63] J. Nadeau and D. Sankoff, "Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution," Genetics **147**, 1259–1266 (1997).

[64] M. Lynch and A. Force, "The probability of duplicate gene preservation by subfunctionalization," Genetics **154**, 459–473 (2000).

[65] T. Ohta, "Time for acquiring a new gene by duplication," Proc. Natl. Acad. Sci. U. S. A. **85**, 3509–12 (1988).

[66] J. B. Walsh, "How often do duplicated genes evolve new functions?" Genetics **139**, 421–8 (1995).

[67] A. Hughes, "The evolution of functionally novel proteins after gene duplication," Proc. R. Soc. London B **256**, 119–124 (1994).

[68] T. A. Gibson and D. S. Goldberg, "Questioning the ubiquity of neofunctionalization," PLoS Comput. Biol. **5**, e1000252 (2009).

[69] M. Goodman and G. W. Moore, "Darwinian evolution in the genealogy of haemoglobin," Nature **253**, 603–608 (1975).

[70] R. A. Jensen, "Enzyme recruitment in evolution of new function," Annu. Rev. Microbiol. **30**, 409–425 (1976).

[71] R. A. Jensen and G. S. Byng, "The partitioning of biochemical pathways with isozyme systems." Isozymes Curr. Top. Biol. Med. Res. **5**, 143–174 (1980).

[72] J. Westin and M. Lardelli, "Three novel Notch genes in zebrafish: implications for vertebrate Notch gene evolution and function," Dev. Genes Evol. **207**, 51–63 (1997).

[73] R. J. DiLeone, L. B. Russell, and D. M. Kingsley, "An extensive 3' regulatory region controls expression of Bmp5 in specific anatomical structures of the mouse embryo," Genetics **148**, 401–8 (1998).

[74] S. Nornes, M. Clarkson, I. Mikkola, M. Pedersen, A. Bardsley, J. P. Martinez, S. Krauss, and T. Johansen, "Zebrafish contains two pax6 genes involved in eye development," Mech. Dev. **77**, 185–96 (1998).

[75] M. Sémon and K. H. Wolfe, "Consequences of genome duplication," Curr. Opin. Genet. Dev. **17**, 505–12 (2007).

[76] B. A. Chapman, J. E. Bowers, F. A. Feltus, and A. H. Paterson, "Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication," Proc. Natl. Acad. Sci. U. S. A. **103**, 2730–5 (2006).

[77] G. Tuskan, S. Difazio, and S. Jansson, "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)," Science **313**, 1596–1604 (2006).

[78] M. A. Nowak, M. C. Boerlijst, J. Cooke, and J. M. Smith, "Evolution of genetic redundancy," Nature **388**, 167–71 (1997).

[79] 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," Dev. Biol. **234**, 275–88 (2001).

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

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

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

[83] J. A. Birchler and R. A. Veitia, "The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution," New Phytol. **186**, 54–62 (2010).

[84] J.-M. Aury, O. Jaillon, L. Duret, B. Noel, C. Jubin, *et al.*, "Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia," Nature **444**, 171–8 (2006).

[85] S. J. Furney, M. M. Albà, and N. López-Bigas, "Differences in the evolutionary history of disease genes affected by dominant or recessive mutations," BMC Genomics **7**, 165 (2006).

[86] M. Pufall and B. Graves, "Autoinhibitory domains: modular effectors of cellular regulation," Annu. Rev. Cell Dev. Biol. **18**, 421–462 (2002).

[87] R. Blekhman, O. Man, L. Herrmann, A. R. Boyko, A. Indap, *et al.*, "Natural selection on genes that underlie human disease susceptibility," Curr. Biol. **18**, 883–9 (2008).

[88] J. J. Cai, E. Borenstein, R. Chen, and D. A. Petrov, "Similarly strong purifying selection acts on human disease genes of all evolutionary ages," Genome Biol. Evol. **1**, 131–44 (2009).

[89] T. Domazet-Loso and D. Tautz, "An ancient evolutionary origin of genes associated with human genetic diseases," Mol. Biol. Evol. **25**, 2699–707 (2008).

[90] J. E. Dickerson and D. L. Robertson, "On the origins of Mendelian disease genes in man: the impact of gene duplication," Mol. Biol. Evol. **29**, 61–9 (2012).

[91] L. W. Berry, B. Westlund, and T. Schedl, "Germ-line tumor formation caused by activation of glp-1, a Caenorhabditis elegans member of the Notch family of receptors," Development **124**, 925–36 (1997).

[92] C. Ciocan, J. Moore, and J. Rotchell, "The role of ras gene in the development of haemic neoplasia in Mytilus trossulus," Mar. Environ. Res. **62**, S147–S150 (2006).

[93] J. Robert, "Comparative study of tumorigenesis and tumor immunity in invertebrates and nonmammalian vertebrates," Dev. Comp. Immunol. **34**, 915–925 (2010).

[94] K. Ise, K. Nakamura, K. Nakao, S. Shimizu, H. Harada, *et al.*, "Targeted deletion of the H-ras gene decreases tumor formation in mouse skin carcinogenesis," Oncogene **19**, 2951–6 (2000).

[95] L. M. Esteban, C. Vicario-abejón, P. Férnandez-Salguero, A. Fernández-Medarde, N. Swaminathan, *et al.*, "Targeted genomic disruption of H-ras and N-ras, individually or in combination, reveals the dispensability of both loci for mouse growth and development," Mol. Cell. Biol. **21**, 1444–1452 (2001).

[96] P. P. Singh, "Expansion of disease gene families by whole genome duplication in early vertebrates," PhD Dissertation (2013).

[97] S. Affeldt, P. P. Singh, I. Cascone, R. Selimoglu, J. Camonis, and H. Isambert, "Evolution and cancer: expansion of dangerous gene repertoire by whole genome duplications," Médecine Sci. M/S **29**, 358–361 (2013).

[98] I. Cascone, R. Selimoglu, C. Ozdemir, E. Del Nery, C. Yeaman, M. White, and J. Camonis, "Distinct roles of RalA and RalB in the progression of cytokinesis are supported by distinct RalGEFs," EMBO J. **27**, 2375–87 (2008).

[99] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, and A. Others, "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis," Science **285**, 901–906 (1999).

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

[101] R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, *et al.*, "Systematic functional analysis of the Caenorhabditis elegans genome using RNAi," Nature **421**, 231–237 (2003).

[102] X. Gu, "Evolution of duplicate genes versus genetic robustness against null mutations," Trends Genet. **19**, 354–6 (2003).

[103] T. Makino, K. Hokamp, and A. McLysaght, "The complex relationship of gene duplication and essentiality," Trends Genet. **25**, 152–155 (2009).

[104] J.-F. Gout, D. Kahn, and L. Duret, "The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution," PLoS Genet. **6**, e1000944 (2010).

[105] T. J. Gibson and J. Spring, "Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins," Trends Genet. **14**, 46–9; discussion 49–50 (1998).

[106] G. Malaguti, P. P. Singh, and H. Isambert, "On the retention of gene duplicates prone to dominant deleterious mutations," Theor. Popul. Biol. **93**, 38–51 (2014).

[107] C. R. Werth and M. D. Windham, "A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression," Am. Nat. pp. 515–526 (1991).

[108] M. Lynch and A. Force, "The origin of interspecific genomic incompatibility via gene duplication," Am. Nat. **156**, 590–605 (2000).

[109] R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon Press, Oxford, 1930).

[110] S. Wright, "Evolution in Mendelian Populations," Genetics **16**, 97–159 (1931).

[111] P. Moran, "Random processes in genetics," Math. Proc. Cambridge Philos. Soc. **54**, 60–71 (1958).

[112] S. Wright, "The differential equation of the distribution of gene frequencies," Proc. Natl. Acad. Sci. **2**, 382–389 (1945).

[113] M. Kimura, "Solution of a process of random genetic drift with a continuous model," Proc. Natl. Acad. Sci. **2**, 144–150 (1955).

[114] S. H. Rice, *Evolutionary Theory: Mathematical and Conceptual Foundations* (Texas Tech University, 2004).

[115] J. Wakeley, "The limits of theoretical population genetics," Genetics **7**, 1–7 (2005).

[116] P. A. P. Moran, "The effect of selection in a haploid genetic population," Math. Proc. Cambridge Philos. Soc. **54**, 463–467 (1958).

[117] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, vol. 1 (Elsevier, 1992).

[118] B. Eldon and J. Wakeley, "Coalescent processes when the distribution of offspring number among individuals is highly skewed," Genetics **172**, 2621–33 (2006).

[119] C. Muirhead and J. Wakeley, "Modeling multiallelic selection using a Moran model," Genetics **1157**, 1141–1157 (2009).

[120] A. M. Etheridge and R. C. Griffiths, "A coalescent dual process in a Moran model with genic selection," Theor. Popul. Biol. **75**, 320–330 (2009).

[121] A. M. Etheridge, R. C. Griffiths, and J. E. Taylor, "A coalescent dual process in a Moran model with genic selection, and the lambda coalescent limit," Theor. Popul. Biol. **78**, 77–92 (2010).

[122] C. Vogl and F. Clemente, "The allele-frequency spectrum in a decoupled Moran model with mutation, drift, and directional selection, assuming small mutation rates," Theor. Popul. Biol. **81**, 197–209 (2012).

[123] M. Lynch, "Evolution of the mutation rate," Trends Genet. **26**, 345–52 (2010).

[124] J. Pearl, *Causality: models, reasoning and inference* (Cambridge University Press, 2nd edition, 2009).

[125] J. Pearl, "Direct and indirect effects," in "Proc. seventeenth Conf. Uncertain. Artif. Intell.", (2001), pp. 411–420.

[126] J. Pearl, "The causal mediation formula–a guide to the assessment of pathways and mechanisms." Prev. Sci. **13**, 426–36 (2012).

[127] J. Pearl, "The mediation formula: A guide to the assessment of causal pathways in nonlinear models," Causality Stat. Perspect. Appl. pp. 151–179 (2012).

[128] M. H. Maathuis, M. Kalisch, and P. Bühlmann, "Estimating high-dimensional intervention effects from observational data," Ann. Stat. **37**, 3133–3164 (2009).

[129] D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz, "Mediation analysis," Annu. Rev. Psychol. **58**, 593–614 (2007).

[130] R. M. Baron and D. A. Kenny, "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations," J. Pers. Soc. Psychol. **51**, 1173 (1986).

[131] J. Robins and S. Greenland, "Identifiability and exchangeability for direct and indirect effects," Epidemiology **3**, 143–155 (1992).

[132] K. Imai, L. Keele, and D. Tingley, "A general approach to causal mediation analysis," Psychol. Methods **15**, 309–334 (2010).

[133] J. Pearl, "Causal inference in statistics: An overview," Stat. Surv. **3**, 96–146 (2009).

[134] I. Shpitser and J. Pearl, "Complete identification methods for the causal hierarchy," J. Mach. Learn. Res. pp. 1941–1979 (2008).

[135] S. Affeldt, *et al.*, "Robust Inference of Causal Graphical Models for Genomics Data," (In preparation).

[136] E. Dijkstra, "A note on two problems in connexion with graphs," Numer. Math. **1**, 269–271 (1959).

[137] J. Pearl, "The do-calculus revisited," arXiv Prepr. arXiv1210.4852 pp. 4–11 (2012).

[138] S. P. Otto and P. Yong, "The evolution of gene duplicates," Adv. Genet. **46**, 451–483 (2002).

[139] F. A. Kondrashov and A. S. Kondrashov, "Role of selection in fixation of gene duplications," J. Theor. Biol. **239**, 141–151 (2006).

[140] J. F. Crow and M. Kimura, "Evolution in sexual and asexual populations," Am. Nat. **99**, 439–450 (1965).

[141] Z. Patwa and L. Wahl, "The fixation probability of beneficial mutations," J. R. Soc. Interface pp. 1279–1289 (2008).

[142] W. Qian and J. Zhang, "Gene dosage and gene duplicability," Genetics **179**, 2319–24 (2008).

[143] P. P. Singh, S. Affeldt, G. Malaguti, and H. Isambert, "Human dominant disease genes are enriched in paralogs originating from whole genome duplication," PLoS Comput. Biol. **10**, e1003754 (2014).

[144] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae," Nature **428**, 617–24 (2004).

[145] B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuvéglise, E. Talla, *et al.*, "Genome evolution in yeasts," Nature **430**, 35–44 (2004).

[146] H. Deng and M. Lynch, "Estimation of deleterious-mutation parameters in natural populations," Genetics **144**, 349–360 (1996).

[147] L. L. Vassilieva, A. M. Hook, and M. Lynch, "The fitness effects of spontaneous mutations in Caenorhabditis elegans," Evolution **54**, 1234–1246 (2000).

[148] H. Deng, G. Gao, and J. Li, "Estimation of deleterious genomic mutation parameters in natural populations by accounting for variable mutation effects across loci," Genetics **162**, 1487–1500 (2002).

[149] J. Fry and S. Nuzhdin, "Dominance of mutations affecting viability in Drosophila melanogaster," Genetics **163**, 1357–1364 (2003).

[150] X.-S. Zhang, J. Wang, and W. G. Hill, "Influence of dominance, leptokurtosis and pleiotropy of deleterious mutations on quantitative genetic variation at mutation-selection balance," Genetics **166**, 597–610 (2004).

[151] N. Phadnis and J. D. Fry, "Widespread correlations between dominance and homozygous effects of mutations: implications for theories of dominance," Genetics **171**, 385–92 (2005).

[152] A. F. Agrawal and M. C. Whitlock, "Inferences about the distribution of dominance drawn from yeast gene knockout data," Genetics **187**, 553–66 (2011).

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

[154] S. Kuraku and A. Meyer, "Whole genome duplications and the radiation of vertebrates," Evol. after Gene Duplic. pp. 299–311 (2010).

[155] H. Innan and F. Kondrashov, "The evolution of gene duplications: classifying and distinguishing between models," Nat. Rev. Genet. **11**, 97–108 (2010).

[156] M. Sémon and K. H. Wolfe, "Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis," Proc. Natl. Acad. Sci. U. S. A. **105**, 8333–8 (2008).

[157] T. Ohta, "Population size and rate of evolution," J. Mol. Evol. **1**, 305–314 (1972).

[158] R. Lande, "Risk of population extinction from fixation of new deleterious mutations," Evolution pp. 1460–1469 (1994).

[159] R. Siegel, "Cancer statistics, 2013," A cancer J. Clin. **63**, 11–30 (2013).

[160] L. L. Capasso, "Antiquity of cancer," Int. J. Cancer **113**, 2–13 (2005).

[161] A. R. David and M. R. Zimmerman, "Cancer: an old disease, a new disease or something in between?" Nat. Rev. Cancer **10**, 728–733 (2010).

[162] B. Faltas, "Cancer is an ancient disease: the case for better palaeoepidemiological and molecular studies," Nat. Rev. Cancer **11**, 76 (2010).

[163] Y. Wang, T. Zhang, and W. Wang, "An old disease, a new disease or something in between: evidence from China," Nat. Rev. Cancer **11**, 76 (2010).

[164] T. Domazet-Lošo, A. Klimovich, B. Anokhin, F. Anton-Erxleben, M. J. Hamm, C. Lange, and T. C. G. Bosch, "Naturally occurring tumours in the basal metazoan Hydra," Nat. Commun. **5**, 4222 (2014).

[165] C. M. Judd and D. A. Kenny, "Process Analysis: Estimating Mediation in Treatment Evaluations," Eval. Rev. **5**, 602–619 (1981).

[166] D. Muller, C. M. Judd, and V. Y. Yzerbyt, "When moderation is mediated and mediation is moderated," J. Pers. Soc. Psychol. **89**, 852 (2005).

[167] P. E. Shrout and N. Bolger, "Mediation in experimental and nonexperimental studies: New procedures and recommendations," Psychol. Methods **7**, 422–445 (2002).

[168] M. L. Petersen, S. E. Sinisi, and M. J. van der Laan, "Estimation of direct causal effects," Epidemiology **17**, 276–84 (2006).

[169] A. N. Glynn, "The product and difference fallacies for indirect effects," Am. J. Pol. Sci. **56**, 257–269 (2012).

[170] D. M. Hafeman and S. Schwartz, "Opening the Black Box: a motivation for the assessment of mediation," Int. J. Epidemiol. **38**, 838–45 (2009).

[171] T. J. VanderWeele, "Marginal structural models for the estimation of direct and indirect effects," Epidemiology **20**, 18–26 (2009).

[172] K. Imai, L. Keele, and T. Yamamoto, "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," Stat. Sci. **25**, 51–71 (2010).

[173] K. Imai and T. Yamamoto, "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments," Polit. Anal. **21**, 141–171 (2013).

[174] C. Avin, I. Shpitser, and J. Pearl, "Identifiability of path-specific effects," in "Proc. Int. Jt. Conf. Artif. Intell.", (2005), pp. 357–363.

[175] J. M. Albert and S. Nelson, "Generalized causal mediation analysis," Biometrics **67**, 1028–38 (2011).

[176] W. J. Ewens, *Mathematical population genetics.* (Springer-Verlag, New York, 1979).

[177] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs and mathematical tables* (Dover, 1964).

[178] B. Houchmandzadeh and M. Vallade, "Alternative to the diffusion equation in population genetics," Phys. Rev. E **82**, 1–8 (2010).

[179] J. Pearl, "Causal diagrams for empirical research," Biometrika (1995).