University of Crete
ICS-FORTH

Master Thesis

Bioinformatics applications in plant molecular research

Author:
Antonios Kioukis

Committee:
Ilias Lagkoubardos
Pavlos Pavlidis
Ioannis Iliopoulos

A thesis submitted in fulfillment of the requirements for the degree of Bioinformatics Msc

in the

Medical School, University of Crete, Greece

Antonios Kioukis
Fall 2018

"Πώς η ανάγκη γίνεται ιστορία
πώς η ιστορία γίνεται σιωπή
τι με κοιτάζεις Ρόζα μουδιασμένο
συγχώρα με που δεν καταλαβαίνω
τι λένε τα κομπιούτερς κι οι αριθμοί "

Άλκης Αλκαίος

Acknowledgements

This work has been carried out at the Institute of Computer Science (ICS) a part of Foundation of Research and Technology Hellas (FORTH), Crete, Greece. I would like to express my gratitude to many excellent people I was lucky to work with. First of all, I would like to thank my supervisor, Dr. Pavlos Pavlidis for his enthusiasm and believe in my work. His thoroughness and patience from the initial to the final level enabled me to develop an understanding of the subjects and made this thesis a reality. His support in difficult moments helped me not to give up.

I am grateful to Dr. Ilias Lagkoubardos for providing the necessary background knowledge and guidance. His support and eye for detail acted as a barrier against the accumulation of small errors that would have culminated to a disaster.

It was a pleasure to work with Dr. Panagiwtis Sarris and his student Vassiliki Michalopoulou on the project of Brassica Cretica. It enabled the application of theoretical knowledge and exposed me to new ideas and procedures.

I would like to express my gratitude to current and past lab members. First of all to Aggelos Koropoulis a friend and a keen listener to any problems I faced through these years. Anna Mathioudaki helped me to develop a deeper understanding of the biological processes and to Yiannis Koutsoukos who help me keep perspective of what is truly important in life.

I want to thank my parents, who motivated me to pursue this masters degree, not to give up, but also to respect and love myself. Finally, my brother was of great support by lending a sympathetic ear when I was stressed and provided comfort.

Contents

List of Figures

Chapter 1

Introduction

Plants are living organisms that can live on land or water. They have many different forms such as huge trees, others are herbs or some have bushy form. The basic food for all organisms is produced by green plants. Plants help in maintaining oxygen balance, the most important gas that enable us to breathe. Animals emit carbon dioxide by taking in oxygen, plants reduce this rise in carbon dioxide levels in air. Removal of carbon dioxide from the atmosphere reduces the greenhouse effect and global warming. It also maintains the ozone layer that helps protect Earth's life from damaging UV radiation. Humans directly or indirectly depend on plants for several of their needs, ranging from food to natural pesticides and even fuel. As a result, it is vital to study and comprehend their molecular mechanisms. Bioinformatics allow us to explore them in new ways allowing us to increase our knowledge at a relatively low cost but at an unprecedented speed. This is achieved by the creation of novel tools which model molecular mechanisms, development of databases for increased interpretability of data and decrease in the time required from sequencing to full genome reconstruction and annotation.

Databases: Data generated in wet-lab experiments are not presented in a human-friendly manner, they are raw, unharnessed information. Meanwhile their sheer volume impedes their processing. Databases have been in use from the infantile steps of the field to aid in the usability of the data. This usability encapsulates initial analysis, creation of search filters and extraction methods. By adding robust concepts of organization schemas in biological databases we can increase their scalability thus improving the level of Human-Computer-Interaction (HCI). Taken all this in account, I have created Crowth a platform for the genetic identification of grapevine, honey and flowers using the ITS1 region of each plant. Crowth aims to empower each producer of these products and increase the confidence of the consumers. Until now, the identification of the cultivar for each of these products was based on physical observation through microscopes. Crowth automates this process, resulting in decreasing the time needed for analysis trying to keep up with the demand. The ITS region was chosen due to its hyper-variability even amongst closely related taxa. This variability of ITS has led to the recommendation of [] to be used as a classification of all plant and animal life in conjunction with the use of CO1. Crowth is available at `https://github.com/antokioukis/crowth`

Simulators: Wet-lab experiments require monetary capital and adequate facilities. Computer programs can simulate complex biological systems precisely and at a fraction of the cost thus providing a worthwhile alternative. In population genetics, the need for simulators is prevalent. Simulators allow researchers to study scenarios that are not tractable mathematically, test infer-

ence algorithms and facilitate the development of analytical models for complex molecular data. The main categories of genetic data simulations are divided in forward-time approaches and backward in time. In forward-time simulations, an initial population is constructed and it evolves forward in time until a threshold is reached. The main advantage of forward-time simulators is flexibility, as it allows the generation of complex models. The main disadvantage is that the whole population needs to be tracked, thus they are expensive in terms of computational resources. EvoNet is a forward in-time simulator that models the evolution of Gene Regulatory Networks (GRN) by extending previous research by modeling a deeper level of gene interaction based on each gene's cis- and trans- regions. EvoNet aims to identify the effect of mutations both beneficial and deleterious as they are cascaded or phased out in the population. EvoNet is available at `https://github.com/antokioukis/evonet`.

Genome: With the advert of next generation sequencing it is a fact that the cost of sequencing drops every day. Empowered by this, faster and more precise genome assemblers have been created resulting in the minimization of the time required to get a full genome from months to hours. By collaborating with the lab of Dr. Sarris, who sequenced four samples of Brassica Cretica we have reconstructed a draft genome and with the use of the dadi pipeline inferred the demography model. This project allowed me to bridge the gap from theory to practice and introduced me to the use of cutting-edge software tools. Brassica cretica Lam. is a wild crop relative of a big number of crops of the genus Brassica, proposed to be the ancestor of broccoli, Brussel sprouts, cabbage, cauliflower, kale, swede, turnip and oilseed rape. Since this species is thought to be the gene donor of many crops of the brassica group, it might contain genes that are not included in the domesticated crops, as well as a different set of NLRs.

Chapter 2

Crowth

Crowth, an identification platform for grapevine, olive and honey

Abstract

The Internal Transcriber Spacer (ITS) region has been proposed to act as the universal DNA barcode for plants. Here, we present Crowth (CRetan grOWTH), a web platform that identifies and quantifies the plant origins of three Cretan products by creating a genetic identity based on their ITS region. Furthermore, each sequence of interest is placed in a phylogenetic tree to allow for broader evidences of similarity. To our best knowledge, Crowth is the first web server dedicated to the identification and quantification of wine, olive oil and honey using the ITS region, and currently hosts more than two hundred plants endemic in Crete. Crowth is available at *http:139.91.68.81/*

2.1 Introduction

Internal Transcriber Spacers (ITS1, ITS2) are spacer DNA located between the small-subunit ribosomal RNA (rRNA) and large-subunit rRNA genes. In plants, ITS1 is located between 18S and 5.8S rRNA genes, while ITS2 is between 5.8S and 26S. ITS1 and ITS2 have long been used as a region for phylogenetic reconstruction of species and genus relationships (Yao et al., 2010a; Coleman, 2003; Coleman, 2007; Coleman, 2009) using comparisons of primary sequence. The usage of ITS makes possible the creation of reliable sequence-structure alignments that take into account the secondary structure of the region due to its high conservation within all eukariotes (Schultz et al., 2005; Schultz et al., 2006; Schultz and Wolf, 2009). The comparison of sequences based on the ITS region is widely used in taxonomy (Yao et al., 2010b) and molecular phylogeny because of several favorable properties. Its small size allows for amplification and association with available highly conserved flanking sequences. It is detectable even from small quantities of DNA due to the high copy number of the rRNA clusters (Song et al., 2012). Unequal crossing-over and gene conversion result in rapid concerted evolution. This promotes intra-genomic homogeneity of the repeat units, although high-throughput sequencing showed the occurrence of frequent variations within plant species. Finally, it has a high degree of variation even between closely related species. This can be explained by the relatively low evolutionary pressure acting on such non-coding spacer sequences. This conservation permits comparisons at deeper taxonomic levels (Chen et al., 2010; Gao et al., 2010; Pang et al., 2011; Luo et al., 2010; Li et al., 2010b; Prasad

et al., 2009a; Prasad et al., 2009b). Based on these facts we created Crowth , a platform for the genetic identification of three Cretan products wine, olive oil and honey. Crowth is based on the Internal Transcriber Spacers(ITS1, ITS2) to create genetic identities for each of the plants. The genetic identity is more specific in grapevine and olive trees differentiating between different cultivars where as the genetic identity of flowers signals a higher taxonomic level. The diversity of Cretan micro-climates combined with the island's altitude differences enhance the diversification of flora and allows for plants of different taxonomic groups to co-exist and mix. Crowth provides the necessary framework to get back at the source of each product and identify whence it came from.

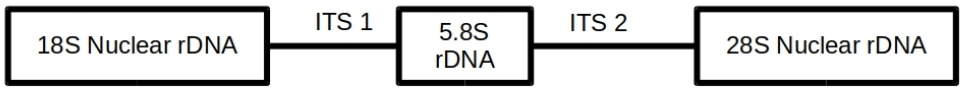

Figure 2.1: ITS locations between 18S,5.8S and 26S genes

2.2 Materials and Methods

2.2.1 Database Schema

The core of Crowth is comprised of three tables (Grapevine, Flower and Olive) that contain each products' available information. For every plant in the database Crowth stores: (i) a unique integer identifier used as the primary key of the table. (ii) The name of the plant originated from the NCBI downloaded file. (iii) The ITS sequence is stored in the sequence field of the table. (iv) The last updated field holds information showing when the table entry was last modified. (v) Cultivar description is also taken from the NCBI file and holds all the information besides name and sequence provided by the NCBI file. (vi) The link field is currently empty but when populated will provide a hypertext link to a page holding general information about the plant in question.

Crowth operates on two categories of queries. Identification Queries accurately predict the closest taxa from the user-provided sequences. Quantification Queries handle metagenomic samples, by processing a FASTQ file containing amplicons from a PCR experiment. The results of all queries are available for download from the main dashboard located in /jobs.html.

2.2.2 Query Analysis Overview

Identification Queries

Identification Queries are further divided in two categories: Simple sequence repeats (SSRs) and ITS, depending on what region will be used for the identification process.

SSRs analysis is composed of two phases, parsing and distance calculation. Crowth currently supports 18 different SSR locations. The first step of each SSR query is to identify which of the 18 markers, currently supported by Crowth , are contained in the input. Locations not present in the input file do not affect the analysis results. It is worth noting that the robustness of the analysis and the confidence in the results are analogous with the number of included SSR

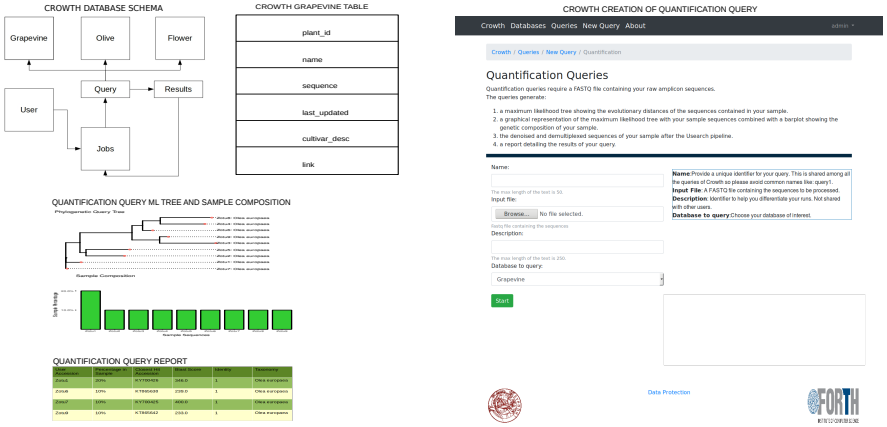

Figure 2.2: Top left: Crowth database schema, showing the process of each new query. Top Middle: Grapevine table field. Right: Query creation screen as viewed by the front-end. Bottom left: Quantification Query results available for download.

locations. Crowth calculates the distance between the provided sample and the locations in the database. The distance metric used is the euclidean distance due to the linearity of the data and its low processing resources requirements. The SSR identification analysis is currently available only for grapevine.

The identification analysis based on the ITS region requires more steps. The input sequences are BLASTn (Camacho et al., 2009) against the target products' database sequences. The top N hits for each sequence are saved in a tab-delimited file. Next, the input sequences are used to create a maximum likelihood phylogenetic tree (Stamatakis, 2014). To do that, the input sequences must be aligned with the N best hit database sequences. For this purpose we use the MAFFT aligner (Katoh and Standley, 2013). The alignment of database and query sequences is calculated from scratch everytime because Crowth has no knowledge a priori of the database sequences that would be the top hits. After the new alignment is done, the phylogenetic tree is created using RAxML (Stamatakis, 2014) and is stored in newick format. The resulting tree is used for a preliminary visualization through the use of a custom R script.

Quantification Queries

Quantification queries are handled by two already developed tools Usearch (Edgar, 2016) and RAxML (Stamatakis, 2014). The first step is dereplication which finds the set of unique sequences from the FASTQ file. Dereplication compares all the sequences from the input file and extracts sequences that match exactly. Denoising (Edgar and Flyvbjerg, 2015) is the next step. Sequence errors from amplicon reads are removed while identifying the correct biological sequences in the reads. The output sequences are now free of errors and are placed in a FASTA file, followed by a BlastN search against the Crowth database. Sequences that match with the database are excluded from further analysis and

are placed in a file available for download when the query has been completed. Not matched sequences, are placed in the database's phylogenetic tree to identify the closer taxa. This action offers additional information for the sample. The final phylogenetic tree is visualized in R.

2.2.3 Implementation

Crowth is implemented using the Django python framework. Django implements a MVC (Model-View-Controller) architecture, consisting of an object-relational mapper that interacts with data models in the relational database ("Model"), a handler of HTTP requests with a web templating system ("View"), and a regular-expression-based URL dispatcher ("Controller"). Crowth currently supports up to 3 concurrent queries independent of the query type. Crowth guarantees to maintain all query results for at least two weeks. The actual processing of the data is handled by custom made python scripts implementing a many to one access to the Crowth database. Each query is scheduled with the use of Celery (FOSS, 1999). Celery is an open-source asynchronous task queue or job queue which is based on distributed message passing. NGINX is used as the back-end HTTP server and reverse proxy. NGINX was chosen because it does not rely on threads to handle requests like traditional servers (ex: apache2). Instead, it uses a scalable event-driven (asynchronous) architecture which uses small, but more importantly, predictable amounts of memory under load. Supervisor is used as a fail-safe to automatically detect anomalies such as the Django back-end or the NGINX front-end shutting down... Its purpose is to restart them thus losing the minimal response time.

2.2.4 Data Export and Search

Crowth enables each user to export the database data. Each table supports a dedicated link to generate a file containing all the available information. This file is either in comma separated format or fasta. Searches on the products database can be conducted using as filter either the NCBI accession name or a part of the description. Downloads of the search results are also possible.

2.3 Server Description

Crowth is currently available at: http://139.91.68.81/.

2.3.1 Input

Crowth can be queried using a DNA sequences in the FASTA format for identification queries. Quantification queries require a FASTQ file of the query sequences. The length of the input sequences is not limited and the time required for identification queries is typically small (¡ 1 min). However, it may take up to several minutes for big inputs. Quantification queries, typically, take longer. A bare bones programming interface is currently being developed.

2.3.2 Output

Crowth provides output for all the three type of queries on the same page. Each output is available for download for two weeks after each query has completed. The query outputs bundle together: a general report explaining in detail

how the output files where generated and what they report, the specific tab-delimited query report file, the phylogenetic trees in newick format as well as their visualizations.

2.4 Conclusion and Future Perspectives

Crowth is a robust solution for identification and quantification for every producer and consumer based on NGS technologies. With the inclusion of SSR methods for backwards compatibility, Crowth seeks to extend the use of NGS identification methods on grapevine, olive oil and honey without alienating current approaches. In the future Crowth will include a picture of each plant as well as geographical links of its known habitats.

Chapter 3

Evonet

Evolution of gene regulatory networks by means of selection and random genetic drift

Abstract

The evolution of a population by means of genetic drift and natural selection operating on a gene regulatory network (GRN) of an individual has not been scrutinized in depth. Thus, the relative importance of various evolutionary forces and processes on shaping genetic variability in GRNs is understudied. Furthermore, it is not known if existing tools that identify recent and strong positive selection from genomic sequences, in simple models of evolution, can detect recent positive selection when it operates on GRNs. Here, we propose a simulation framework, called EvoNet , that simulates forward-in-time the evolution of GRNs in a population. Since the population size is finite, random genetic drift is explicitly modeled. The fitness of a mutation is not constant, but we evaluate the fitness of each individual by measuring its genetic distance from an optimal genotype. Mutations and recombination may take place from generation to generation, modifying the genotypic composition of the population. Each individual goes through a maturation period, where its GRN reaches equilibrium. At the next step, individuals compete to produce the next generation. As time progresses, the beneficial genotypes push the population higher in the fitness landscape. We examine properties of the GRN evolution such as robustness against the deleterious effect of mutations and the role of genetic drift. We confirm classical results from Andreas Wagner's work that GRNs show robustness against mutations and we provide new results regarding the interplay between random genetic drift and natural selection.

Crowth, an identification platform for grapevine, olive and honey

November 14, 2018

3.1 Background

3.1.1 Introduction

The path from genotype to phenotype is characterized by an immense number of direct and indirect gene interactions. The relationship between genotype and phenotype has long been of interest to geneticists, developmental biologists and evolutionary biologists. This is partially because the relationship between genotypes and phenotypes is ambiguous and non-linearities appear often. The same phenotype can be produced by a range of genotypes and a single genotype can result in different phenotypes due to the environmental effects (Sansom and Brandon, 2007). Natural selection operates on various levels of genomic organization, from single nucleotides, genes, networks of genes to complex phenotypes. Phenotypic variation is the first of the three principles required for the action of natural selection (Lewontin, 1970). Thus, it may seem inconsistent that tests for localizing the action of natural selection, *i.e.* selective sweeps, use solely genotypic information, in models that incorporate no gene interactions or genotypic-phenotypic relations. In contrast, they utilize the concept of constant selection coefficient, which can be understood as a summary of the dynamics of the allele under selection, but lacks a clear biological meaning Chevin, 2008. If a genomic region is localized as the target of positive selection, the next step usually comprises an extensive literature search in an effort to connect the genotype to phenotype, and thus build plausible narratives that explain the action of positive selection (Pavlidis et al., 2012).

Chevin (2008) extended the theory of selective sweeps to the context of a locus that affects a quantitative trait, thus a phenotype, that harbors background genetic variation due to other, unlinked and no-interacting, loci. They assumed a large number of background loci with a small effect on the phenotype. Even though the increase in frequency of a beneficial mutation is slower than the classical one-locus selective sweep, they showed that under such a model, selective sweeps can still be detected at the focal locus, especially if the genetic variation of the background is not too large. Pavlidis, Metzler, and Stephan (2012) showed that when the train under selection is controlled by only a few loci (up to 8 in their simulations), it is possible that an equilibrium is reached, and thus no fixation of an allele. Such an equilibrium scenario happens more frequently when loci are characterized by having a similar effect on the phenotype. Contrariwise, if the population is far from the optimum and the focal allele has relatively large effect, then it will reach fixation. In general, multilocus model allow competition between loci, thus whether a locus will reach

fixation fast, and thus a selective sweep will be detected, depends crucially on the initial conditions.

To our knowledge, the first attempt to understand the evolution of regulatory networks was done in the seminal work by Wagner (1996). Wagner evolved numerically a network of genes that assume binary states (either on or off). He studied whether a population of such networks can buffer the (detrimental) effect of mutations after it evolves to reach its optimum. Indeed, he found (Figure 2 in (Wagner, 1996)) that after evolving a network of genes by means of natural selection (stabilizing selection), the effect of mutations is considerably lower than a system where evolution has not occurred yet. Natural selection, combined with neutral processes, modifies gene expression and in consequence the properties of GRNs. Ofria, Adami, and Collier (2003), using computer simulations, demonstrated that when the mutation rate is greater than zero, selection favors GRN variants that have similar phenotypes. Wagner (2008) demonstrated that neutral variants with no effect on the phenotype facilitate evolutionary innovation because they allow for thorough exploration of the genotype space. These ideas can be directly applied to GRNs by employing the concepts of robustness and redundancy. Robustness refers to the resilience that GRNs exhibit with respect to mutations. One mechanism for maintaining robustness is redundancy. Redundancy may be caused by/implemented by gene duplication or by unrelated genes that perform similar functions (Nowak et al., 1997).

Computational tools for detecting positive selection have been developed (Nielsen et al., 2005; Alachiotis, Stamatakis, and Pavlidis, 2012; Pavlidis et al., 2013) based on the "hitchhiking" or "selective sweep" theory (Maynard Smith and Haigh, 1974; Stephan, Wiehe, and Lenz, 1992). Three deviations from classic selective sweep theory are possible because of positive selection effects on GRNs: i) variation in selection intensity through time; ii) 'soft' sweeps that start with several favorable alleles; and iii) overlapping sweeps (Hermisson and Pennings, 2005). Since more than one network configuration can give rise to the same phenotype, the polymorphic patterns at the genome level are not necessarily expected to match the expected polymorphic pattern distribution that is caused by a strong beneficial mutation in just a single, independent gene. This has been shown for selective sweeps on a quantitative trait locus (Pavlidis, Metzler, and Stephan, 2012). Adaptation may often be based on pre-existing genetic variation of the population (standing genetic variation), rather than single, new mutations. Thus, it is expected that the new allele may originated from multiple initial alleles, which will in turn weaken the signal of positive selection (Przeworski, Coop, and Wall, 2005). Finally, if hitchhiking, as is widely believed, dominates the pattern of neutral diversity, the genome may be subject to multiple overlapping sweeps. Barton (1995) has extended earlier branching-process methods to determine how overlapping sweeps reduce mean coalescence time as well as how they reduce the fixation probability of favorable alleles.

In this work, we study via a forward-in-time simulator, named EvoNet , the evolution, by means of random genetic drift and selection, of a population of GRNs. We extend Wagner's classical model (Wagner, 1996) and subsequent extensions (e.g. (Siegal and Bergman, 2002)) by allowing cyclic equilibria during the maturation period and a different recombination model. We provide results about the robustness of the network to mutations, and its properties during evolution in a fitness landscape (e.g. genetic diversity). Furthermore, we study the Site Frequency Spectrum (SFS) signatures that the process leaves on neutral genomic regions linked with the genes of the GRN while the pop-

ulation climbs up the fitness landscape. In other words, we study whether we can use SFS-based neutrality tests, such as SweeD (Pavlidis et al., 2013), or SweepFinder (Nielsen et al., 2005), to detect the effects of selection.

Methods

The model

Regulatory regions define interactions: We assume a population of N individuals. Each individual comprises a set of n genes consisting of *cis* and *trans* binary regulatory regions, each of length L. A *cis* regulatory region is defined as the region upstream the gene on which other genes of the GRN can bind. Let $R_{i,c}$ be the *cis* region of the gene i and $R_{j,t}$ the *trans* region of gene j. Then, we define a function $I(R_{i,c}, R_{j,t})$ that receives as arguments two binary vectors and returns a float number in the $[-1, 1]$ representing the interaction strength. Negative values model suppression, positive values activation, whereas 0 means no interaction. Any function that takes as arguments binary vectors and returns a value in the $[-1, 1]$ could be used as the I function. Here, for the absolute value of interaction, we use the Equation 3.1:

$$|I(R_{i,c}, R_{j,t})| = \begin{cases} \frac{pc(R_{i,c}[1:L-1] \& R_{j,t}[1:L-1])}{L} \\ 0 : \text{no regulation} \end{cases} \tag{3.1}$$

where pc is the popcount function, which counts the number of set bits (i.e. 1s) that are common in the two vectors. The occurrence of interaction, as well as, the $+$ or $-$ *sign*, is defined by the last bit of the $R_{i,c}$ and $R_{j,t}$ vectors as:

$$\begin{array}{ll} 0, & R_{i,c}[L] = 0 \\ +, & R_{i,c}[L] = R_{j,t}[L] = 1 \\ -, & R_{i,c}[L] = 1 \text{ and } R_{j,t}[L] = 0 \end{array} \tag{3.2}$$

In other words, the first $L-1$ bits define the strength of the interaction, which is proportional to the number of common set bits (i.e. common 1s). The last (L^{th}) bit in each vector determines if the interaction is present and if it is suppression or activation. If the last bit of the *cis* element is '0' then it does not 'accept' any regulation. If it is '1', then regulation can be either positive or negative, depending on the last bit of the *trans* element.

The above representation of regulation enables a more realistic representation of regulation than Wagner's model (Wagner, 1996) and its more recent extensions (Siegal and Bergman, 2002; Huerta-Sanchez and Durrett, 2007). A single mutation in the *cis* region of a gene can affect its regulation by all other genes, and a mutation in the *trans* region of a gene can affect the way it regulates all other genes (see also the section 'Mutation model of regulatory regions').

Interaction matrix and expression levels: Interaction values are stored in a square $M_{n\times n}$ matrix of real values in the $[-1, 1]$, where n is the number of genes in the network. A positive M_{ij} value indicates that gene j activates gene i, a negative value indicates suppression and 0 represents no interaction. Thus, the row $M_{i\cdot}$ represents the interaction between all trans regulatory elements and the *cis*-regulatory region of gene i. Gene expressions are represented by a vector E_n of n elements. In the general case, the expression level E_j of the j_{th} gene can be a real positive number. Here, however, E is a binary vector, indicating only if a gene is switched on or off. Such a representation is more

efficient computationally. A similar approach has been used by Wagner (1996) and Siegal and Bergman (2002).

Inheritance of regulation and recombination: Each child inherits from his parents (the model allows for two parents or a single mother) the *cis* and *trans* regulatory regions. The initial values of expression levels (at birth) are defined solely by the environment , and here they are initialized to a constant binary vector. If the model allows for two parents, then recombination is possible to occur. We have implemented two recombination models. The first is similar to Wagner (1996)'s model that swaps rows of the interaction matrix of parents to form children. Such a model corresponds to tight linkage between the *cis* regulatory elements of a gene and recombination between genes. Wagner's model of recombination may be however unrealistic because it allows the some *cis* regulatory regions to be exchanged, however the *trans* regulation does not change. Thus, the *cis* regions can be exchanged but not the genes that correspond to the *cis* regions(Figure 3.1 top panel). In Wagner (1996), the interaction values between genes in the recipient and donor genomes remain unchanged after recombination (Figure 3.1, upper panel A). We implemented Wagner's model of recombination, but we re-estimated the interaction values between genes in the donor and the recipient genomes. This is necessary because *cis* and *trans* interactions are modified after recombination (Figure 3.1, upper panel B). We implemented an additional recombination model that allows cross-over events between parental genomes as follows: Assuming that n genes exist in the genome (members of the GRN), choose j, $0 < j < n$ an integer breakpoint. Then, the first j genes inherit both the *cis* and the *trans* regions from one parent, and the last $n - j$ genes inherit *cis* and *trans* regions from the other parent. The regulation between the first j and the last $n - j$ genes is re-computed from their regulatory regions (Figure 3.1, bottom panel).

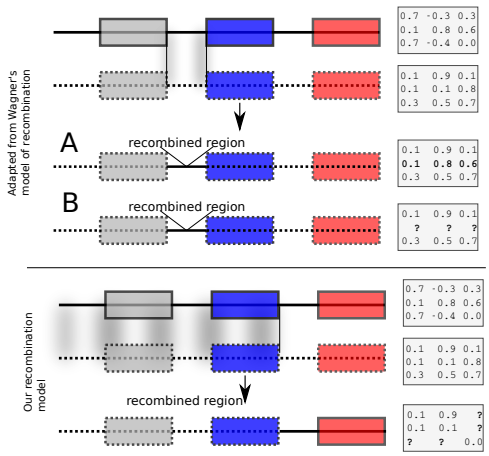

Figure 3.1: Recombination models implemented by EvoNet . Shaded areas show the gnomic regions that are exchanged due to the recombination process. At the upper panel, Wagner's model is illustrated, where *cis* regulatory regions can be swapped between individuals of the population. At the bottom panel, our model is shown. In our model, recombination is implemented via a recombination break-point. All genes at its left side inherit both the *cis* and the *trans* regions from one parent, whereas the genes on the right inherit *cis* and *trans* regions from the other parent. The interaction matrix is re-evaluated after recombination.

Mutations: Mutations take place in the *cis* and *trans* regulatory regions during offspring generation. Since regulatory regions are implemented as binary vectors, a mutation can change a position in a region by modifying a 0 to 1 and *vice versa*. On one hand, if a mutation will affect a *cis* region, then all interactions between this *cis* and all *trans* regions might be modified (i.e., the row of the interaction matrix will be affected). On the other hand, if a mutation will change a *trans* region, all interactions between this *trans* and all other *cis* regions might be modified (i.e., the column of the interaction matrix). For each individuals, the number of mutations is drawn from a Poisson distribution with parameter *mu* (mutation rate per genome per generation), and then mutations (if any) are placed uniformly among the *cis* and *trans* regulatory regions.

For example, let $R_{i,cis}$ be the *cis* regulatory region of gene i that is going to be mutated. $R_{i,cis}$ comprises two parts: the $[1 : L-1]$ part, which controls the strength of interactions and the L position that controls the type of interaction as described in *Regulatory regions define interactions*. Since mutations in the L position may have a dramatic effect, changing the type of interaction (e.g. a repressor might become activator or regulation can be silenced), we implemented two different mutation rates for these two parts of the regulatory regions. Mutations in the first $[1 : L-1]$ part are distributed uniformly. We model with 1% chance the probability that a mutation occurs and the trans region changes its behavior. This modeled the biological fact that mutations that change the nature of an established relationship of two genes is very rare in contrast to changing the strength of the respective relationship.

Selection: Selection operates on expression levels. In every generation selection is applied to select each parent of an individual. Let E_{opt} represent the optimal vector of expression values for the n genes, that is the optimal expression level for the first gene is $E_{opt,1}$, for the second gene $E_{opt,2}$ and so on. The fitness of an individual with expression values defined by the E_n vector is defined by:

$$F(E_n) = e^{-\|E_n - E_{opt}\|/\sigma^2} \tag{3.3}$$

where $\|E_n - E_{opt}\|$ is a norm of the difference between E_n and E_{opt} expression vectors (here the Euclidean distance is used). Parents are chosen proportionally to their fitness value $F(E_n)$.

Maturation and equilibria: Every 'new-born individual' has inherited the regulatory regions from its parents (potentially with mutations) and in addition it has acquired an initial expression vector (expression values for all genes) that is constant for all individuals. Since genes may interact with each other, we have implemented an additional 'maturation' process. During the maturation process the expression levels of genes change as a result of gene-gene interactions until either an equilibrium point or a cyclic equilibrium is reached. At the $t+1$ step of the process a new expression vector $E_n(t+1)$ is obtained using the expression vector of the t_{th} step and the interaction matrix M:

$$E_n(t+1) = ME_n(t) \tag{3.4}$$

Equivalently, the i^{th} element $E_n(t+1)[i] = \sum_{j=1}^{n} M_{i,j}E_n(t)[j]$. Depending on the interaction matrix M and the initial value of the expression vector E_n, there are 3 possible outcomes of this process.

$$\begin{array}{lll} (i) & E_n(t) = E_n(t+1) = E_n(t+2) = \ldots & \\ (ii) & E_n(t) = E_n(t+k) = E_n(t+2k) = \ldots, & k > 1 \\ (iii) & E_n(t) \neq E_n(t+j), & \text{for each } t, j \end{array} \tag{3.5}$$

In Wagner's model (Wagner, 1996) as well as in Huerta-Sanchez and Durrett (2007), only case (i) in Equation 3.5 is considered viable. Case (i) facilitates fitness evaluation of the individual using Equation 3.3. Individuals with a maturation process that concludes in (ii) or (iii) were removed from the population. Here, motivated by Pinho, Borenstein, and Feldman (2012) who suggested that in Wagner's model most networks are cycling, we developed a circadian framework to evaluate the fitness of individuals that conclude in cyclic equilibria during the maturation step. Individuals that conclude in case (iii), or individuals that conclude in case (ii) but the period k is greater than an upper threshold (here 10,000 steps) were considered non-viable and were removed from the population. Thus, if the maturation process concludes in case (ii), with $E_n(t) = E_n(t+k) = E_n(t+2k) = \ldots$ and $k > 1$, we evaluated the fitness of the individual as the minimum fitness value during the period of a cycle.

Results

Comparisons between Neutral Evolution and Selection Scenarios

Simulations setup

To explore the gene expression differences between neutral evolution and evolution under directional selection, we simulated neutral datasets and datasets with selection. For the two scenarios, command line arguments were identical except the random number generator seed and the binary flag that denotes whether simulation is neutral. All command lines are provided in the Supplement. Both models were evolved for 15,000 generations. Each individual network comprises 10 genes, each with 30-bit long *cis* and *trans* regulatory elements. The last bit of each regulatory element is responsible for the type of regulation (positive or negative; see Methods) and the remaining 29 bits determine the strength of the interaction, if any. In generation 0, all *cis*-regulatory elements were set to $000\ldots01000$, that is, initially they can not accept any regulation. In contrast, all *trans*-elements were set to $000\ldots01001$,*i.e.*, they are activators, thus they can regulate a *cis* element positively (provided that the last bit of the *cis*-element is 1). After maturation (see Methods), the expression vector was converted to binary format (the expression value is 1 if the expression is positive and 0 otherwise). Thus, initially all expression vectors v were equal to $\mathbf{0}$. The fitness of each person was evaluated after maturation. The optimum was set to the state were all genes were expressed (*i.e.*, state 1 for all genes). For the simulations with selection, the selection intensity $1/sigma^2$ (see Methods) was set to 1/5. The population size was set to 100 haploid individuals and remained constant throughout the entire simulation.

Optimum is gradually reached in a ladder-like fashion

We evaluated whether, and how, the population reaches the optimum state. Given that the initial state was 00000000 (i.e., all genes inactive) and the optimum state was 11111111 (*i.e.*, all genes active), the population had to experience the appropriate changes in its *cis*- and *trans*- regulatory elements, and consequently the GRN, to achieve the activation of all genes. We observed a ladder-like behavior for the average fitness (Figure 3.2); that is, networks were successively replaced by fitter networks in discrete steps.

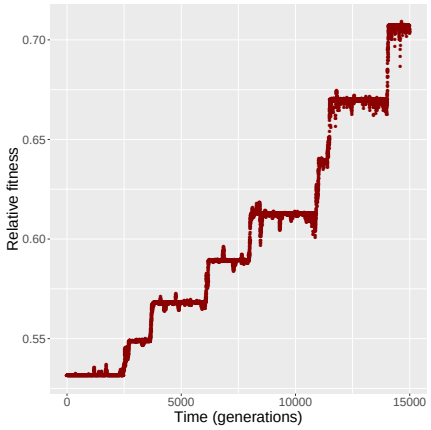

Figure 3.2: The increment in relative fitness of the population is taking place in discrete steps, in a ladder-like fashion.

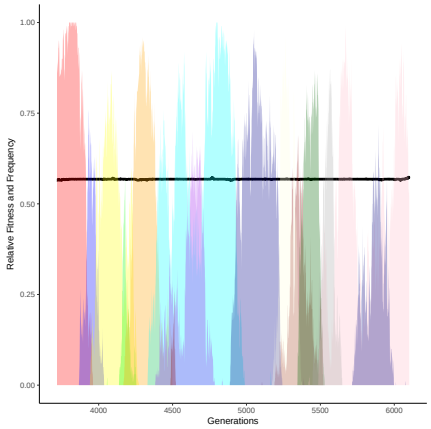

Figure 3.3: Alternating frequency-trajectories of the various regulatory networks at a certain fitness level (0.5679; black thick horizontal line). All networks have the same fitness. Here, we show only networks with frequency at least 50%. There are 14 different networks.

At every step of the 'ladder', the average population fitness remains approximately constant. After reaching each fitness step, the population starts exploring different GRN topologies until a fitter genotype establishes in the population. While exploring candidate topologies, genetic drift acts and it is therefore possible that the population will not incorporate every novel beneficial network topology that it will encounter. If a beneficial topology overcomes drift, its frequency increases and the average population follows. Finally, when the new topology reaches fixation, the population has reached the next step in the fitness 'ladder' (Figure 3.3).

Mutations are the driving force behind the exploration of the topology space, since each mutation may represent a novel network topology. By increasing the mutation rate, the number of novel explored topologies increases and waiting times between each step are decreased. (Figure 3.4).

Recombination rates also affect the time required for each step. Recombination allows the parental networks to be combined resulting in enhancement of the network variability in the population, thus the optimum can be reached faster. In our simulations our proposed model R1R2 swapping reaches optimum faster than the row-swapping model proposed by Wagner Wagner, 1996 (Figure 3.5).

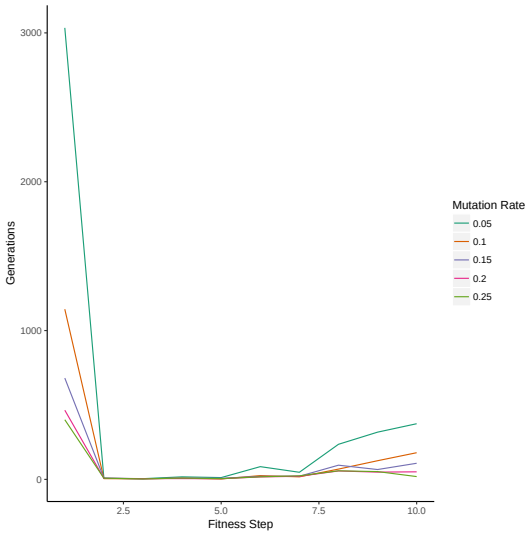

Figure 3.4: An increase to mutation rate reduces the time needed to take the next step on the fitness landscape.

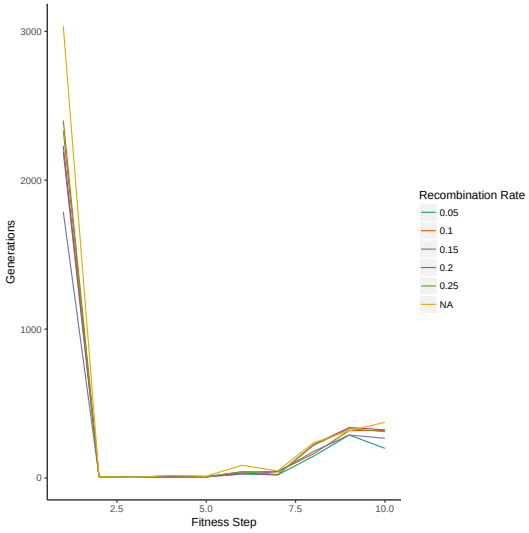

Figure 3.5: Recombination rates and time needed to take the next step on the fitness landscape. Especially for the first step, which takes most of the time, the least time is achieved when recombination rate is 0.15, i.e. intermediate between the minimum and the maximum.

Size of the regulatory space in neutrality and selection

We assessed how the population explores the state space of regulatory networks during its evolution, by evaluating the number of different genotypes individuals obtain. We studied whether neutrality or selection explores the space more efficiently, *i.e.*, which of the two processes allow the population to explore a higher number of genotypes on average.

During the course of evolution, for 15,000 generations, both neutral and selection scenarios experienced a multitude of GRNs. In the selection scenario, the population encountered 17,110 different networks; under neutrality the population experienced only 5,105 GRNs. This means that under selection the population is able to explore a greater part of the space of GRNs than under neutrality. Due to selection pressure, the population moves towards the optimum via genotypes that are optimal at the given time point. Then, due to drift, it explores genotypes with the same fitness (*i.e.*, effectively neutral) until a new optimal genotype overcomes drift and brings the population to the next fitness level.

Under neutrality, the behavior of the population was different. With the selection pressure absent, the fate of genotypes was affected solely by genetic drift. In the limited amount of generations (15,000), the population explored a small fraction of the genotypic space centered around the initial state.

3.1.2 Choice of recombination model and shape of fitness landscape affect time to reach optimum fitness

Different optimal states model different fitness landscapes. EvoNet will reach the optimal state regardless of the shape of the fitness landscape. However the time needed for each landscape change is based on the optimal state. A population following our R1R2 recombination model reaches the optimum faster than a non-recombining population in the cases of the optimal states 1111111111 and 1111100000 (Figure 3.6). On the other hand, for the optimal states 1100110011 and 1010101010, recombination makes the population reach the optimum slower than the non-recombination scenario.

3.1.3 Robustness of Gene Regulatory Network

Robustness to the (phenotypic) effect of mutations has been studied in the framework of GRNs Wagner, 1996, demonstrating that GRNs which reached the phenotypic optimum are less sensitive to mutations, a phenomenon named epigenetic stability. Thus, epigenetic stability was attributed to the evolution of GRNs via the selection process. At discrete time-points EvoNet clones the evolving population ('core' population) creating a 'branch' population. Each 'core' individual has an interactions matrix M_i shared with its 'clone'. The 'branch' population mutates further and then both populations start the maturation progress. The interaction matrices are, then, discretized (D_i, $D^{'}{}_i$).

We assess the GRN robustness at two levels, topology and phenotype. Each GRN has a unique network topology characterizing the strength and effect of all gene interactions. In EvoNet , the topologies are modelled by the interactions matrix, so the additional mutations occuring in the 'branch' population have the potential to change the network's topology. The topology robustness measures if the 'core' and 'branch' networks represent the same network topology after the incorporation of the additional mutations on the 'branch' population.

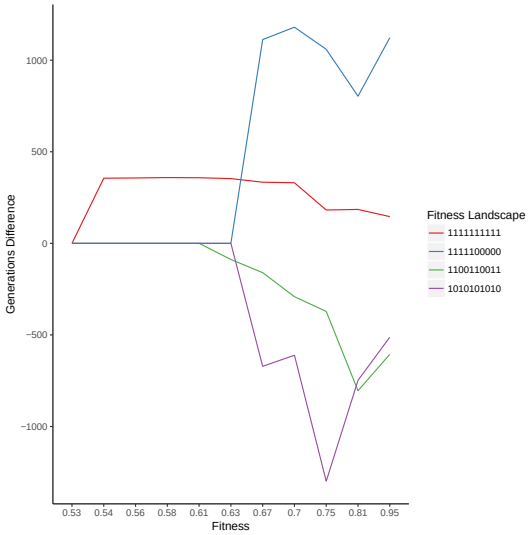

Figure 3.6: Non-recombining population needs more time to navigate the landscape than recombining population for the 1111111111 and 1111100000 cases. On the other hand the optimum is reached faster for the non-recombining populations when the optimum is set to 1100110011 and 1010101010.

Phenotypic robustness measures differences in the (binary) expression vector between the two populations after every branching. (Figure 3.7).

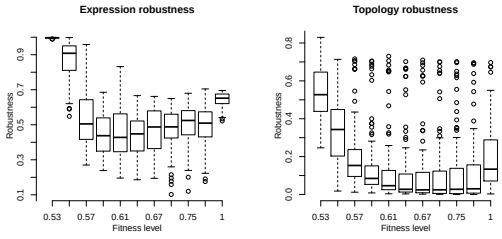

Figure 3.7: Robustness of the (binary) expression vector and network topology. Initially, the robustness of the expression vector is very high due to the initialization of the simulator. The initial interaction matrix results in the $00\ldots0$ expression vector. Since no interaction is possible in the beginning, the initial state is robust to mutations. Robustness falls dramatically after the initialization step and increases as fitness increases. The maximum robustness is achieved when the optimum has been reached, on average. The topology is less robust than then expression vector (bottom plot). However, robustness of topology also increases when the population has reached the maximum fitness level.

3.1.4 Effect of neutral genes

All genes in a GRN are not subject to the same evolutionary pressure. Often, a subset of the GRN is evolving under neutrality while other parts are under selection. Using EvoNet we inferred that the interactions between neutrally evolving genes and selected genes are negatively correlated with the average population fitness. When the fitness is low, there are multiple interactions between the two parts, due to the fact that a beneficial mutation in the neutral cluster has a positive effect on the GRN. In contrast, as the fitness increases,

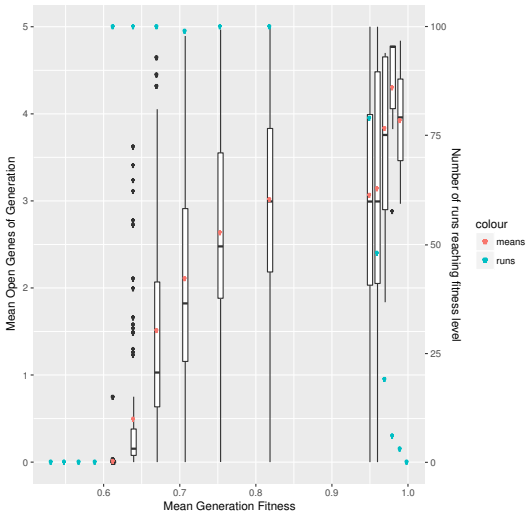

Figure 3.8: It is beneficial for the GRN to have open connections (red points) with neutrally evolving genes outside the GRN when the population is ascending the fitness landscape (bars). Upon reaching optimum fitness those interactions tend to be discarded. Barplots depict the results of 100 simulations, where the majority (blue points) reached each fitness step.

the majority of mutations, on either part, are deleterious resulting in disadvantageous interactions. Since mutations happen with the same rate across all clusters, the GRN minimizes the chance that a deleterious mutation will affect it, by gradually discarding the interactions between the different clusters. By doing so, the network avoids the consequences of deleterious mutations on the neutral cluster while protecting the selection cluster. (Figure 3.8).

3.1.5 Competition Between GRNs of Different Length

We examined whether the size of the GRN is itself a feature on which selection may operate. Thus, we created two distinct GRNs and we let them evolve in the same population. The first GRN, G_s, consists of five genes under selection. The second GRN, G_l consists of seven genes. In both GRNs the rest of the genes (five and three, respectively) evolve neutrally. In addition, G_s could not regulate the *trans* region of half of the neutral-evolving genes to simulate a slower mutation rate outside the GRN, whereas the second GRN was free to regulate everything. During the competition between the GRNs, G_l dominated G_s even though G_s had fewer genes under selection so deleterious mutations occurred less frequently. The lack of regulation on the trans-region prohibited G_s from reaching the critical fitness level after which the neutral gene interaction are phased out.

3.1.6 GRN effect

Robustness against mutations is an emergent feature of the GRN (Krishnan, Tomita, and Giuliani, 2008). By comparing EvoNet with another algorithm that omits the GRN and directly switches on and off the genes, we demonstrate that the existence of the GRN gives rise to mutational robustness and therefore reaching the fitness optimum faster at high mutation rates. For small mutation rates, robustness and the resulting buffering of mutations happening in EvoNet hinders the acquisition of fitness optimum. When the mutational load increases,

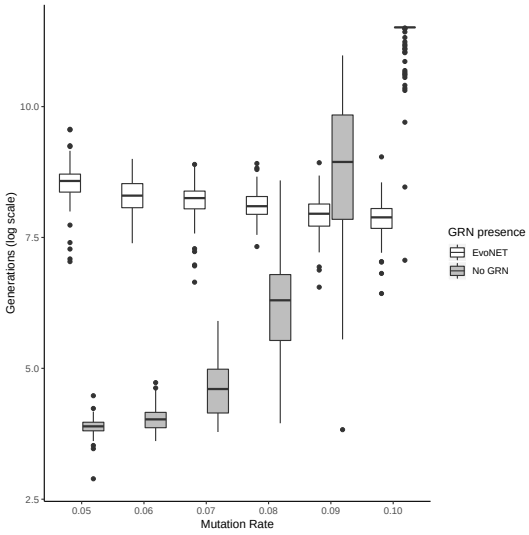

Figure 3.9: Comparison between the time (in generations) needed to reach the fitness optimum between EvoNet (white) and a similar simulator that directly switches on and/or off genes without employing a GRN (gray boxes). For lower mutation rates, robustness buffers the immediate effect of mutations. Thus, EvoNet reaches the optimum slower than the alternative approach that does not employ GRNs. When the mutation rate increases, mutations, on one hand slow down the simulator without the GRNs. On the other hand, they do not have a detrimental effect on EvoNet due to the buffering effect of the GRN.

however, EvoNet reaches optimum fitness faster due to the robustness created by the GRN. (Figure 3.9)

3.2 Discussion

In this study, we developed EvoNet that creates detailed models of GRNs, thus, enabling the investigation of GRN evolution in population level. EvoNet extends the algorithm proposed by Wagner Wagner, 1996, by simulating the *cis* and *trans* gene regions creating a more realistic model of GRN. The regulatory *cis* and *trans* regions interact to create the gene interaction matrix which was the basis of Wagner's model (Wagner Wagner, 1996 directly mutates the interaction matrix). EvoNet employs the following processes in every discrete generation: birth (with or without recombination), mutation, maturation and fitness calculation. The birth phase is represented by the inheritance of the *cis* and *trans* regions from the previous generation. We introduced a new recombination model (R1R2) that is more realistic than the previously used row-swapping model by Wagner, 1996. The R1R2 model has a similar behaviour with Wagner's row swapping model regarding the average time needed for every fitness level (Figure 3.10). Next, mutations happen, affecting the *cis* and *trans* regions. *cis* and *trans* regions interact to create a new interaction matrix. EvoNet models the type of interaction using the formula shown in Equation 3.1. In the maturation phase, the phenotype is obtained. In contrast to previous studies, we handled the cyclic equilibria instead of discarding them (Wagner, 1996) and we evaluated their fitness, making the evolution model more realistic.

In the simulations where the mutation rate is sufficiently low, we observed that the fitness landscape takes a ladder-like shape. The steps of the ladder represent the time (measured in generations) that the population explores the

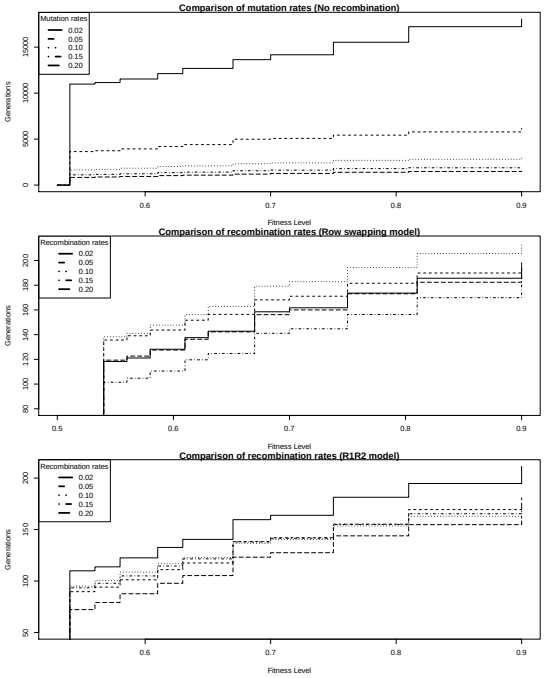

Figure 3.10: Comparison between the no-recombination model, the Wagner's swapping and the R1R2 recombination model regarding the generations required to reach optimum fitness

genotype space by searching different network topologies allowing for the next step in the fitness 'ladder". Small increases in the mutation and recombination rates result in a decrease in the exploration time (Figure 3.10) due to the increased number of mutations permitting a quicker exploration of candidate network topologies.

We explored the role of robustness of the GRNs while they undergo selection. Robustness is important because it verifies the existence of phenotypically neutral mutations and allows for complex biological structures that are robust to the detrimental effects of mutations. There are two robustness levels acting as canalization attempts, the network topology and the phenotype. Phenotype is more robust to mutations than network topologies, since topology is more directly related to the regions affected by mutations. By comparing EvoNet with a GRN-less simulation program we conclude that these robustness levels permit the GRN to increase its fitness even under high mutation rate. In lower mutation rates, robustness acts as a barrier on the effect of all mutations driving the population to a flat network space thus avoiding perturbations (Lenski, Barrick, and Ofria, 2006)]. In contrast, when the mutation rate increases, the GRN robustness limit is overcome and deleterious mutations, which are more frequent, are immediately affecting the network. GRNs are able to buffer the detrimental effect of mutations, highlighting their biological significance.

Each GRN interacts with other genes and GRNs, which may be evolving with the same rate as our principal GRN or not. By using EvoNet to simulate neutrality and selection acting on parts of the GRN we can draw conclusions on these interactions' effect. While those interactions are beneficial at a lower average population fitness level, they are disadvantageous at higher levels of fitness. As the GRN ascends the fitness landscape, those interactions are removed. A plausible explanation is that the GRN manages to achieve higher robustness level by removing unnecessary genes and also avoids the effect of deleterious mutations happening on the additional genes of the GRN.

3.3 Conclusion

Gene Regulatory networks play a vital role in the development of evolutionary advantageous traits for all organisms. In this study we have presented EvoNet , a versatile simulator for the evolution of GRNs through means of genetic drift and selection. Through the use of EvoNet we were able to identify new levels of genetic robustness as well as verify the findings of previous research. EvoNet is freely available for download and modification from `https://github.com/antokioukis/evonet`.

Chapter 4

Brassica Cretica

Draft Genome Sequence, natural selection targets and demography history of the wild crop relative Brassica cretica Lam.

Abstract

Wild crop relatives contain an incredible amount of genetic diversity, representing an invaluable resource for crop improvement. Many of their traits have the potential to help crops become more resistant and resilient, and adapt to the new conditions that they will experience due to climate change. An impressive global effort occurs for the conservation of various wild crop relatives and facilitate their use in crop breeding for food security. The genus Brassica is listed in Annex I of the International Treaty on Plant Genetic Resources for Food and Agriculture. Brassica oleracea (or wild cabbage) is a species native to coastal southern and Western Europe that has become established as an important human food crop plant because of its large reserves stored over the winter in its leaves. Brassica cretica Lam. is a wild relative crop in the brassica group. There have been proposed three native subspecies in Europe: B. cretica ssp. aegaea; B. cretica ssp. cretica, and B. cretica ssp. nivea. The species B. cretica has been proposed as a potential gene donor to a number of crops in the brassica group, including broccoli, Brussels sprout, cabbage, cauliflower, kale, swede, turnip and oilseed rape. Here, we present the draft de novo genome assemblies of four B. cretica individuals, including two B. cretica ssp. nivea and two B. cretica ssp. cretica. De novo assembly of Illumina MiSeq genomic shotgun sequencing data yielded 243,461 contigs totalling 412.5 Mb in length were generated, corresponding to 122% of the estimated genome size of Brassica cretica (339 Mb). According to synteny mapping and phylogenetic analysis of conserved genes, B. cretica is related to ??? proteins.

4.1 Introduction

4.1.1 Wild crop relatives

Many plant species are used in the food and agriculture market, however, 30 crops account for the 95% of food production worldwide (Brozynska, Furtado, and Henry, 2016). Domesticated crops, used in the food production, show reduction in the genetic diversity, compared to their respective Wild Crop Relatives (CWRs). During the last years of continuing growth trend of productivity and crop uniformity, this "domestication bottleneck" (Tanksley and McCouch, 1997) may lead to loss of valuable genetic alleles. On the other hand, during

the domestication process of cultivated varieties with wild species, additional genetic diversity may arise (Hufford et al., 2013; Sawler et al., 2013).

As wild species of crops, in nature, continue to evolve under abiotic and biotic stresses, it is very important to conserve this genetic biodiversity, which can be useful for agriculture (in situ conservation). Seed banks or germplasm collections are also important to preserve as another resource for agriculture (ex situ conservation). The total genome sequencing of several CWRs may be used first to characterize wild populations and help their conservation. While, from the other hand, the analysis of the sequence will point out the genetic variation and important genetic characters, which probably have been lost during domestication, and that could be transfer into crop species to support food security, climate adaptation and nutritional improvement (Brozynska, Furtado, and Henry, 2016). Since the improvement of the newest technologies, regarding the precision and sequencing read lengths (e.g. Illumina technologies: MiSeq, HiSeq, or the Pac Bio platform), sequencing of a bigger number of wild species, respective to domesticated crops, is now achievable.

During the last decades, there are some remarkable examples of introducing favored traits, from CWRs in their respective domesticated crop plants. In most cases, these traits concern about resistance to biotic stresses, such as late blight resistance to Phytophthora infestans from the wild potato Solanum demissum Lindl. (Prescott-Allen and Prescott-Allen, 1986; Witek et al., 2016). Besides biotic tolerance, there have been identified and/or introduced many quantitative trait loci, regarding the grain quality for increased yield, such as from Oryza rufipogon, a wild species of rice, to Oryza sativa (Eptiningsih and Trijatmiko, 2003)and grain hardness from Hordeum spontaneum (wild barley) (Li et al., 2010a).

4.1.2 Brassica oleracea

Brassica oleracea is a very important domesticated plant species, comprising of many vegetable crops as different cultivars, such as cauliflower, broccoli, cabbages, kale, Brussel sprouts, savoi, kohlrabi and gai lan. Brassica oleracea or wild cabbage belongs to the family of Brassicaceae and is found in coastal Southern and Western Europe. The species has become very popular because of its high content to nutrients, such as vitamin C, its anticancer properties (Higdon et al., 2007), as well as the high food reserves in its leaves.

B. oleracea constitutes the one of the three diploid Brassica species in the classical triangle of U (1935) (genome: CC), that contains nine chromosomes. The other two species in this group are B. rapa (genome: AA) with 10 chromosomes and B. nigra (the black mustard) (genome: BB) with 8 chromosomes. These three species, as they are closely related, gave rise to new allotetraploids species that are very important oilseed crops, the B. juncea (genome: AABB), B. napus (genome: AACC) and B. carinata (genome: BBCC). There is evidence for each of the Brassica genomes to have undergone a whole-genome duplication (Bowers et al., 2003; Jiao et al., 2011) and a Brassicae-lineage-specific-whole-genome triplication, which was followed after the divergence from Arabidopsis lineage (Lysak et al., 2005; Wang et al., 2011).

In 2014, Liu and et al ((Liu et al., 2014)) reported a draft genome of B. oleracea var. capitata and a genomic comparison with its very close sister species B. rapa. A total of 45,758 protein-coding genes were predicted, with mean transcript length of 1,761 bp and 3,756 non-coding RNAs (miRNA, tRNA, rRNA and snRNA). It is observed that there is a greater number of transposable

elements (TEs) in B. oleracea than in B. rapa as a consequence of continuous amplification over the last 4 million years (MY), the time that the two species were diverged from a common ancestor, whereas in B. rapa the amplification is made mostly in the recent 0.2 MY (Fig. 2b, (Liu et al., 2014). Moreover, there has been observed massive gene loss and frequent reshuffling of triplicated genomic blocks, which favored over-retention of genes for metabolic pathways.

Brassica cretica Lam. is a wild crop relative of a big number of crops of the genus Brassica, proposed to be the ancestor of broccoli, Brussel sprouts, cabbage, cauliflower, kale, swede, turnip and oilseed rape. The species is found in Eastern Meditteranean region, mainly on Crete and the surrounding Aegean islands, where it grows in isolated populations in cliff systems and ravines (Snogerup, Gustafsson, and Von Bothmer, 1990). There are three subspecies, B. cretica ssp. aegaea; B. cretica ssp. cretica and B. cretica ssp. nivea. Since this species is thought to be the gene donor of many crops of the brassica group, it might contain genes that are not included in the domesticated crops, as well as a different set of NLRs. The analysis of the NLRsome of wild species will help us find which genes or locus are responsible for the recognition of effectors from important phytopathogens and thus create resistant plants in the field via transfer of these favored genes/locus (Chen et al., 2013).

Here, we present the first draft de novo genome assemblies of four individual Brassica cretica species (two B. cretica ssp. nivea and two B. cretica ssp. cretica).

4.2 Materials and Methods

4.2.1 Genomy Assembly

Prior to assembly, Illumina MiSeq sequence reads were filtered on quality scores and trimmed to remove adapter sequences using trimgalore (`https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`) with q = 30. Reads were assembled into contigs using SOAPdenovo2 (Luo et al., 2012) with k = 127. Contigs shorter than 500 bp in length were removed.

For comparison with Brassica oleracea var. oleracea (wild cabbage) and for variant calling between B. cretica individuals, we aligned B. cretica MiSeq reads against the previously published reference genome sequence (GenBank: GCA_000695525.1) (Parkin et al., 2014) using BWA (Li and Durbin, 2009). SNP calling was performed as previously described () Yematew et al. 2018).

Genome annotation was performed using the MAKER pipeline (Campbell et al., 2014; Cantarel et al., 2007). Ab initio gene prediction was performed using Augustus (Stanke and Waack, 2003) trained on Arabidopsis. Amino acid sequences predicted by MAKER were subjected to analysis with PfamScan to identify those predicted proteins containing an NB-ARC domain (Finn et al., 2013).

4.3 Results and Discussion

4.3.1 Detection of natural selection

After the genome assembly of Brassica cretica, we mapped the resulting contigs on the Brassica oleracea genome. This allowed us to use all specimens without sacrificing one as a reference and also provided us with a closely related out-

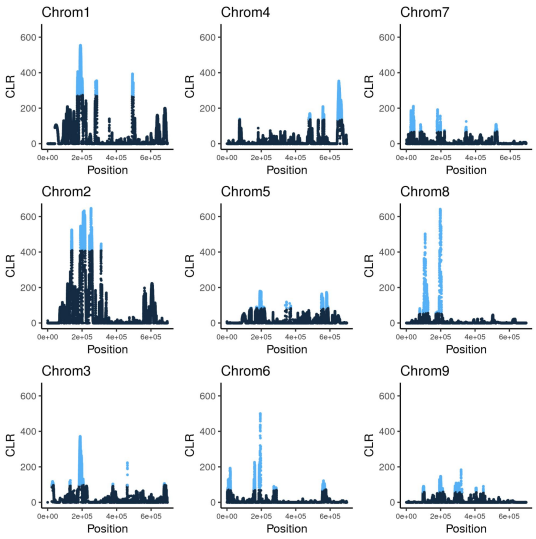

Figure 4.1: Selective sweeps of brassica cretica reads mapped Brassica oleracea. Ciel points indicate high likelihood of selection, blue indicates low likelihood.

group. Following the GATK best practices pipeline Auwera et al., 2013, this mapping resulted in approximately six million single nucleotide polymorphisms (SNPs). Brassica olearacea has been examined thoroughly in the past and there is a gene list of the organism organised into chromosomes. We used this list to exclude SNPs with a distance less than 10kb from those coding regions. This process of removing SNPs is necessary due to hitchhiking effect and later stages of our analysis pipeline require it.

To detect selection targets we used the selective sweep theory (Maynard Smith and Haigh, 1974) implemented in SweeD (Pavlidis et al., 2013). When a beneficial mutation arises and starts spreading in the population, the hitchhiking effect reduces genetic variation around the point of mutation thus creating a so-called selective sweep. After the fixation of the beneficial allele there is no diversity in the selected site and patterns of linkage disequilibrium (LD) emerge around the target site of the beneficial mutation. Searching for selective sweeps around the SNPs we can identify regions where natural selection has acted since a selective sweep increases linked neutral or weakly selected variants. This hitchhiking effect drastically reduces genetic variation near the positively selected site, thus creating a selective sweep. Creating an allele frequency spectrum(AFS) on the whole genome enables us to identify selective sweeps and as a result define where natural selection acted.

Allele frequency, is the relative frequency of an allele at a particular locus in a population, expressed as a percentage. AFS is the the histogram of these frequencies, with each entry grouping all the loci with that frequency. (singletons, doubletons...) Each locus contributing to the AFS is assumed biallelic and neutral to the changes of frequencies of other loci.

4.3.2 Demographic Model Inference

The SNPs were converted to ms format (Hudson, 2002) for speeding up the rest of the analysis pipeline. We applied logistic Principal Component Analysis (`http://arxiv.org/abs/1510.06112v1`) (logPCA) for differentiating the number of the different populations of the plants (Figure 4.2)

Based on the logPCA results we identified 2 populations. The first comprising three individuals (A,B,D) and the second containing one (C). It is important

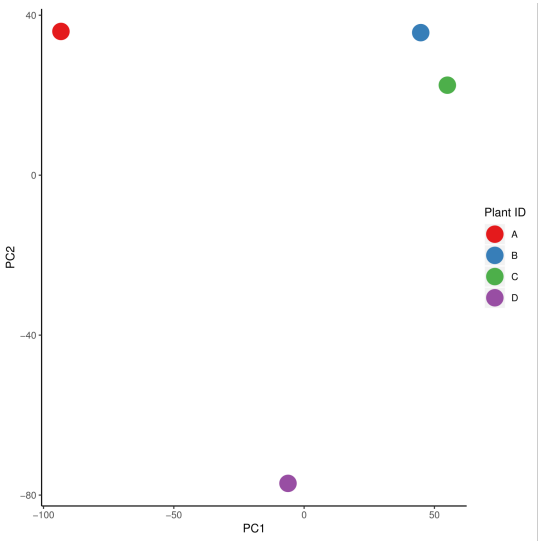

Figure 4.2: logPCA on binary SNPs plotted on 2 dimensions clearly shows 2 populations (A,B,D) and (C) along the eigenvector 1 axis.

to note that despite the fact that the A,B,C plants were sampled from Central Greece and D from Crete, logPCA shows that the cretan individual is genetically closer to A than the rest of its geographic neighbors.

The difference of euclidean distance of A,D plants to the cluster of B,C is small so we can not be certain of the final clustering. In our analysis we have used a second population schema where the B,C,D individuals compose one population and A another. We have used both to identify different demographic models.

Inferring a demographic model consistent with a particular data set requires random walks into a large parameter space by simulating the model using Monte Carlo coalescent-theory based approaches. The main handicap of these methods is their scalability to genome-wide size data sets. Another issue arises when multiple populations are free to interact through migration (either symmetric or asymmetric) resulting in an increase of the required complex calculations. These complexities hinder any effort to thorough explain the statistical properties of the summary statistics produced during the walks. To avoid these problems we based our demographic model inference on the multi-population allele frequency spectrum (AFS) (Bustamante et al., 2001; Caicedo et al., 2007; Hernandez, Williamson, and Bustamante, 2007; "Darwinian and demographic forces affecting human protein coding genes" 2009). Allele frequency, is the relative frequency of an allele at a particular locus in a population, expressed as a percentage. AFS is the the histogram of these frequencies, with each entry grouping all the loci with that frequency. (singletons, doubletons...) Each locus contributing to the AFS is assumed biallelic and neutral to the changes of frequencies of other loci. The demographic history of a population and selection affect allele frequency dynamics, reflected in the allele frequency spectrum. By comparing the different spectra produced by simulations and observations we can access the model's goodness of fit and estimate the best parameters for each model.

In spite of the existence of efficient algorithms for the simulation of a single population AFS (Adams and Hudson, 2004; Marth et al., 2004) (Williamson et al. 2005), the joint AFS between two populations still requires very computationally intensive coalescent simulations. However, approximations of the joint-AFS using a numerical solution of a diffusion equation have been used

extensively in the past (Der Sarkissian et al., 2015), enabling simulations of a joint-AFS for more than two simultaneous populations in a reasonable computation time. Although the diffusion approach neglects linkage we can use composite likelihood function as a consistent estimator for evaluating genetic scenarios. Concerns about the use of composite likelihood in population genetics are overcome by allowing conventional and parametric bootstrap of the data.

The dadi python package (Gutenkunst et al., 2009) implements these approximations and in conjunction with the dadi pipeline described in (Portik et al., 2017) allows for adequate exploration of the parameter space. The dadi pipeline consists of three optimization rounds and a final plotting step. We used 30 demography models ranging from simple (populations never diverge) to complex (ancient divergence with asymmetric migrations between the two populations) to find the best fitting model.

The initial two rounds of optimizations search the parameter space for the parameter set that best describes the data under each of the models. For every model we sampled 50 different parameter sets performing 50 repetitions of the each set to get the actual global maximum for each model while avoiding local maxima. We based our selections of the best parameter sets on the AIC score for each model. To assess which demographic model better reflects the true demographic history of the Brassica cretica population a simple comparison between the respective AIC scores from each model is not valid because AIC is not comparable between different models. We compared the models using relative AIC weights. For each model, we calculated the differences in AIC with respect to the AIC of the best candidate model. With a simple transformation we can calculate an estimate of the relative likelihood Li of each model. By dividing each Li with the sum of L we can normalize the weights and compare the models.

After this calculation, we selected the Founder event and discrete admixture, two epoch model, as the most possible demography model for the first population schema and Divergence with continuous symmetric migration and instantaneous size change.(Figure 4.3). The first model specifies that the original population split into two subgroups that allowed symmetric migration between them, continuing the population size of each subgroup changed, whereas the second model allows the subpopulations to migrate as the time progresses and the second subpopulation experiences a population size change. (Figure 4.4).

Figure 4.3: From top right to bottom left: 1) Calculated AFS from B.cretica data sets, split by populations. 2) Simulated AFS of the best fitting model from final dadi simulations. 3) Heatmap of the residual errors from the comparison between real and simulated AFS. 4) Barplot of the same comparisons.

Figure 4.4: Schematic of the proposed demography model showing the creation of the 2 different subpopulations, the different time periods before and after complete isolation and the migration occurring during T1 time period.

We have not assigned biological meaning to any of the parameters because we aimed at model selection.

The top 3 AIC relative weights models for each population schema are:

ABD-C clusters	
Model Name	Relative AIC weight
Founder event and discrete admixture	1
Divergence with ancient symmetrical migration	1.67e-91
Divergence with no migration	5.08e-211

BCD-A clusters	
Model Name	Relative AIC weight
Divergence with symmetric migration, size change	1
Vicariance with late discrete admixture	9.59e-20
Founder event and discrete admixture, two epoch	4.11e-75

In tables (4.3.2,4.3.2) we show the top 3 AIC relative weights models.Based on the differences of the relative AIC weights of the top 3 models for each schema we are confident that the selected models are the most accurate simulation of the demographic history of Brassica cretica.

4.4 Conclusion

Brassica Cretica is an important as a wild relative of the Brassica taxa and its possible commercial use. Due to its wild status, it is a viable candidate for detection of NLR genes that are not preserved in the domesticated species. Detection of those genes is only possible through the genome assembly we have performed. The demography model of Brassica Cretica is vital in helping us understand the population schema and help us identify the history of domestication of the related species. We hope that this work will be used as the foundation for further examination of Brassica Cretica such as gene annotation.

Chapter 5

Conclusion

The identification of the products we consume is increasingly important due to globalization. With the use of NGS technologies, Crowth aims to empower every consumer and producer to accurately identify and quantify their products by providing a user-friendly and easily-accessible solution. By specializing in grapevine, olive oil and honey we have set a baseline of what is achievable by using data currently available. Crowth is easily extendable for inclusion of different products and even use of another genetic neighborhoods for identification besides the current ITS regions.

Gene Regulatory networks affect every stage of an organism's life from aging to sexual attractiveness. EvoNet simulates their evolution through means of genetic drift and selection. By using EvoNet , we shed light to new levels of genetic robustness and their biological significance. The models build upon Dr. Wagner's research by adding two regions *cis*and *trans*that interplay to create the genetic interactions matrix instead of starting at the matrix level. EvoNet incorporates cyclic equilibria that previous research discarded, while removing the limitation of the population size and generations forward in-time by adhering to the view that the current generation is only affected by the most recent generation and not the others.

The Brassica taxa contains some of the planet's most commercially important plants. Brassica Cretica is a wild relative of this taxa and is a source for the detection of new NLR genes. By assembling its genome and identifying its demographic history we can harness this knowledge for improving our domesticated species. We hope that the work presented here will be used as a foundation for further examination of Brassica Cretica

Bibliography

Adams, Alison M and Richard R Hudson (2004). "Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms". In: *Genetics* 168.3, pp. 1699–1712.

Alachiotis, N., A. Stamatakis, and P. Pavlidis (2012). "OmegaPlus: A scalable tool for rapid detection of selective sweeps in whole-genome datasets". In: *Bioinformatics* 28.17. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts419.

Auwera, Geraldine A Van der et al. (2013). "From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline". In: *Current protocols in bioinformatics* 43.1, pp. 11–10.

Barton, N H (1995). "Linkage and the limits to natural selection". In: *Genetics* 140.2, pp. 821–841. ISSN: 0016-6731. URL: http://www.ncbi.nlm.nih.gov/pubmed/7498757.

Bowers, John E et al. (2003). "Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events". In: *Nature* 422.6930, p. 433.

Brozynska, Marta, Agnelo Furtado, and Robert J Henry (2016). "Genomics of crop wild relatives: expanding the gene pool for crop improvement". In: *Plant Biotechnology Journal* 14.4, pp. 1070–1085.

Bustamante, Carlos D et al. (2001). "Directional selection and the site-frequency spectrum". In: *Genetics* 159.4, pp. 1779–1788.

Caicedo, Ana L et al. (2007). "Genome-wide patterns of nucleotide polymorphism in domesticated rice". In: *PLoS genetics* 3.9, e163.

Camacho, Christiam et al. (2009). "BLAST+: architecture and applications". In: *BMC bioinformatics* 10.1, p. 421.

Campbell, Michael S et al. (2014). "Genome annotation and curation using MAKER and MAKER-P". In: *Current Protocols in Bioinformatics* 48.1, pp. 4–11.

Cantarel, Brandi L et al. (2007). "MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes". In: *Genome research* 18.1, pp. 000–000.

Chen, Guangdeng et al. (2013). "A novel and major quantitative trait locus for Fusarium crown rot resistance in a genotype of wild barley (Hordeum spontaneum L.)" In: *PloS one* 8.3, e58040.

Chen, Shilin et al. (2010). "Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species". In: *PloS one* 5.1, e8613.

Chevin, Luis-Miguel et al. (2008). "Selective sweep at a quantitative trait locus in the presence of background genetic variation". In: *Genetics* 180.3, pp. 1645–1660.

Coleman, Annette W (2003). "ITS2 is a double-edged tool for eukaryote evolutionary comparisons". In: *TRENDS in Genetics* 19.7, pp. 370–375.

— (2007). "Pan-eukaryote ITS2 homologies revealed by RNA secondary structure". In: *Nucleic Acids Research* 35.10, pp. 3322–3329.

Coleman, Annette W (2009). "Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide". In: *Molecular Phylogenetics and Evolution* 50.1, pp. 197–203.

"Darwinian and demographic forces affecting human protein coding genes" (2009). In: *Genome research*, gr–088336.

Der Sarkissian, Clio et al. (2015). "Evolutionary genomics and conservation of the endangered Przewalski's horse". In: *Current Biology* 25.19, pp. 2577–2583.

Edgar, Robert C (2016). "UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing". In: *bioRxiv*. DOI: 10.1101/081257. eprint: https://www.biorxiv.org/content/early/2016/10/15/081257.full.pdf. URL: https://www.biorxiv.org/content/early/2016/10/15/081257.

Edgar, Robert C. and Henrik Flyvbjerg (2015). "Error filtering, pair assembly and error correction for next-generation sequencing reads". In: *Bioinformatics* 31.21, pp. 3476–3482. DOI: 10.1093/bioinformatics/btv401. eprint: /oup/backfile/content_public/journal/bioinformatics/31/21/10.1093_bioinformatics_btv401/2/btv401.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btv401.

Eptiningsih, EM and KR Trijatmiko (2003). "Identification of quantitative trait loci for grain quality in an advanced backcross population derived from the Oryza sativa variety IR64 and the wild relative O. rufipogon". In: *Theor Appl Genet* 107, pp. 1433–1441.

Finn, Robert D et al. (2013). "Pfam: the protein families database". In: *Nucleic acids research* 42.D1, pp. D222–D230.

FOSS (1999). *Celery: Distributed Task Queue*. URL: http://www.celeryproject.org/ (visited on 10/02/2018).

Gao, Ting et al. (2010). "Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2". In: *Journal of ethnopharmacology* 130.1, pp. 116–121.

Gutenkunst, Ryan N et al. (2009). "Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data". In: *PLoS genetics* 5.10, e1000695.

Hermisson, Joachim and Pleuni S Pennings (2005). "Soft sweeps: molecular population genetics of adaptation from standing genetic variation". In: *Genetics* 169.4, pp. 2335–2352. ISSN: 0016-6731. DOI: 10.1534/genetics.104.036947. URL: http://www.genetics.org/cgi/content/abstract/169/4/2335http://www.ncbi.nlm.nih.gov/pubmed/15716498.

Hernandez, Ryan D, Scott H Williamson, and Carlos D Bustamante (2007). "Context dependence, ancestral misidentification, and spurious signatures of natural selection". In: *Molecular biology and evolution* 24.8, pp. 1792–1800.

Higdon, Jane V et al. (2007). "Cruciferous vegetables and human cancer risk: epidemiologic evidence and mechanistic basis". In: *Pharmacological Research* 55.3, pp. 224–236.

Hudson, Richard R (2002). "Generating samples under a Wright–Fisher neutral model of genetic variation". In: *Bioinformatics* 18.2, pp. 337–338.

Huerta-Sanchez, Emilia and Rick Durrett (2007). "Wagner's canalization model." In: *Theoretical population biology* 71.2, pp. 121–30. ISSN: 0040-5809. DOI: 10.1016/j.tpb.2006.10.006. URL: http://www.ncbi.nlm.nih.gov/pubmed/17178139.

Hufford, Matthew B et al. (2013). "The genomic signature of crop-wild introgression in maize". In: *PLoS Genetics* 9.5, e1003477.

Jiao, Yuannian et al. (2011). "Ancestral polyploidy in seed plants and angiosperms". In: *Nature* 473.7345, p. 97.

Katoh, Kazutaka and Daron M. Standley (2013). "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability". In: *Molecular Biology and Evolution* 30.4, pp. 772–780. DOI: 10.1093/molbev/mst010. eprint: /oup/backfile/content_public/journal/mbe/30/4/10.1093/molbev/mst010/3/mst010.pdf. URL: http://dx.doi.org/10.1093/molbev/mst010.

Krishnan, Arun, Masaru Tomita, and Alessandro Giuliani (2008). "Evolution of gene regulatory networks: Robustness as an emergent property of evolution". In: *Physica A: Statistical Mechanics and its Applications* 387.8-9, pp. 2170–2186.

Lenski, Richard E, Jeffrey E Barrick, and Charles Ofria (2006). "Balancing robustness and evolvability". In: *PLoS biology* 4.12, e428.

Lewontin, Richard C (1970). "The units of selection". In: *Annual review of ecology and systematics* 1.1, pp. 1–18.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14, pp. 1754–1760.

Li, Wei-Tao et al. (2010a). "Genetic analysis and ecological association of Hina genes based on single nucleotide polymorphisms (SNPs) in wild barley, Hordeum spontaneum". In: *Hereditas* 147.1, pp. 18–26.

Li, Yanwei et al. (2010b). "COI and ITS2 sequences delimit species, reveal cryptic taxa and host specificity of fig-associated Sycophila (Hymenoptera, Eurytomidae)". In: *Molecular ecology resources* 10.1, pp. 31–40.

Liu, Shengyi et al. (2014). "The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes". In: *Nature communications* 5, p. 3930.

Luo, Kun et al. (2010). "Assessment of candidate plant DNA barcodes using the Rutaceae family". In: *Science China Life Sciences* 53.6, pp. 701–708.

Luo, Ruibang et al. (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler". In: *Gigascience* 1.1, p. 18.

Lysak, Martin A et al. (2005). "Chromosome triplication found across the tribe Brassiceae". In: *Genome research* 15.4, pp. 516–525.

Marth, Gabor T et al. (2004). "The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations". In: *Genetics* 166.1, pp. 351–372.

Maynard Smith, J and J Haigh (1974). "The hitch-hiking effect of a favourable gene." In: *Genetical research* 23.1, pp. 23–35. ISSN: 0016-6723. URL: http://www.ncbi.nlm.nih.gov/pubmed/4407212.

Nielsen, Rasmus et al. (2005). "Genomic scans for selective sweeps using SNP data." In: *Genome research* 15.11, pp. 1566–75. ISSN: 1088-9051. DOI: 10.1101/gr.4252305. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1310644{\&}tool=pmcentrez{\&}rendertype=abstract.

Nowak, M A et al. (1997). "Evolution of genetic redundancy". In: *Nature* 388.6638, pp. 167–171. ISSN: 0028-0836. DOI: 10.1038/40618. URL: http://www.ncbi.nlm.nih.gov/pubmed/9217155.

Ofria, Charles, Christoph Adami, and Travis C Collier (2003). "Selective pressures on genomes in molecular evolution". In: *Journal of Theoretical Biology* 222.4, pp. 477–483. ISSN: 0022-5193. URL: http://www.ncbi.nlm.nih.gov/pubmed/12781746.

Pang, Xiaohui et al. (2011). "Applying plant DNA barcodes for Rosaceae species identification". In: *Cladistics* 27.2, pp. 165–170.

Parkin, Isobel AP et al. (2014). "Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea". In: *Genome biology* 15.6, R77.

Pavlidis, P. et al. (2013). "SweeD: Likelihood-based detection of selective sweeps in thousands of genomes". In: *Molecular Biology and Evolution* 30.9. ISSN: 07374038. DOI: 10.1093/molbev/mst112.

Pavlidis, Pavlos, Dirk Metzler, and Wolfgang Stephan (2012). "Selective sweeps in multilocus models of quantitative traits". In: *Genetics* 192.1, pp. 225–239.

Pavlidis, Pavlos et al. (2012). "A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans". In: *Molecular biology and evolution* 29.10, pp. 3237–3248.

Pinho, Ricardo, Elhanan Borenstein, and Marcus W Feldman (2012). "Most networks in Wagner's model are cycling". In: *PloS one* 7.4, e34285.

Portik, Daniel M et al. (2017). "Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection". In: *Molecular ecology* 26.19, pp. 5245–5263.

Prasad, Pramod Kumar et al. (2009a). "Phylogenetic reconstruction using secondary structures and sequence motifs of ITS2 rDNA of Paragonimus westermani (Kerbert, 1878) Braun, 1899 (Digenea: Paragonimidae) and related species". In: *BMC genomics*. Vol. 10. 3. BioMed Central, S25.

— (2009b). "Use of sequence motifs as barcodes and secondary structures of internal transcribed spacer 2 (ITS2, rDNA) for identification of the Indian liver fluke, Fasciola (Trematoda: Fasciolidae)". In: *Bioinformation* 3.7, p. 314.

Prescott-Allen, Christine and 1942 Prescott-Allen Robert (1986). *The first resource : wild species in the North American economy*. English. "Published with support from the World Wildlife Fund and Philip Morris Incorporated." New Haven : Yale University Press. ISBN: 0300032285 (alk. paper).

Przeworski, Molly, Graham Coop, and Jeffrey D Wall (2005). "The signature of positive selection on standing genetic variation". In: *Evolution; International Journal of Organic Evolution* 59.11, pp. 2312–2323. ISSN: 0014-3820. URL: http://www.ncbi.nlm.nih.gov/pubmed/16396172.

Sansom, Roger and Robert N Brandon (2007). *Integrating evolution and development: From theory to practice*. MIT Press.

Sawler, Jason et al. (2013). "Genomics assisted ancestry deconvolution in grape". In: *PLoS One* 8.11, e80791.

Schultz, Jörg and Matthias Wolf (2009). "ITS2 sequence–structure analysis in phylogenetics: a how-to manual for molecular systematics". In: *Molecular Phylogenetics and Evolution* 52.2, pp. 520–523.

Schultz, Jörg et al. (2005). "A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota". In: *Rna* 11.4, pp. 361–364.

Schultz, Jörg et al. (2006). "The internal transcribed spacer 2 database—a web server for (not only) low level phylogenetic analyses". In: *Nucleic Acids Research* 34.suppl_2, W704–W707.

Siegal, Mark L and Aviv Bergman (2002). "Waddington's canalization revisited: developmental stability and evolution". In: *Proceedings of the National Academy of Sciences* 99.16, pp. 10528–10532.

Snogerup, Sven, Mats Gustafsson, and Roland Von Bothmer (1990). "Brassica sect. Brassica (Brassicaceae) I. Taxonomy and variation". In: *Willdenowia*, pp. 271–365.

Song, Jingyuan et al. (2012). "Extensive Pyrosequencing Reveals Frequent Intra-Genomic Variations of Internal Transcribed Spacer Regions of Nuclear Ribosomal DNA". In: *PLOS ONE* 7.8, pp. 1–12. DOI: 10.1371/journal.pone.0043971. URL: https://doi.org/10.1371/journal.pone.0043971.

Stamatakis, Alexandros (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9, pp. 1312–1313.

Stanke, Mario and Stephan Waack (2003). "Gene prediction with a hidden Markov model and a new intron submodel". In: *Bioinformatics* 19.suppl_2, pp. ii215–ii225.

Stephan, Wolfgang, Thomas H E Wiehe, and Marcus W Lenz (1992). "The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory". In: *Theoretical Population Biology* 41.2, pp. 237–254. ISSN: 0040-5809. DOI: 10.1016/0040-5809(92)90045-U. URL: http://www.sciencedirect.com/science/article/B6WXD-4F1Y9N0-3M/2/1245281bba0c6b542457fdd75c343edf.

Tanksley, Steven D and Susan R McCouch (1997). "Seed banks and molecular maps: unlocking genetic potential from the wild". In: *Science* 277.5329, pp. 1063–1066.

Wagner, Andreas (1996). "Does evolutionary plasticity evolve?" In: *Evolution* 50.3, pp. 1008–1023.

— (2008). "Neutralism and selectionism: a network-based reconciliation". In: *Nature Reviews. Genetics* 9.12, pp. 965–974. ISSN: 1471-0064. DOI: 10.1038/nrg2473. URL: http://www.ncbi.nlm.nih.gov/pubmed/18957969.

Wang, Xiaowu et al. (2011). "The genome of the mesopolyploid crop species Brassica rapa". In: *Nature genetics* 43.10, p. 1035.

Witek, Kamil et al. (2016). "Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing". In: *Nature biotechnology* 34.6, p. 656.

Yao, Hui et al. (2010a). "Use of ITS2 Region as the Universal DNA Barcode for Plants and Animals". In: *PLOS ONE* 5.10, pp. 1–9. DOI: 10.1371/journal.pone.0013102. URL: https://doi.org/10.1371/journal.pone.0013102.

Yao, Hui et al. (2010b). "Use of ITS2 region as the universal DNA barcode for plants and animals". In: *PloS one* 5.10, e13102.