Mathematical modelling and integration of complex biological data : analysis of the heterosis phenomenon in yeast

Marianyela Petrizzelli

NNT : 2019SACLS204

Thèse de doctorat

Mathematical modelling and integration of complex biological data: analysis of the heterosis phenomenon in yeast

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

Ecole doctorale n°577 Structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat: Sciences de la vie et de la santé

Thèse présentée et soutenue à Gif-sur-Yvette, le 8 Juillet 2019, par

Marianyela Petrizzelli

Composition du Jury :

Christophe Giraud Professeur à l'Université Paris-Sud (LMO, Orsay)	Président
Marie-Laure Martin-Magniette Directrice de recherche, INRA (MIA, Paris)	Rapportrice
Henrique Teotónio Maître de conférences, ENS (IBENS, Paris)	Rapporteur
Delphine Ropers Chargée de recherche, INRIA (IBIS, Grenoble)	Examinatrice
Philippe Marullo Chargé de recherche, BioLaffort (ISVV, Villenave d'Orson)	Examinateur
Christine Dillmann Professeure à l'Université Paris-Sud (GQE-Le Moulon Gif-sur-Yvette)	Directrice de thèse
Dominique de Vienne Professeur émérite à l'Université Paris-Sud (GQE-Le Moulon, Gif-sur-Yvette)	Directeur de thèse

"*In the physical sciences, mathematical theory and experimental investigation have always marched together. Mathematics has been less intrusive in the life science because they have been largely descriptive, lacking the invariance principles and fundamental constants of physics.*
Increasingly, in recent decades, however, mathematics has become pervasive in biology, taking many different forms: statistics in experimental design; pattern seeking in bio-informatics; models in evolution, ecology and epidemiology; and much else... "

R. M. May

Acknowledgements

First of all I would like to express my sincere thanks and appreciations to my Ph. D. supervisors, Prof. Christine Dillmann and Prof. Dominique de Vienne. Your immense knowledge, rigor and constant orientation towards new tracks have allowed me to achieve original results, to prepare my thesis work and to grow as a research scientist. I deeply thank you for your guidance and

I feel lucky to have met you.

You have transmitted me your passion and constant love for the life-science and you have supported me continuously throughout this journey. Christine, I deeply feel that you have offer me your sincere friendship and I hope to be as lively, enthusiastic, and energetic as you are always. Dominique, your sensibility in understanding me have been supportive. You have been patient and your eternal good mood has let me keep a smile on my face even in the hardship.

I will forever be thankful to my former master research advisor Olivier Martin. You have been present since the beginning of my research career, before working together and after continuously providing me your advice. You were the reason why I decided to go to pursue a career in research and you gave me the opportunity to meet Christine and Dominique. You were and remain for me a model as a scientist and a mentor.

I would like to express my sincere gratitude to Marie-Laure Martin-Magniette and to Henrique Teotónio for accepting to be rapporteurs, along with Delphine Ropers, Christophe Giraud and Philippe Marullo for have me had the honor of judging this work.

I would also like to thank my theses committee, Anne Goelzer, Warren Albertin, Sylvie Huet, Bruno Bost and Jean-Pierre Mazat for their helpful advices and suggestions. Moreover, I would like to thank Delphine Sicard and Telma da Silva for the experimental construction that they have developed and that I have used in this study, along with Thibault Nidelet, Arnaud Le Rouzic and Monique Bolotin for the material they have passed me.

Thanks to all the member of my team Biology and Adaptation, Systems in Evolution for the continuum exchanges and interesting scientific discussions. I would like to add a thanks to the ammnistrative staff of my laboratory *Génétique Quantitative et Evolution* of Le Moulon and to my doctoral school Structure et Dynamique des Systèmes Vivants. I would also like to thank all the members of my laboratory for the enjoyable moments we have shared, and in particular to the basketball team of Le Moulon. I would specially thank you, Natalia Martinez. I have met you at the beginning of Ph. D. thesis and you have become a real friend to me.

Aside, I would like to thanks the Kavli Institute of Theoretical Physics for awarding me with a Fellowship that let me spend a period in California during my Ph. D. thesis. This opportunity let me growth as a scientist, meeting peoples all around the world with whom I had interesting discussions.

A special thanks goes to my family. Words cannot express how grateful I am to my mother, father, sister and brother for your support. I would also like to thank my beloved boyfriend, Guillaume Panthou. I can't thank you enough for encouraging me throughout this experience. You have been a true and great supporter and you have unconditionally loved me during my good and bad times, being non-judgmental of me and instrumental in instilling confidence. You have faith in me and my intellect and I thank you for that.

Contents

List of Figures

List of Tables

This work is dedicated to my family. Their support, encouragement, and constant love have sustained me throughout my life

Preface

Mathematics has long played a dominant role in our understanding of physics, chemistry and other physical sciences. In biology it has first been confined to some particular disciplines such as population genetics, but for some decades, this situation is changing at a fast pace. Mathematical methods are increasingly used for model construction and to deepen our knowledge of living systems.

The growing interest for mathematical biology may be partly explained by the advent of innovative profiling technologies that have led to high-throughput production of different types of biological data at different spatial and temporal scales. In this context, conceptual developments are essential to organize data and extract relevant information to analyze the interactions between the components of the systems and understand their behavior.

The novel technological tools for quantitative biology ranges from DNA sequencing for genomics to high-throughput phenotyping for ecology. Since the first genomic sequencing of the bacteria *Haemophilus influenzae* (Fleischmann et al., 1995) and that of the human genome (Lander et al., 2001; Venter et al., 2001), whole genome sequencing is now commonplace. This constitutes a major milestone for understanding organism biology, as whole genome provides a catalogue of all genes and associated molecules that are required for creating a living being, and carries information on the functioning of the organism under different developmental stages and conditions.

Other techniques such as spectroscopy, electro-chemistry and crystallography have enabled researchers to monitor complex cellular processes by using absorption and emission spectral methods, flow cytometric analysis, etc. The data generated by the so-called *–omics* technologies, such as transcriptomics, proteomics and metabolomics, shift cell analysis toward its interpretation. Indeed, transcriptomics sheds light on which genes are active in a given cell at a given time, proteomics reveals which proteins are present in a cell and in what amount, and metabolomics gives access to the metabolic processes at work in a cell under different conditions. All these components do not work in isolation but are connected at various levels in networks of varying complexity (Fischer, 2008).

Accurate high-throughput phenotyping strategies have been developed to highlight the quantitative phenotypic variation across cells, organs and tissues, developmental stages, years, environments and species. This wealth of data challenges systems biologists, quantitative geneticists, medical researchers and breeders not only to understand the genetic bases of complex trait variation, but also to use that knowledge to efficiently prevent diseases or derive crop varieties.

In this context, there is a crucial need for model construction to analyze and interpret the biological systems under investigation. Mathematics, hand-in-hand with the development of new statistical and computational tools, plays a key role in unifying concepts that allow researchers to get new insights about the biology of living systems and the regulation of their underlying complex mechanisms. The models can also help researchers to design further experiments for addressing new biological questions.

Due to experimental constraints, structures and parameters of biological systems can often not be assessed directly. Instead, they have to be inferred from limited, noise-corrupted data. Chance

plays a role in the variability of the biological phenomena, through experimental noise, chance during the reproduction process, stochasticity in the sub-cellular reactions, etc. These sources of stochasticity are not independent from each other, structuring the datasets and leading to similarities. The genotypes are linked by common evolutionary history, the genes are not independent within the genome, the traits are, in a large extent, pleiotropically connected, etc. The integration of dependence structures in models has both methodological and algorithmic costs (obtaining estimators in a realistic computation time is challenging).

In addition, multi-scale approaches are necessary for modeling the biological systems, which intrinsically and irreducibly integrate processes of various natures at various levels, and can be described within various frameworks (Lesne, 2013). Biological scales include atomic, molecular, molecular complexes, sub-cellular, cellular, multi-cellular systems, tissue, organ, multi-organ systems, organism, population (Prokop and Michelson, 2012). To achieve a holistic understanding of biological systems a wide range of models have been proposed (Hasenauer et al., 2015). They are obtained by coupling models at different scales, and accordingly, the most naive approach is to perform parameter estimation and model selection at each scale. Within a single quantity, the relevant parameters can encapsulate the net result of various processes, such as a structural feature, an interaction or the effect of an evolutionary pressure at a higher level.

Such data-driven models can be enriched with available knowledge about the biological processes, by integrating both experimental data from other scales and biological knowledge from the literature. Consider for example the work of Renaud et al. (2006) in which they coupled knowledge about teeth morphology of fossil rodents with the well-known model of response to selection (Lande, 1979), in order to study the evolutionary pressures constraining their development. They showed that the patterns of intra-specific phenotypic variation were conserved over long evolutionary time-scales and that departures were caused by climate-related selective pressure.

Finally, biological systems are essentially characterized by an entanglement of bottom-up and top-down influences following from their evolutionary history. The overall behavior of a system cannot be intuitively understood in terms of the individual components or interactions, and the qualitative nature of their behavior can depend on quantitative differences in their structure. Moreover, models must be specific to the investigated issue. They are designed to focus on certain aspects of the object of study, the other aspects being not considered. For instance, the familiar ball-and-stick model of chemical structure focuses on a molecule's chemical bonds. So it does not capture the resulting polarity in the molecule's atoms. Thus the models should ignore degrees of freedom irrelevant to the issue under study and should focus on the characteristic scales. Similarly, a multi-scale model should not intend to keep track of all details at all scales but only of the relevant features, whatever their scales, essential to address a particular biological question.

In this context, the focus of my Ph. D. work is to address the general question of the genotype-phenotype relationship, with particular attention to the study of hybrid vigor (or heterosis), relying on a big dataset obtained during a previous ANR project *HeterosYeast: exploitation of the heterosis phenomenon for wine yeast improvement*. In this project, a set of heterogeneous data, corresponding to different levels of cellular organization (quantitative proteomics, fermentation and life-history traits), was collected on a diallel-cross design constructed by pairwise crossing a series of strains belonging to two yeast species, *Saccharomyces cerevisiae* and *S. uvarum*, under two different growing conditions.

The approach involved a combination of mathematical, statistical and computational methods merged with biological knowledge of the system under study. Multi-scale and model testing

approaches have been employed for the prediction and understanding of the variation of the integrated phenotypes from protein abundance data and metabolic (flux) traits.

This work is organized as follows:

Chapter 1 provides an overview of the genetic bases of phenotypic variation from a quantitative and a population genetics perspective. The chapter is build gradually, from the concept of phenotypic variation to its driving evolutionary forces. I finish with the description of particular experimental designs and of statistical methods for the inference of genetic components.

Chapter 2 covers the experimental material and the analyses already performed on the HerosYeast dataset. For the sake of completeness, a brief overview of the phylogeny and domestication of *S. cerevisiae* and *S. uvarum* is first presented. Secondly, the chapter provides an overview of the diversity in life-history traits and fermentation in yeast, highlighting the well studied trade-off between life-history traits. Finally it presents the HerosYeast dataset, the previously achieved results and the aims of my Ph. D. work.

Chapter 3 presents the first part of my thesis work consisting in the identification of the genetic and molecular bases of phenotypic variation through the analysis of the diallel data. Among others the most striking finding has been the decoupling of the variances of heterosis and inbreeding effects (published in *Genetics*, Petrizzelli et al. (2019)).

Chapter 4 provides an introduction to constraint-based modeling (CBM) and to the methods I have used to investigate the molecular bases of phenotypic variation in the HerosYeast dataset. The chapter is organized as follows: first I present the formulation of CBM modeling in a mathematical framework, secondly I review methods for integration of proteomic data into CBM models, then I compare classical methods used to infer metabolic models with the one I use in chapter 5.

Chapter 5 presents the second part of my thesis work consisting in finding predictors of fermentation and life-history traits through the inference of metabolic fluxes. Statistical approaches have allowed to integrate the three different levels of cellular organization to gain information on the metabolic and molecular predictors of the integrated traits. This work, *Data integration uncovers the metabolic bases of phenotypic variation in yeast*, will soon be submitted to *Molecular Systems Biology*.

I conclude with **Chapter 6** were I draw my final conclusions and propose some future prospects.

Appendix A provides the supplementary material of the published article "Decoupling the Variances of Heterosis and Inbreeding Effects Is Evidenced in Yeast's Life-History and Proteomic Traits" (Chapter 3).

Appendix B provides the supplementary material of the article "Data integration uncovers the metabolic bases of phenotypic variation in yeast" (Chapter 5).

Appendix C, *Probabilities of multilocus genotypes in SIB recombinant inbred lines*, covers an additional research work on which I have worked during my master thesis and that is going to be submitted soon to *Frontiers in genetics*. It tackles the biological question of multi-locus frequencies in sibling (SIB) recombinant inbreed lines (RILs) by means of applied mathematical tools. This additional chapter provides the general formulation of the problem and the detailed description of the model for the computation of multi-locus probabilities for any number of loci in SIB-RILS.

Bibliography

Fischer, H. P. (2008). Mathematical Modeling of Complex Biological Systems, *Alcohol Research & Health* **31**(1): 49–59.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. and Al, E. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd, *Science* **269**(5223): 496–512.
URL: *http://science.sciencemag.org/content/269/5223/496*

Giraud, C. (2014). *Introduction to high-dimensional statistics*, Monographs on statistics and applied probability, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Boca Raton.

Hasenauer, J., Jagiella, N., Hross, S. and Theis, F. J. (2015). Data-driven modelling of biological multi-scale processes, *arXiv:1506.06392 [q-bio]* . arXiv: 1506.06392.
URL: *http://arxiv.org/abs/1506.06392*

Lande, R. (1979). Quantitative Genetic Analysis of Multivariate Evolution, Applied to Brain: Body Size Allometry, *Evolution* **33**(1): 402–416.
URL: *http://www.jstor.org/stable/2407630*

Lander, E. S., Linton, L. M., Birren, B., International Human Genome Sequencing Consortium et al. (2001). Initial sequencing and analysis of the human genome, *Nature* **409**(6822): 860–921.

Lesne, A. (2013). Multiscale Analysis of Biological Systems, *Acta Biotheoretica* **61**(1): 3–19.
URL: *http://link.springer.com/10.1007/s10441-013-9170-z*

Petrizzelli, M., Vienne, D. d. and Dillmann, C. (2019). Decoupling the Variances of Heterosis and Inbreeding Effects Is Evidenced in Yeast's Life-History and Proteomic Traits, *Genetics* **211**(2): 741–756.
URL: *https://www.genetics.org/content/211/2/741*

Prokop, A. and Michelson, S. (2012). *Systems Biology in Biotech & Pharma: A Changing Paradigm*, Springer Science & Business Media. Google-Books-ID: zSAjcvhwiNMC.

Renaud, S., Auffray, J.-C. and Michaux, J. (2006). Conversed phenotypic variation patterns, evolution along lines of least resistance, and departure due to selection in fossil rodents, *Evolution* **60**(8): 1701–1717.
URL: *http://doi.wiley.com/10.1111/j.0014-3820.2006.tb00514.x*

Venter, J. C., Adams, M. D., Myers et al. (2001). The sequence of the human genome, *Science (New York, N.Y.)* **291**(5507): 1304–1351.

Chapter 1

Chapter 1

The genetic bases of phenotypic variation

1.1 Phenotypic diversity

Phenotypic diversity, *i.e.* the fact that different individuals of a given species exhibit distinct phenotypes, is very common in natural populations. Understanding how phenotypic variation emerges and how it is maintained is of fundamental significance in the study of evolution and in its implications in plant/animal breeding and conservational biology (Andersson, 2001; Forsman, 2014). In this context quantitative genetics plays a key role for understanding the main factors affecting quantitative traits.

Fisher (1919) was the first to propose a mathematical formalism to tackle this question. Uniting Mendelian and quantitative genetics, he assumed that trait value is influenced by a large number of Mendelian genes and by a random environmental variation (Fisher, 1919). In particular, assuming that the overall population was panmictic, he proposed a probabilistic model for the decomposition of trait value taking into account the transmission mode of genetic information from one generation to another. At a reference generation $g = 0$, he parametrized the phenotypic value, P_i, of a trait observed for an individual, i, with the additive (A_i) contribution of a large number of genetic loci, allowing for dominance (D_i) within each locus and epistasis (I_i) between the loci. He further considered Mendelian segregation, *i.e.* the meiotic effect (W_i) resulting from the random choice of a gene in a locus out of two during meiosis, and an environmental (ϵ_i) effect:

$$g = 0, \quad P_i = A_i + D_i + I_i + W_i + \epsilon_i \tag{1.1}$$

The genetic and non-genetic effects were modeled as continuous random variables. Therefore, he could express offspring phenotypic value, between two randomly chosen individuals, i and j, from the reference generation as

$$g = 1, \quad P_o = \frac{1}{2}A_i + \frac{1}{2}A_j + D_o + I_o + W_o + \epsilon_o \tag{1.2}$$

Indeed, each offspring inherits one gamete (half of the genetic information) from its parents ($\frac{1}{2}A_i + \frac{1}{2}A_j$). Dominance effects are not predictable, in a panmictic population, since it is not possible to infer the allele received from one parent knowing the one inherited from the other. Similarly, epistatic and meiotic effects are specific to each individual, and in absence of environmental correlations, the offspring-parent environmental effects are independent.

Quantitative traits and Quantitative genetics

Quantitative traits, also known as complex or polygenic traits, usually show a continuous range of variation as they are influenced by both environmental and genetic factors.

Quantitative genetics is the study of the inheritance of quantitative traits, such as height or biomass, as opposed to discretely identifiable phenotypes, such as eye-color.

This simple model allows explaining the phenotypic resemblance between relatives, since they share common genes inherited from their kin, and the diversity observed on the whole population as a direct consequence of the genetic variation and of environmental factors, which provides the basis for evolution.

1.1.1 Components of phenotypic variation

Fisher (1919) proposed to characterize quantitative traits by their frequency distribution in a population. His model decomposed the mean phenotypic value of a trait in a population into a sum of random variables. The relative importance of the genetic and non-genetic effects can therefore be accessed by the relative ratio of the variances associated to each component and the phenotypic variance observed in the population. The model can be rearranged by grouping the additive and non-additive genetic effects, and the genetic and non-genetic effects. In a given environnement, we have:

$$P = A + NA + \epsilon = G + \epsilon \tag{1.3}$$

where $NA = D + I + W$ is the mean value of non-additive genetic effects, $G = A + NA$ is the the mean value of genetic effects and ϵ is the micro-environmental effect experienced by each individual, due to measurement errors, local effects and/or epigenetic factors. Therefore, the phenotypic variance can be expressed as

$$\mathrm{Var}(P) = \mathrm{Var}(G) + \mathrm{Var}(\epsilon) + 2\mathrm{Cov}(G, \epsilon) \tag{1.4}$$

It is immediate to remark that in a population composed only of individuals with the same genotype (*e.g.* clones, $F1$ offspring between two pure lines) phenotypic variation is still possible.

Without loss of generality, we can assume that the expected value of ϵ is zero for all genotypes, thus eq. 1.4 reads

$$\mathrm{Var}(P) = \mathrm{Var}(G) + \mathrm{Var}(\epsilon) \tag{1.5}$$

The component of phenotypic variation explained by genetic effects is the only component carrying the genetic information inherited from one generation to the other. The relative ratio between the genetic and phenotypic variance is called *broad-sense heritability*:

$$H^2 = \frac{\mathrm{Var}(G)}{\mathrm{Var}(P)} \tag{1.6}$$

Since genetic effects are independent by definition, $\mathrm{Var}(G) = \mathrm{Var}(A) + \mathrm{Var}(NA)$. Additive genetic effects are the only effects carrying transmissible information on a phenotypic trait. Therefore, the portion of the total phenotypic variance of a quantitative trait that is transmissible from generation to generation is

$$h^2 = \frac{\mathrm{Var}(A)}{\mathrm{Var}(P)} \tag{1.7}$$

generally referred to as *narrow sense heritability*.

1.1.2 Genotype by environment interactions

The ability of one genotype to express multiple phenotypes as a response to different environments is defined as phenotypic plasticity (fig. 1.1). In the general case, where the population is composed of genetically different individuals, genotypes may be more or less sensitive to the so-called macro-environmental effects, noted E. Macro-environmental effects are those that are common to a given location at a given time to all genotypes. The variance of E may vary with the genotype (Lerner, 1954; Crow, 1960), so the model becomes:

$$P = G + E + G \times E + \epsilon \tag{1.8}$$

where $G \times E$ is the genotype $\times$ environment interaction effect.

Phenotypic plasticity can therefore increase phenotypic variation in populations under divergent selection or create convergence of phenotypes within genetically diverse populations exposed to the

same selective pressure. In this context, the extent to which phenotypic plasticity is a heritable character and acts upon adaptive evolution is an open issue (Chevin and Lande, 2015).

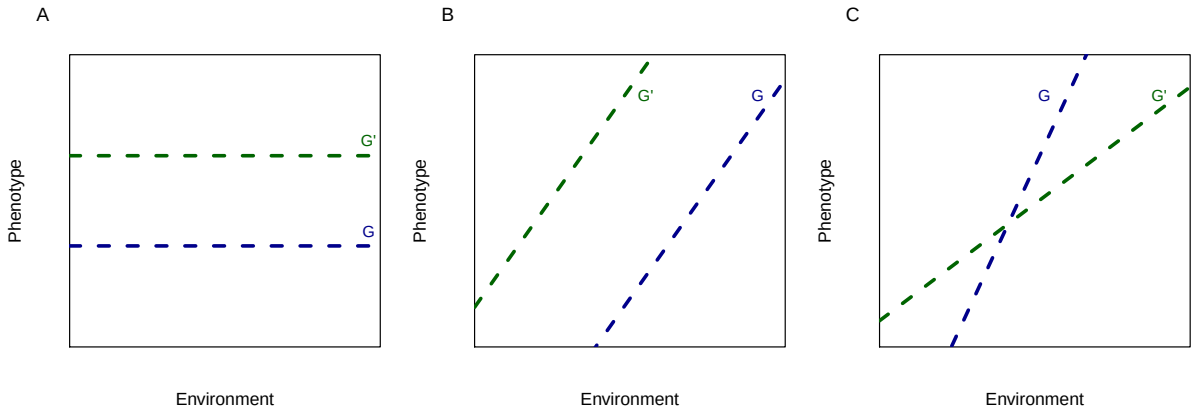


Figure 1.1: Phenotypic plasticity and genotype × environment interactions. Each dashed line represents a different genotype, G and G'. A, the lines are horizontal: the phenotypic values are not influenced by environmental changes (E is null). B, the lines are slanted and parallel: environmental variation produces the same phenotypic variation on the two genotypes ($G \times E$ is null, but not E). C, the lines intersect: the environment influences phenotypic variation in a genotype-dependent manner (E and $G \times E$ are not null).

1.1.3 Parent-offspring regression and the breeder's equation

Narrow sense heritability, h^2, is of particular interest mostly because it represents a quantitative measure of the quality of prediction of offspring's phenotypes from parental phenotypes, of resemblance between relatives and of the rate of short-term response to natural or artificial selection from standing variation, without knowing the details of the underlying genes.

Consider for instance the parent-offspring regression: $P_o = \mu + \beta P_p$, where μ is the mean phenotypic value of the offspring and β the regression slope:

$$\beta = \frac{\mathrm{Cov}(P_p, P_o)}{\mathrm{Var}(P_p)} \tag{1.9}$$

It is possible to obtain an estimate of the covariance between the parental and offspring phenotypic value using Fisher's model. Assuming a panmictic population of large effective size, with no selection, no genotype-environment interaction and independent environmental effects, and keeping in mind that genetic effects are independent by definition, the covariance between parents and offspring is:

$$\mathrm{Cov}(P_p, P_o) = \mathrm{Cov}(A_p + D_p + I_p + W_p + E_p, \frac{1}{2}A_p + \frac{1}{2}A_m + D_o + I_o + W_o + E_o) = \frac{1}{2}\mathrm{Var}(A) \tag{1.10}$$

where A_m is the additive effect of the second parent. Therefore

$$\beta = \frac{\mathrm{Var}(A)}{2\mathrm{Var}(P)} = \frac{1}{2}h^2 \tag{1.11}$$

Parent-offspring regression can be used to describe the phenotypic value of offspring from one generation to the other. Let X_t denote the phenotypic value of an individual at generation t. In a panmictic population with no overlapping generations, the phenotypic value of an offspring at

generation $t+1$, X_{t+1}, can be expressed as:

$$X_{t+1} = \mu_t + h^2(\frac{X_t^m + X_t^p}{2} - \mu_t) + \epsilon \tag{1.12}$$

The mean phenotypic value of offspring, conditionally to the value of its parents, is therefore

$$\mathrm{E}(X_{t+1}|X_t^m, X_t^p) = \mu_t + h^2(\frac{X_t^m + X_t^p}{2} - \mu_t) \tag{1.13}$$

Integrating over the phenotypic values of parents contributing to the next generation allows to estimate the mean phenotypic value at the next generation. When all parents do not contribute to the next generation, *e.g.* due to selection, we obtain:

$$\mu_{t+1} = \mu_t + h^2(\mu_{s,t} - \mu_t) \quad \Leftrightarrow \quad \mu_{t+1} - \mu_t = h^2(\mu_{s,t} - \mu_t) \tag{1.14}$$

where $\mu_{s,t}$ is the mean phenotypic value of parents contributing to the next generation (fig. 1.2).

Eq. 1.14 is generally referred to as *Breeder's equation*. In its most common formulation, eq. 1.14 writes

$$R = h^2 S \tag{1.15}$$

where S is the selection differential, the average phenotypic value of selected parents expressed as a deviation from the mean phenotypic value in the population, and R the response to selection, the average expected phenotypic value of offspring at the next generation expressed as a deviation from the previous generation, fig. 1.2.

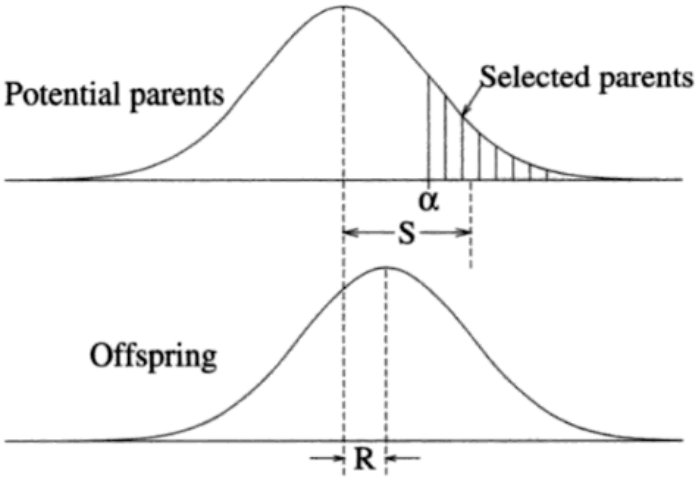


Figure 1.2: Response to truncation selection (Gillespie, 2004). Above: Phenotypic distribution of the selected trait in the parent population; α is the selection threshold and S the selection differential. Below: Phenotypic distribution of offspring. R is the difference of mean phenotype from one generation to the other.

Breeder's equation is an accurate description of the response to selection in a single generation. It is widely used in evolutionary biology to study the adaptation of natural populations. Its elegance resides in the fact that the complexity of multi-locus inheritance are aggregated into h^2. However, it is not necessarily an accurate predictor of the progress of selection over several successive generations because each generation of selection changes h^2 in ways that are impossible

to predict (Gillespie, 2004). Indeed, heritability changes with any modification in either additive, non-additive and/or environmental variances. When there is very little variation of additive effects with respect to the total phenotypic variance, $h^2 \sim 0$, and all phenotypic variation is attributed to chance. There can be extensive selection and yet no evolution.

1.1.4 Selection gradients

It is the amount of additive variance that determines the rate of evolutionary change. Directly comparing variances on multiple traits is difficult because they are not dimensionless and therefore vary with the scale of the trait or organism being measured. Yet, a different formulation of equation 1.15, and its multivariate version, have been proposed by R. Lande (Lande, 1976, 1979), coupling mean fitness value in a population to the mean phenotypic value of a trait.

Let $w_t(x)$ and $f_t(x)$ be the fitness and the probability density function, respectively, associated to phenotype x at generation t. We can assume that $w_t(x)$ is a continuous function, integrable on the domain of variation of x, $\mathcal{D}_x$. In particular, the population mean fitness is:

$$\bar{w}_t(x) = \int_{\mathcal{D}_x} w_t(x) f_t(x) dx \tag{1.16}$$

Changes in mean fitness due to selection can be related to changes in the mean value of fitness related traits. Assume a normal distribution for individual phenotypes

$$X_t \sim \mathcal{N}(\mu_t, \mathrm{Var}_t(P)) \tag{1.17}$$

Therefore, since $w_t(x)$ does not depends on the mean phenotypic value μ_t,

$$\frac{d\bar{w}_t}{d\mu_t} = \int w_t(x) \frac{df_t(x)}{d\mu_t} dx \tag{1.18}$$

Given that

$$\frac{df_t(x_t)}{d\mu_t} = f_t(x_t) \frac{x_t - \mu_t}{\mathrm{Var}_t(P)} \tag{1.19}$$

and

$$\mu_{s,t} = \frac{\int x_t w(x) f_t(x) dx}{\int w(x) f_t(x) dx} \tag{1.20}$$

in a panmictic population for which there are not effects of sex on x, equation 1.18, after arrangements, can be written as

$$\mu_{s,t} - \mu_t = \mathrm{Var}_t(P) \frac{d\ln(\bar{w}_t)}{d\mu_t} \tag{1.21}$$

Substituting, eq. 1.21 in eq. 1.14, a novel formulation of the response to selection is obtained:

$$\mu_{t+1} - \mu_t = \mathrm{Var}(A) \frac{d\ln(\bar{w}_t)}{d\mu_t} \tag{1.22}$$

This equation shows that the response to selection depends on the additive genetic variance of the trait of interest and on the selection gradient $\frac{d\ln(\bar{w}_t)}{d\mu_t}$. **Changes of fitness-related mean trait value at each generation are a result of selection driving fitness value towards a local maximum.**

Extension to the multivariate case in which multiple are the traits correlated to fitness is straight-forward, just by assuming normality for these traits. Letting

$$\Delta\overrightarrow{\mu} = \begin{pmatrix} \mu^1_{t+1} - \mu^1_t \\ \mu^2_{t+1} - \mu^2_t \\ \dots \\ \mu^n_{t+1} - \mu^n_t \end{pmatrix} \quad G = \begin{pmatrix} \mathrm{Var}(A^1) & \mathrm{Cov}(A^1, A^2) & \dots & \mathrm{Cov}(A^1, A^n) \\ \mathrm{Cov}(A^1, A^2) & \mathrm{Var}(A^2) & \dots & \mathrm{Cov}(A^2, A^n) \\ \dots & \dots & \dots & \dots \\ \mathrm{Cov}(A^1, A^n) & \dots & \dots & \mathrm{Var}(A^n) \end{pmatrix} \quad \overrightarrow{\beta} = \begin{pmatrix} \frac{d\ln(\bar{w}_t)}{d\mu^1_t} \\ \frac{d\ln(\bar{w}_t)}{d\mu^2_t} \\ \dots \\ \frac{d\ln(\bar{w}_t)}{d\mu^n_t} \end{pmatrix} \tag{1.23}$$

The equation can be written as

$$\Delta\overrightarrow{\mu} = G\overrightarrow{\beta} \tag{1.24}$$

where $\Delta\overrightarrow{\mu}$ is the vector of responses of multiple traits, G the variance-covariance matrix of additive genetic effects and $\overrightarrow{\beta}$ the vector of selection gradients.

Thus, selection on one trait can result in selection on another trait due to correlations between them. Similarly, if there is a correlation between additive genetic effects associated to different traits, a correlation between selection responses can be observed. The β coefficients determine the adaptive landscape, while the G matrix determines the direction of the phenotypic evolution following the axes of greater genetic variation.

? Question

The response to selection, R, depends on the additive genetic variance, $\mathrm{Var}(A)$. But how do we understand $\mathrm{Var}(A)$ in terms of allele frequencies and their additive effects on the phenotype?

1.1.5 One locus case

The relative portion of genetic variation explained by additive effects depend on the amount of genetic interactions, on population mating system and on alleles frequencies in the population. To show this last point, consider a panmictic population (in Hardy-Weinberg equilibrium) of genotypes with a single biallelic diploid locus. Alleles A_1 and A_2 are assumed to have frequency p and q, respectively, and genotypes A_1A_1, A_1A_2 and A_2A_2 to take the genotypic values a, d, $-a$, *i.e.* $d = a$ corresponds to complete dominance of allele A_1 over allele A_2, and $d = 0$ to additivity (tab. 1.1).

Genotype	A_1A_1	A_1A_2	A_2A_2
Value	$G_{A_1A_1} = a$	$G_{A_1A_2} = d$	$G_{A_2A_2} = -a$
Frequency	p^2	$2pq$	q^2

Table 1.1: Example. Genotypic values and genotypic frequencies for a single biallelic diploid locus.

The mean genotypic value of the population is thus

$$\mu = a(p^2 - q^2) + 2pqd = (p - q)a + 2pqd \tag{1.25}$$

The allele substitution effect, *i.e.* the change in mean genotype value when an allele A_2 is substituted by allele A_1, is

$$\alpha = a + d(q - p) \tag{1.26}$$

The additive effects for allele A_1 and A_2 are

$$\alpha_{A_1} = q\,(a + (q - p)d) \tag{1.27}$$

$$\alpha_{A_2} = -p\,(a + (q - p)d) \tag{1.28}$$

and the dominance effects

$$\delta_{A_1A_1} = -2q^2d \tag{1.29}$$

$$\delta_{A_1A_2} = 2pqd \tag{1.30}$$

$$\delta_{A_2A_2} = -2p^2d \tag{1.31}$$

The calculation of the additive and dominance components of genetic variance are straightforward:

$$\mathrm{Var}(A) = 2pq\,(a + (q - p)d)^2 \tag{1.32}$$

$$\mathrm{Var}(D) = 4p^2q^2d^2 \tag{1.33}$$

Additive and dominance genetic variances are expressed as a function of allele frequencies and genotypic values of the individuals in the population. If no dominance effect is present, $d = 0$, $\mathrm{Var}(D) = 0$ and $\mathrm{Var}(A) = 2pqa^2$.

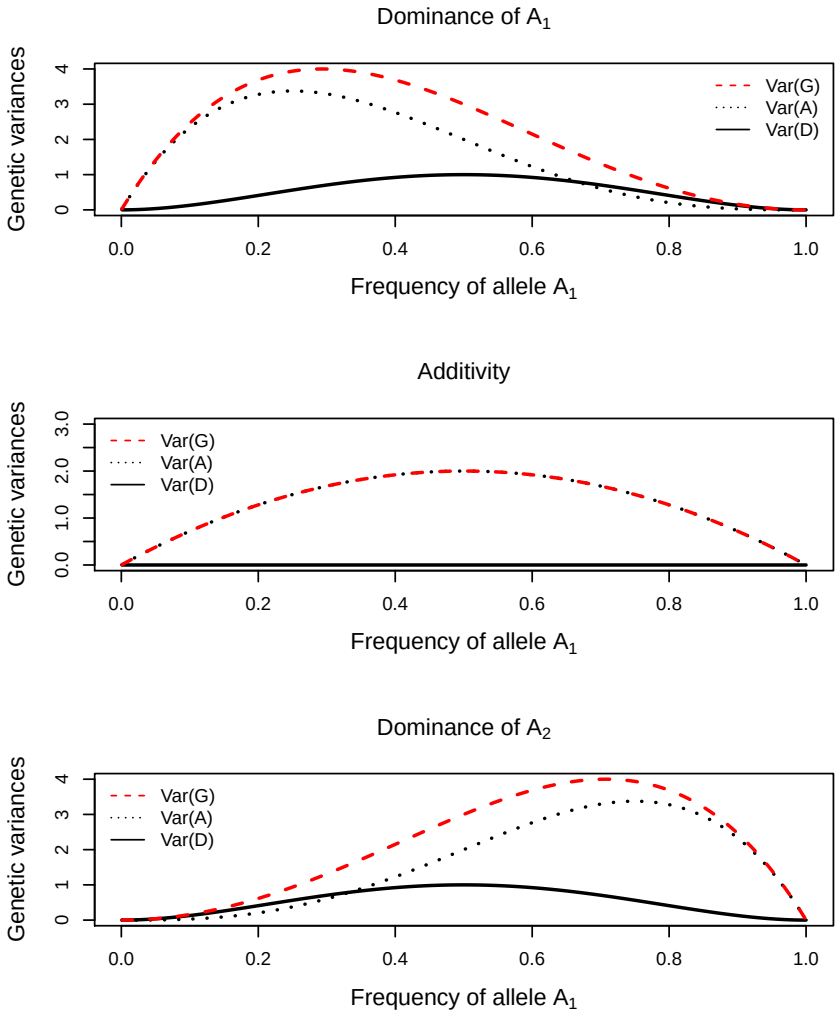


Figure 1.3: Genetic variance components in the biallelic case. Additive and dominance variances are calculated as a function of the frequency of allele A_1 through eq. 1.32- 1.33 (black dotted and solid lines, respectively). Red line represent the total genetic variance. Top: Complete dominance of A_1 ($d = a$). Center: No dominance ($d = 0$). Bottom: Complete dominance of A_2 ($d = -a$).

It is important to realize that the dominance effect contributes to the additive genetic variance. Suppose for instance that $a = -a = 0$, then $\mathrm{Var}(A) = 2pqd^2(q-p)^2$. Furthermore, when there is complete dominance ($d = a$ or $d = -a$), the ratio between the additive and dominance variances depends only on allele frequencies: $\mathrm{Var}(A)/\mathrm{Var}(D) = 2q/p$.

In addition, additive genetic variance is higher for low frequencies of the dominant allele (fig. 1.3). When the frequency of the dominant allele goes to zero, the genetic variance vanishes, dropping faster than when it goes to 1. The additive variance is always the major component of the total genetic variation, and when no dominance effect is present it is maximum for intermediate frequencies ($p = q = 1/2$).

In this simple example, the dominance effect contributes to the additive genetic variance. Introducing additional genetic interaction effects will contribute similarly to the additive variance component. Therefore, **additive effects must be thought as the average genetic effect transmitted from generation to generation, rather than the mean additive value of alleles participating to the considered trait value.**

Finally, we have seen that the genetic variance depends both on allele frequencies and genotypic values. If allele frequencies and genotypic values are constant across generations, there will be no phenotypic evolution of the population. If there is no genetic variation, the phenotypic variation of the population can only be due to environmental and/or epigenetic factors. As long as these factors do not vary, phenotypic evolution is not possible, even under selection.

How phenotypic diversity evolves and is maintained depends on the underlying mechanism responsible for changes in allele frequencies while preserving genetic diversity. As pointed out above, for a species to evolve there must be heritable phenotypic variation on which selection can act.

? Question

How does selection act on a fraction of segregating alleles within a population? How does this not lead to an extremely high variance and an intolerably large number of genetic deaths?

1.2 Genetic Polymorphism

Genetic variation is the result of processes generating variability (mutation, migration, segregation) and of demographic processes (selection and genetic drift). Mutations are changes in allele sequences through deletion, insertion, or, more commonly, substitutions of single DNA base pairs. They furnish an almost infinite field of possible gene variations. Migration (or gene flow) is the movement of genes into or out of a population. The allele frequencies of both the population they leave and the population they enter will change in relation to the rate of migration. Genetic drift is a random change in allele frequencies that is specially noticeable in small populations, in populations experiencing a bottleneck (the population suddenly gets much smaller), or in case of founder effect (a few individuals leave their population and found a new population). Segregation is the apportionment of alleles among the genotypes of the progeny resulting from the meiosis-fertilization process.

Natural and artificial selection act on genotypes by changing their probability to participate to the next generation. Artificial selection is due to the action of plant/animal breeders in choosing the parents of the next generation. Natural selection is due to differential mortality or fertility in the population, *i.e.* to selection of individuals showing higher fitness. Their common feature is that the parents of the next generation are a selected subgroup of the whole population.

More specifically, fitness is defined as the average contribution to the gene pool of the next generation that is made by an average individual of specified genotype. Given that the fitness of a given genotype is manifested through its phenotype, which is affected by the environment it experiences during its development, its fitness can be different in different selective environments.

Natural selection acts on a population as long as genetic variation exists and this standing variation is associated with fitness-related traits (Darwin, 1859). The problem amounts to understanding relationship between genetic variation and fitness. For the sake of simplicity, assume a diploid population whose fitness depends on one biallelic locus (tab. 1.2).

Genotype	A_1A_1	A_1A_2	A_2A_2
Fitness	$w_{A_1A_1}$	$w_{A_1A_2}$	$w_{A_2A_2}$
Frequency	p^2	$2pq$	q^2

Table 1.2: Fitness and frequencies of genotypes in a mono-locus biallelic model.

At each generation, if selection acts, allele frequencies will change as (Hartl and Clark, 1997)

$$\Delta p = \frac{pq}{\bar{w}}[p(w_{A_1A_1} - w_{A_1A_2}) + q(w_{A_1A_2} - w_{A_2A_2})] = \frac{pq}{2\bar{w}}\frac{d\bar{w}}{dp} = \frac{pq}{2}\frac{d\ln(\bar{w})}{dp} \tag{1.34}$$

where $d\ln(\bar{w})/dp$ is the selection gradient, $pq/2$ reflects the additive genetic variation of the allele frequency p in the population and $\bar{w}$ denotes the average fitness

$$\bar{w} = w_{A_1A_1}p^2 + 2w_{A_1A_2}pq + w_{A_2A_2}q^2 \tag{1.35}$$

It is easy to see that natural selection will increase the frequency of allele A_1 as long as

$$p > \frac{w_{A_2A_2} - w_{A_1A_2}}{w_{A_1A_1} - 2w_{A_1A_2} + w_{A_2A_2}} \quad \text{if} \quad w_{A_1A_1} - 2w_{A_1A_2} + w_{A_2A_2} > 0 \tag{1.36}$$

$$p < \frac{w_{A_2A_2} - w_{A_1A_2}}{w_{A_1A_1} - 2w_{A_1A_2} + w_{A_2A_2}} \quad \text{if} \quad w_{A_1A_1} - 2w_{A_1A_2} + w_{A_2A_2} < 0 \tag{1.37}$$

Assuming heterozygote inferiority, ineq. 1.36 applies and natural selection will drive allele frequencies towards an extreme value (0 or 1), depending on the initial allele frequencies in the population, thus eliminating genetic variation (fig. 1.4-A). Assuming heterozygote superiority, ineq. 1.37 applies and natural selection will drive allele frequencies towards an intermediate value, preserving genetic diversity (fig. 1.4-B). Nevertheless, heterozygote superiority and/or inferiority are not well documented (Fiévet et al., 2018). In the dominance case, selection leads to the fixation of the advantageous alleles (fig. 1.4-C).

In general, it is assumed that, in a stable environment, the additive genetic variance of a panmictic population of finite size will decline over time, due to genetic drift. Under directional selection, additive genetic variance is assumed to decline due to both drift and fixation of favorables alleles, while the speed of fixation is modulated by generation of gametic disequilibrium (Bulmer, 1971).

How genetic variation is maintained depends on the interplay between these different mechanisms. Haldane and Jayakar (1963) have argued that, in natural populations, genetic polymorphism is the result of conflicting evolutionary pressures, identifying five main conflicts:

The conflict between selection and mutation, or mutation-selection balance. Since most mutations that affect fitness are deleterious, selection will balance the effects of mutations, and genetic polymorphisms may be maintained in the population. For instance, consider the mono-locus biallelic model in tab. 1.2, where the A_2 allele is produced by mutation of the A_1 allele, at a rate μ, with relative fitnesses $w_{A_1A_1} = 1$, $w_{A_1A_2} = 1 - hs$ and $w_{A_2A_2} = 1 - s$, s being the selection coefficient against the A_2A_2 genotype and h the degree of dominance of allele A_1. An equilibrium can be attained and allele frequencies will be $p = 1 - \sqrt{\frac{\mu}{s}}$ under complete dominance, and $p = 1 - \frac{\mu}{hs}$

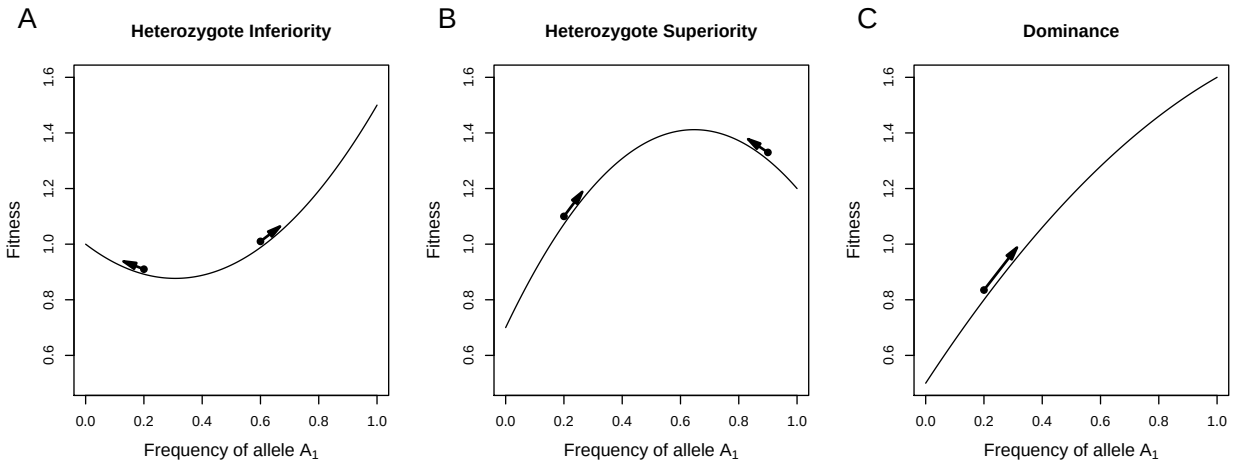


Figure 1.4: Examples of adaptive landscape in the one locus case. Frequencies of allele A_1 against mean fitness in the population. Arrows indicate the direction of selection, *i.e.* changes in allele A_1 frequency due to selection. Natural selection will drive allele frequencies towards: **A**, an extreme value, in case of heterozygote inferiority, depending on the initial allele frequencies in the population; **B**, an intermediate value, in case of heterozygote superiority; **C**, 1 for the strongest allele and 0 for the lowest, in case of dominance. Overall, natural selection drives allele frequencies towards the closest fitness local optimum.

under partial dominance assuming $hs \gg \mu$. Similarly, if we consider haploid organisms, a genetic polymorphism can be maintained in the population if $\mu < s$ (Haldane, 1937).

The conflict between selection and segregation. The most common example is when the heterozygote has a higher fitness than either of the two homozygotes (fig. 1.4-B). For instance, if $w_{A_1A_1} = 1 - k$, $w_{A_1A_2} = 1$ and $w_{A_2A_2} = 1 - s$, k and s being the selection coefficients against A_1A_1 and A_2A_2, respectively, an equilibrium can be reached when $p = 0$ or $p = 1$ or $p = \frac{s}{s+k}$. The first two equilibria are unstable, while the latter is stable and corresponds to the case in which the average fitness is maximized in the population.

The conflict between fitness and frequency. Selected polymorphisms can be maintained through negative frequency-dependent selection, *i.e.* the fitness of a genotype decreases as it becomes more frequent. As an example, we can consider that the fitness of a genotype decreases proportionally to its frequency at a constant c, thus, $w_{A_1A_1} = 1 - cp^2$, $w_{A_1A_2} = 1 - 2cpq$ and $w_{A_2A_2} = 1 - cq^2$. An equilibrium can be reached when $p = 0$, $p = 1$ or $p = 1/2$. As before, the first two equilibria are unstable, while the latter is stable and corresponds to the case in which both alleles are present in equal proportion. As an example, consider the conflict between sexes under optimal mating rate with costly male sexual harassment. In this case, polymorphism can emerge through negative frequency-dependent selection on fecundity (Iserbyt et al., 2013).

The conflict between selection and migration. The relative fitness of genotypes may vary according to different environments. If each genotype is favored in a different subset of environments, within subdivided populations, local adaptation would have tendency to fix different alleles in different geographic location, thus allowing the maintenance of genetic diversity between demes at the level of the whole population. Inter-deme migration or colonization is therefore the main mechanism to maintain genetic variation, importing new genetic material within a deme.

The conflict between selection in the diploid and the haploid phases or between the

two sexes. When the fitness associated with the diploid and the haploid phases differs (or similarly, the fitness between genotypes differs between sexes), the relative magnitude of the fitness of the two states can attain an equilibrium. Interestingly, with an appropriate choice of fitnesses, it is possible to have more than one stable polymorphism (Otto et al., 2015).

Genetic polymorphism
From a practical point of view, genetic polymorphism is the occurrence of different alleles at a locus within a population at a rate of at least 1%.

1.2.1 Genetic load

All mechanisms that generate selected polymorphism are necessarily accompanied by the apparition of genetic load, defined as the proportion by which the average fitness in the population is decreased in comparison with what it would be if the factor under consideration were absent (Crow, 2001), *i.e.*

$$L^{(f)} = 1 - \frac{\bar{w}}{w_{max}} \tag{1.38}$$

where f denotes the factor of interest and w_{max} the maximum fitness.

The reduction in mean fitness of a population that is caused by deleterious mutations is called mutation load; by the recreation of Hardy-Weinberg genotype frequencies in sexual organisms, segregation load (for instance, in random mating populations the homozygous state will be regenerated by segregation); by unfavorable alleles increasing in frequency due to drift in small populations, drift load; and by immigrants adapted to a different environment, migration load.

The apparition of genetic load is naturally associated to heterosis (or hybrid vigor). At a population level, heterosis is defined as the increase in mean fitness of offspring with respect to the parental population. Indeed, as long as genetic load exists, the population has not reached its maximum fitness and there is place for heterosis.

Haldane (1937) suggested that this loss of fitness is the price paid by a population for its capacity for further evolution. Indeed, the apparition of genetic load has important evolutionary consequences for instance on the fate of small populations, in the evolution of sex and in the evolution of mating systems. All mechanisms able to reduce the genetic load will be favored by natural selection.

1.2.2 Segregation load

Segregation load is the reduction on mean fitness of a population that is caused by the recreation of Hardy-Weinberg genotype frequencies in sexual organisms. Reconsidering the mono-locus biallelic case presented above (tab.1.2) with $w_{A_1A_1} = 1$, $w_{A_1A_2} = 1 - hs$ and $w_{A_2A_2} = 1 - s$. For $h < 0$, the common example of heterozygote superiority applies for any value of s, and **at the selection-segregation equilibrium the genetic load would be null** (fig. 1.5, blue dotted line). However, for $h > 0$ heterozygosity inferiority applies, and genetic load at equilibrium can reach high values. In general, genetic load is minimum at the selection-segregation equilibrium.

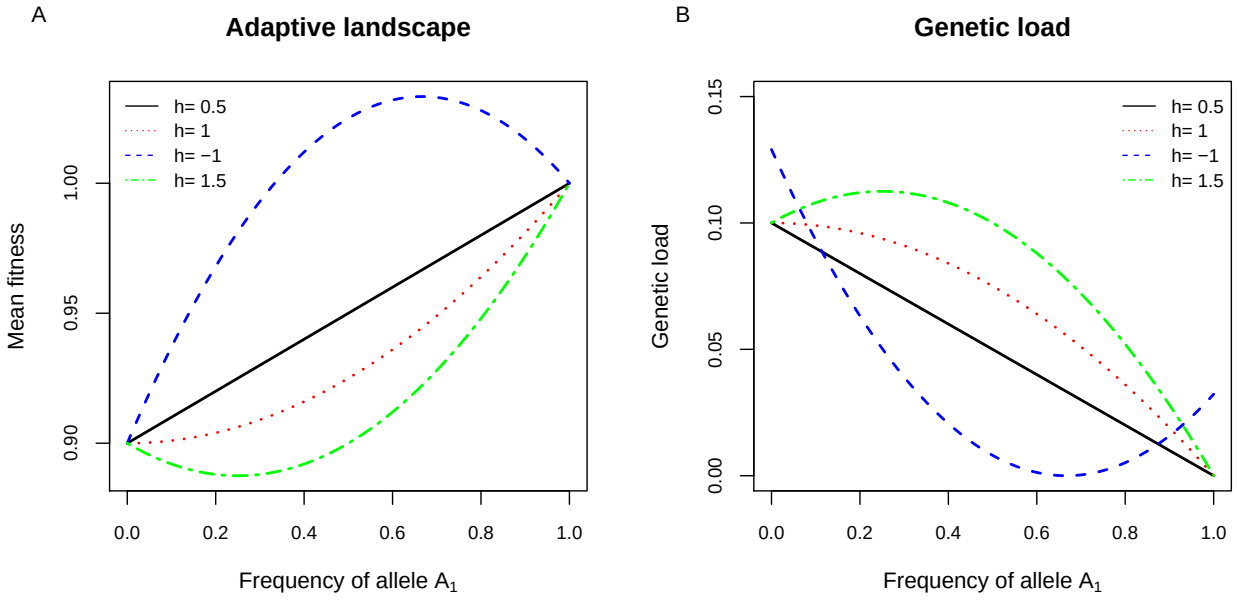


Figure 1.5: Mean fitness and genetic load. Frequencies of allele A_1 against: **A**, mean fitness in the population; **B**, genetic load. The selection coefficient s is set to 0.1, the degree of dominance of allele A_2 is let to vary: $h = 0.5$ no dominance (black), $h = 1$ dominance of the recessive allele (red), $h = -1$ heterozygote superiority (blue), $h = 1.5$ heterozygote inferiority (green). The figure shows that genetic load is minimum when allele frequencies reach their equilibrium value in the population.

1.2.3 Mutational load and drift

Mutational load is the reduction in mean fitness of a population that is caused by deleterious mutations. In the simple mono-locus biallelic case presented above, assuming that the maximum relative fitness of a genotype without mutations is equal to 1 ($\bar{w}_{\text{no mut}} = 1$), the average fitness of the population will be $\bar{w}_{\text{mut}} = 1 - 2pqsh - q^2 s$ under mutation-selection balance. Thus, the mutational load is

$$L^{(m)} = 2pqsh + q^2 s \tag{1.39}$$

Under partial dominance, assuming that selection is stronger than the mutation rate ($hs \gg \mu$), $p \sim 1$, $q^2 \sim 0$ and $L^{(m)} \simeq 2\mu$; under complete dominance, $L^{(m)} \simeq \mu$, *i.e.* selection removes two copies of mutation at once. Therefore, at a first-order approximation, the mutational load depends only on the mutation rate at the locus. This implicates that the harmful effect of an increase in the mutation rate is the same with respect to the case in which the produced mutations are mildly or severely deleterious. Their effect indeed counterbalance because a more detrimental mutation comes at lower rate equilibrium frequency.

A generalization to the multi-locus case can be made assuming no epistatic interaction for fitness between deleterious mutations (no genetic interaction between loci carrying mutations affecting fitness). The mean fitness in the population can be expressed through a multiplicative fitness function as the product of the mean fitness effects at each locus:

$$\bar{w} = \prod_{l=1}^{L} (1 - 2\mu) \simeq e^{-\sum_{l=1}^{L} 2\mu} = e^{-U} \tag{1.40}$$

where the product runs over the $L \gg 0$ loci, and $U = \sum_{l=1}^{L} 2\mu$ denotes the genome wide mutation rate of alleles affecting fitness. Note that assuming mutations follow a Poisson distribution of parameter U, this term correspond to the probability of no mutation. Therefore, the mutational

load is

$$L^{(m)} = 1 - e^{-U} \tag{1.41}$$

When the equilibrium between selection and mutation is reached, even in an excess of mutation over selection, a population will not degenerate. If however the population is subject to drift and reproduce asexually, selection, even strong, will not be able to keep the population at equilibrium (Muller, 1964).

To show this point, consider an initial finite haploid (or diploid without dominance) population reproducing asexually and released at the peak of the multiplicative fitness landscape, *i.e.* individuals do not carry any mutation and thus have maximum fitness, $\bar{w} = 1$ and $w_{max} = 1$. The population reproduces randomly, and at each generation it undergoes selection, mutation and drift. Under the action of mutation some individuals will soon acquire deleterious mutant alleles. Let's consider classes of individuals having the same number of mutations (under a multiplicative fitness landscape it does not matter the position of the mutant allele in the genome, but only its number). The fittest class, holding zero mutation, will participate to the next generation only if at least one individual *(i)* does not experience mutation and *(ii)* it is sampled by random drift. If it is not sampled, the zero mutation class disappears, and the fittest class becomes the class holding one mutation. Indeed, due to the unidirectional nature of mutations, fittest genotypes can never be restored, inducing a decline of the mean fitness in the population, and therefor an increase of the mutational load, in a ratchet-like manner (fig. 1.6).

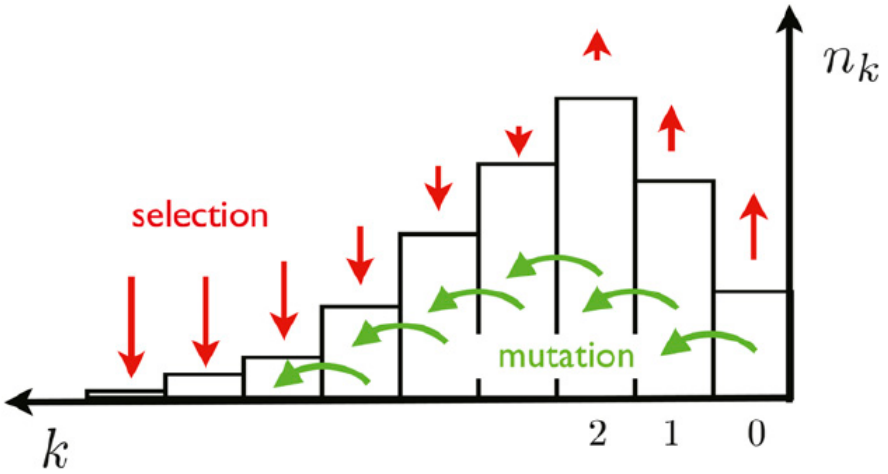


Figure 1.6: Muller's ratchet. k is the number of mutations and n_k the number of individuals with k mutations. The red arrows indicate positive (upwards) or negative (downwards) selection. Initially, accumulation of deleterious mutations will be accompanied by selection of the fittest class of individuals. Drift continuously removes individuals, and in the long term, the class with less mutations disappears (due to both drift and mutations). As deleterious mutations accumulates, selection will act on the opposite direction, until extinction.

A fundamental difference could be obtained with a sexual reproductive regime. Haag and Roze (2007), using single-locus models, have explored the combined effects of segregation, selection, and drift in finite populations of sexual and asexual individuals. For partly recessive deleterious alleles, they found that segregation affected changes in allele frequencies resulting in a greater mutation load in asexuals than in sexuals. This arises primarily because, in the absence of segregation, heterozygotes may reach high frequencies due to drift, which is not possible with segregation, as mating between heterozygotes constantly produces new homozygotes which are efficiently selected

against. Further, they proposed an extension of their model to the multi-locus case under a multiplicative fitness landscape, that could substantially reduce genetic load for sexuals. Indeed, genetic drift is accompanied by the apparition of random associations between loci (positive and negative). Positive associations are rapidly fixed by selection while negative are broken by recombination (Hill and Robertson, 1966), therefore generating a selective advantage for sexual reproduction.

The ratchet-like phenomenon is therefore more pronounced with asexual than sexual reproduction, as pointed out by Muller (1964). A process termed *Muller's Ratchet* describes the phenomenon of almost irreversible (other than exact reverse mutations) accumulation of deleterious mutations in asexual populations.

1.2.4 The evolution of sex

Stochastic effects occurring in any finite population tend to generate negative associations between loci (Hill and Robertson, 1966). Breaking these negative associations increases the variance in fitness among offspring and the efficiency of natural selection, that is the role of recombination which therefore increases the rate of adaptation, as is the case in Muller's Ratchet.

On the other hand, sexual reproduction costs in terms of energy required to find a mate, increased risk of predation and disease transmission, investment into males (the two-fold cost of sex) or time (sexually reproducing organisms tends to have fewer offspring and takes much longer to grow).

A first concern is about the modification of a reproductive system. In nature obligate sexuals persist, yet organisms able to alternate between reproductive modes exist, *e.g.* yeasts, lettuce-leaf aphids or rotifers. To investigate selection for sex in finite populations, numerous theoretical models have been proposed. Roze (2014) addressed this question by modeling sex rate as a quantitative trait on a finite population consisting of haploid individuals. The relative investment into sexual and asexual reproduction was assumed to depend on one locus and, at each generation, the probability of an individual to participate to the next generation depended on its fitness and on its role in the production of offspring (*i.e.* reproducing asexually or on being the female or male for sexual reproduction). Individual fitness was assumed to be multiplicative and depended on the number of accumulated mutations and on the selection coefficient against the deleterious mutations. This study showed that alleles increasing sex rate escape more easily from low-fitness genetic backgrounds than alleles coding for lower rates of sex. Furthermore, at mutation-selection balance, where selection is strong enough to outweigh a substantial cost of sex, interactions between selected loci had a stronger effect that the sum of individual effects of each locus. This means that selection on a sex-related allele resulted from its effect in pairwise associations with other loci. Overall deleterious mutations tend to favor small rates of sex in the presence of strong direct costs. However, population structure should enhance indirect selection due to stochastic effects and allow higher rates of sex to be maintained.

Vanhoenacker et al. (2018) proposed a model to account for epistatic interactions for the study of sex evolution of a haploid population under an isotropic model for stabilizing selection. The fitness of an individual depended on a variable number of phenotypic traits. Sex was modeled as a phenotypic trait, and trait values depended on the additive contribution of a large number of loci, and of a random environmental effect. No covariance between traits was assumed and epistasis was defined as a deviation from additivity of mutational effects on the (log) fitness genotype. They showed that positive rates of sex are maintained in the population at equilibrium. Selection of sex depended on the dimensionality of the pleiotropic fitness landscape and, for weak selection and not too low rates of sex, on negative linkage disequilibrium caused by epistasis. They further highlighted that selection gradients exist for sex, since sex breaks the associations between alleles at different loci generated by selection, increasing the genetic variance among offspring, and allowed for a better response to directional selection.

1.2.5 The evolution of mating systems

Another major point to account for is, in sexually reproducing populations, the appearance of inbreeding depression and the reduction in fitness due to inbreeding. Yet among sexual species, many reproduce with both selfing and out-crossing and others have developed mechanisms to avoid selfing, such as self-incompatibility, dioecy, heterostyly or dichogamy.

To investigate this issue, Lande and Schemske (1985) have proposed a multi-locus model for the study of the evolution of selfing rate. Inbreeding depression was allowed to change with the mean of selfing rate in a population incorporating recessive mutations and partially dominant lethal and sub-lethal alleles at many loci. Selfing rate was supposed to depend on one locus, while fitness depended on an infinite number of loci, with small effect. Letting $\bar{w}_0$ and $\bar{w}_1$ denote the mean fitnesses of out-crossed and selfed progeny in the population. Inbreeding depression, the reduction in mean fitness in the population caused by inbreeding, can be expressed as:

$$\delta = 1 - \frac{\bar{w}_1}{\bar{w}_0} \tag{1.42}$$

Assuming that all genotypes produce the same amount of pollen, and that any seeds which are not derived from out-crossing are self-fertilized, the expected fitness of genotypes with selfing rate r is

$$w = r\bar{w}_1 + \frac{1}{2}(1-r)\bar{w}_0 + \frac{1}{2}(1-\bar{r})\bar{w}_0 \tag{1.43}$$

where the first two terms are components of fitness from selfed and out-crossed seeds, and the last term is that from pollen fertilizing ovules of other plants ($\bar{r}$ denoting the mean rate of selfing in the population).
The condition for the evolution of the selfing rate, r, is therefore

$$\frac{dw}{dr} > 0 \quad \Leftrightarrow \quad \delta < \frac{1}{2} \tag{1.44}$$

i.e. there is selection for increased selfing if the inbreeding depression is less than 50%, *i.e.* the two-fold cost of sex. Under different hypotheses on the causality between mating systems different threshold values for inbreeding depression are found.

In addition, Lande and Schemske (1985) showed that at the selection-mutation balance, under complete dominance, the mean fitness value of progeny was equal to the mean mutational fitness for any value of the selfing rate. Under random mating, selfing rate is assumed to be small ($r \sim 0$) and the mean fitness of the out-crossed and of the selfed progeny is

$$\bar{w}_0 = \prod_i \bar{w}_0(i) = \prod_i (1-\mu), \tag{1.45}$$

$$\bar{w}_1 = \prod_i \bar{w}_1(i) = \prod_i (1 - \frac{\sqrt{\mu s}}{2}) \tag{1.46}$$

where i is the locus index, μ and s are the mutation rate and the selection coefficient at a locus, respectively, and where it is assumed that selection acts independently on each locus and so fitness effects are multiplicative across loci. Therefore, inbreeding depression is

$$\delta_0 = 1 - \frac{\prod_i (1 - \frac{\sqrt{\mu s}}{2})}{\prod_i (1-\mu)} \simeq 1 - e^{-\sum_i (1 - e^{-(\frac{\sqrt{\mu s}}{2} - \mu)})} \tag{1.47}$$

On the other hand, when the rate of selfing is appreciably high, $r \gg 4\sqrt{\frac{\mu}{s}}$,

$$\bar{w}_0 = 1 \tag{1.48}$$

$$\bar{w}_1 = \prod_i (1 - \frac{\mu}{r}) \tag{1.49}$$

$$\delta_r = 1 - \prod_i (1 - \frac{\mu}{r}) \simeq 1 - e^{-\sum_i \frac{\mu}{r}} \tag{1.50}$$

These formulas confirm that a small amount of selfing greatly reduces the equilibrium frequency of recessive lethals through purging of recessive lethals ($\delta_r < \delta_0$ already for $r = 1\%$) (fig. 1.7). Qualitatively, a similar result is obtained for partially dominant lethal mutations, where the inbreeding depression rapidly decreases as selfing rate increases. In comparison the equilibrium inbreeding depression in a random mating population greatly decreases, allowing for selection of selfing. In addition, when there is variation in the degree of dominance of the deleterious mutants, an excess of inbreeding depression can be produced.

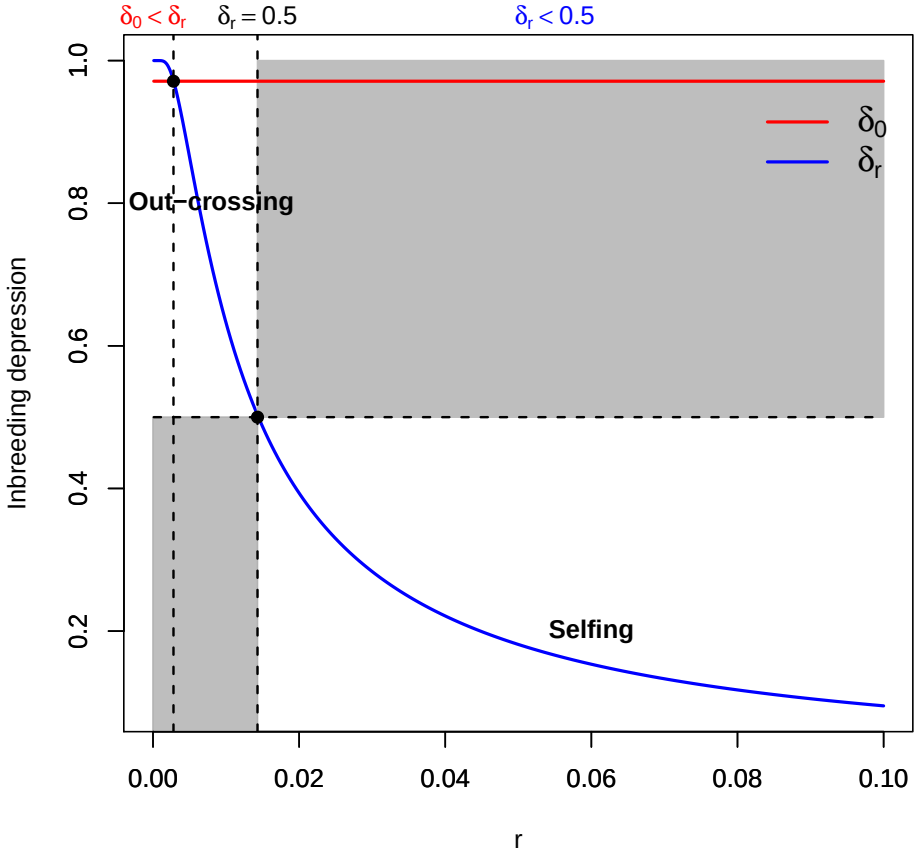


Figure 1.7: Inbreeding depression versus the rate of selfing. As in Lande and Schemske (1985), the mutation rate is set at $\mu = 2 \cdot 10^{-6}$, the number of loci at $n = 5000$ and mutations are assumed to be lethal ($s = 1$) and fully recessive. In populations allowed to self reproduce, inbreeding depression is a decreasing function of the selfing rate (blue dotted line, δ_r) and rapidly falls to zero (for $r = 10\%$, $\delta_r = 0.095$). Under random mating, inbreeding depression does not depend on the selfing rate (red line, δ_0). Out-crossing is selected for $\delta_r > 0.5$, while selfing for $\delta_r < 0.5$ (horizontal dotted line). Grey rectangles feature parameters values that cannot be encountered.

Overall, if the selfing rate is under polygenic control and its evolution proceeds by small steps, there is a bimodal distribution of the selfing rates towards their extreme values: either close to 0 for outcrossing or close to 1 for highly selfed populations. Nevertheless, under random mating, sporadic events are likely to drive out-crossing species to self-fertilization, rather than vice-versa.

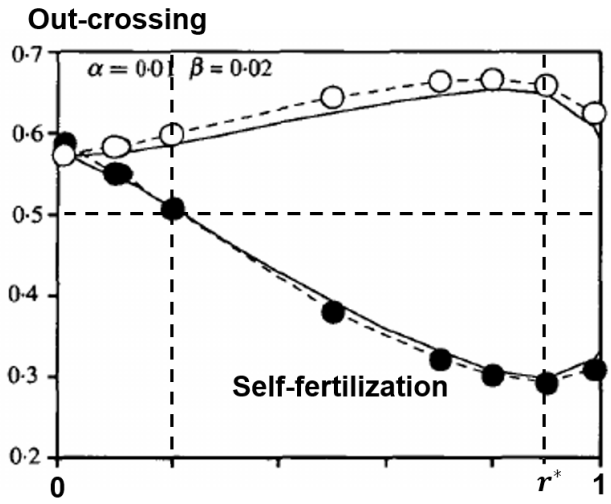


Figure 1.8: Relation between inbreeding depression (filled circles), mean fitness (open circles) and selfing rate (in abscissa) in equilibrium populations with synergistic epistasis. For low selfing rates, mean fitness increases and inbreeding depression decreases with the selfing rate. For high selfing rates, mean fitness may decrease. $U = 1$, $h = 0.2$, $\alpha = 0.01$, $\beta = 0.02$ (Charlesworth et al., 1991).

On the other hand, this model assumes a high mutation rate (instead of $\sim 10^{-8}$) and a multiplicative landscape for non interacting multi-locus effects. Interactions between loci may lead to higher order of inbreeding depression with comparable mean fitness levels.

Charlesworth et al. (1991) proposed a model of synergistic fitness interactions to explain the maintenance of high inbreeding depression and out-crossing under the mutational load model. The proposed model allowed for homozygote and heterozygote mutants, and mutations at multiple loci were supposed to lower the fitness value relative to the case of independence between loci. The fitness value of an individual was therefore modeled as:

$$w_n = \exp[-(\alpha n + \frac{\beta n^2}{2})] \tag{1.51}$$

where $n = hz + y$ is the effective number of mutations, expressed as the sum of the number of mutations in the heterozygous state, z, weighted by the dominance effect of heterozygous loci, h, and the number of mutations in the homozygous state, y; α is a measure of the strength of selection and β is a measure of the interaction between loci. They showed that the mean number of mutations per individual at equilibrium decreased with increased selfing, as for the multiplicative model, and with increased synergism. This induces a higher mean fitness under the synergistic model than under the multiplicative model. Synergism reduced the fall-off of inbreeding depression and increased genetic load with increased selfing. In addition, there can be evolutionarily stable states at values of selfing rate slightly below complete selfing (fig. 1.8).

Interestingly, figure 1.8 shows that with epistasis, the equilibrium inbreeding depression can be significant even in predominantly selfing populations. Recall that $\delta > 0$ means that the average fitness of outcross progenies is higher than the average fitness of selfed progenies. Because the level of heterozygosity of outcross progenies is expected to be higher than the one of selfed progenies, this corresponds to heterosis at the population level.

As stated by Charlesworth et al. (1991), all this together helps to explain the persistence of high heterosis in predominantly self-fertilizing populations, without the need for invoking a general heterozygote advantage, for which there is little evidence.

1.2.6 Population structure, inbreeding depression and heterosis

Both the spatial distribution of organisms and their mode of reproduction have important effects on the change in allele frequencies within populations. In the previous section, we have discussed on the direct advantages associated to selfing, and the evolution of the mating systems in terms of the cost of out-crossing and inbreeding-depression. Here, we discuss the effects of population structure under mutation-selection balance on inbreeding depression and heterosis.

Individuals from the same species are generally found in different geographical areas, forming subgroups from the same population. The spatial distribution of these subgroups and the way they interact define a metapopulation. Metapopulations are described by their patch size and by the degree of isolation between its subunits, *i.e.* they may or may not interact as individual members move from one population to another (fig. 1.9).

Real metapopulations belong to the entire set of possible metapopulations whose extremes can be described as *Patchy*, *Classical*, *Mainland-island* or *non-equilibrium* metapopulations (Harrison and Taylor, 1997). Patchy populations are featured by high dispersal between habitat patches so much that individuals from different patches mix freely, forming effectively a single population. Classical metapopulations have habitat patches with similar probabilities of extinction and the persistence of these metapopulations is dependent on the recolonization of locally extinct patches. Mainland-island metapopulations have a local population that is extinction resistant (*i.e.* mainland) and other local populations that have much higher extinction probabilities (*i.e.* island), but are maintained by dispersal from Mainlands.

Population subdivision naturally gives the opportunity to define different forms of inbreeding depression and heterosis. Roze and Rousset (2004) investigated the combined effect of population structure and rate of selfing on the efficiency of selection against recurrent deleterious mutations, assuming an island model of population structure. They defined within-deme inbreeding depression as the fitness reduction of selfed progeny relative to out-crossed progeny from the same deme, between-deme inbreeding depression as the reduction in fitness of selfed progeny relative to progeny obtained by out-crossing randomly over the whole meta-population and heterosis as the difference between the fitness of the out-crossed progeny within deme and the out-crossed progeny over the whole meta-population.

They showed that selfing reduced within-deme inbreeding depression, between-deme inbreeding depression and heterosis. Between-deme inbreeding depression decreased with the degree of subdivision of the meta-population while within-deme inbreeding depression and heterosis increased. Hence, from a population genetics point of view, **heterosis is expected even in predominantly selfing species in subdivided population.** Thus it is important to note that heterosis and inbreeding depression are not mirror images of each other. Heterosis arises when deleterious, recessive mutations fixed within parental populations are in the heterozygous state by out-crossing, while inbreeding depression is usually attributed to the expression of recessive deleterious mutations when they become homozygous in inbred individuals.

Remarks

- Sex and mating systems can be viewed as quantitative traits that evolve to minimize the genetic load.
- Epistasis and recombination may explain the persistence of inbreeding depression at equilibrium for intermediate levels of sex/out-crossing rates.
- At a metapopulation level, inbreeding depression and heterosis do not evolve in the same manner.

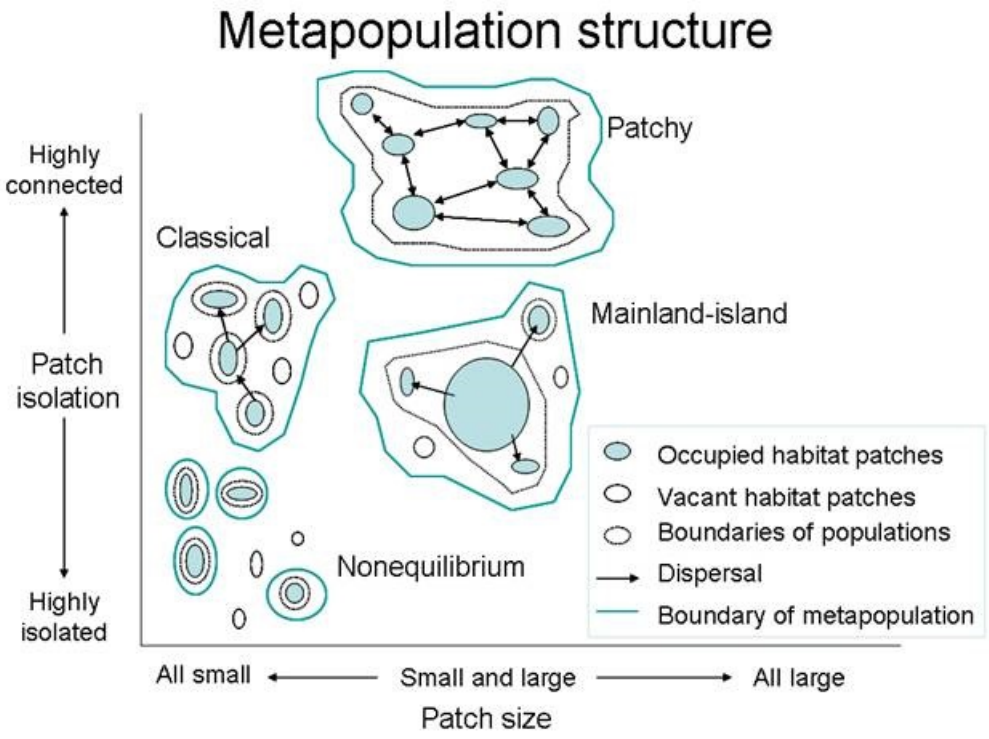


Figure 1.9: Metapopulation structure based on Stith et al. (1996) and Harrison and Taylor (1997). Circles in light blue represent occupied habitat patches, white circles represent vacant (unoccupied) habitat patches. Green (black) closed lines represent the boundaries of local metapopulations (populations, respectively) and arrows represent dispersal. Metapopulation structure is defined by means of patch size and patch isolation. Patch size and degree of isolation of the metapopulation are a measure of its probability of extintion.

1.3 Inbreeding depression and heterosis, the breeder's perspective

The relative parts of additive, inbreeding and heterosis effects on phenotypic variation are crucial for understanding the evolutionary potential of a population. Numerous have been the experimental designs and the statistical methods proposed to address this question (Cochran and Cox, 1950)

In a breeding perspective, Shull (1908) was the first to record experiments on heterosis and inbreeding depression, observing that when plants were self-pollinated, offspring performance declines in terms of growth and grain yield. However, when unrelated inbred lines were crossed the growth and yield performances of the hybrid progeny usually exceeded that of the best parent. His pioneer work in maize predicted that given the large amounts of heterosis within this species, the best way to maximize yield was to create inbreds from existing population varieties in order to seek for the best hybrid combinations. To this end, Sprague and Tatum (1942) developed quantitative genetic techniques to assess the relative importance of additive and non additive effects in trials of single-cross hybrids. In particular, they proposed to move from the analysis of population varieties to hybrid varieties, by estimating parameters on single lines that could be used for the selection of parents and the development of new lines.

They designated *General Combining Ability* (GCA) the average performance of a line in hybrid combinations, and *Specific Combining Ability* (SCA) the difference between the mean phenotypic value of the progeny and the average performance of the parental lines.

Subsequently, diallel designs were popularized as the most comprehensive designs for estimating genetic effects, predicting hybrid values and generating breeding populations to be used as basis

for selection and development of elite varieties (*e.g.* Hallauer and Filho (1988)).

The simplest and most popular decomposition of genetic effects in diallel designs is that of Griffing (1956), in which the mean phenotypic value of a cross between lines i and j is modeled as:

$$y_{ij} = \mu + GCA_i + GCA_j + SCA_{ij} \tag{1.52}$$

where μ is the mean phenotypic value of the population.

Diallel designs

Diallel designs are mating schemes used by plant/animal breeders and geneticists to investigate the genetic underpinnings of quantitative traits. They are constructed by pairwise crossing a set of inbred lines to obtain F1 hybrids. In a full diallel, all parents are crossed to make hybrids in all possible combinations. Variations include half-diallels with or without parents, omitting reciprocal crosses.

A few years later, Eberhart and Gardner (1966) stated that when are included in the diallel both "crossed varieties" and "selfed varieties", combining abilities can be separated to include heterosis and inbreeding effects. The model writes:

$$y_{ij} = \mu + \frac{1}{2}(a_i + d_i) + \frac{1}{2}(a_j + d_j) + \gamma(h_{ij} + \bar{h} + h_i + h_j) \tag{1.53}$$

where a_i (a_j, respectively) is the average performance of line i (j) in hybrid combinations, d_i (d_j, respectively) is the variety inbreeding; h_{ij} is the specific heterosis (difference between the hybrid and all hybrids sharing at least one parent); $\bar{h}$ is the average heterosis (average difference between inbreds and outbreds) and h_i (h_j) is the variety heterosis (average difference between the inbred parent i (j) and all crosses sharing the same parents); γ is an indicator variable that takes value 1 if $i = j$ and 0 otherwise.

Numerous other extensions have been proposed to extract other effects, such as maternal and paternal effects or sex-linked variations (Cockerham and Weir, 1977; Bulmer, 1980; Zhu and Weir, 1996; Greenberg et al., 2010). Recently, Lenarcic et al. (2012) have proposed a comprehensive model able to decompose the diallel into multiple genetic effects: additive, inbreeding and dominance, parent of origin (mitochondrial), symmetric and asymmetric interactions and sex specific effects. The full model of the phenotypic value of a cross between parents i and j, in replica k, reads:

$$\begin{aligned} y_{ijk} = \mu + \underbrace{\boldsymbol{x}_k^\top \boldsymbol{\beta}}_{\text{user fixed}} + \underbrace{\sum_{r=1}^{R} u_k^{(r)}}_{\text{user random}} + \underbrace{a_{i[k]} + a_{j[k]}}_{\text{additive}} + \underbrace{I_{i[k]=j[k]}(\beta_{\text{inbred}} + b_{i[k]})}_{\text{inbred penalty}} + \underbrace{m_{i[k]} - m_{j[k]}}_{\text{maternal}} + \\ \underbrace{I_{i[k]\neq j[k]} v_{(ij[k])}}_{\text{symmetric}} + \underbrace{I_{i[k]\neq j[k]} w_{(ij[k])}}_{\text{asymmetric}} + \underbrace{\psi(\text{sex}_k)(\phi^a_{i[k]} + \phi^a_{j[k]})}_{\text{sex-specific additive}} + \\ \underbrace{\psi(\text{sex}_k) I_{i[k]=j[k]}(\beta_{\text{female inbred}} + \phi^b_{i[k]})}_{\text{sex-specific inbred penalty}} + \underbrace{\psi(\text{sex}_k)(\phi^m_{i[k]} + \phi^m_{j[k]})}_{\text{sex-specific maternal}} + \\ \underbrace{\psi(\text{sex}_k) I_{i[k]\neq j[k]} \phi^v_{ij[k]}}_{\text{sex-specific symmetric}} + \underbrace{\psi(\text{sex}_k) I_{i[k]\neq j[k]} \phi^w_{ij[k]}}_{\text{sex-specific asymmetric}} + \epsilon_i \end{aligned} \tag{1.54}$$

The model allows for inclusion of fixed covariates $\boldsymbol{x}_k$ and R random-effect components

$$u_k^{(r)} \sim \mathcal{N}(0, \tau_r^2), \quad \forall r \in 1, \ldots, R \tag{1.55}$$

other than genetic effects. Along with the model, Lenarcic et al. (2012) proposed a hierarchical and Bayesian approach for the estimation of the parameters of interest. In particular, genetic effects

are modeled hierarchically and as drawn from a common normal distribution, *i.e.* additive genetic effects are assumed $a_i \sim \mathcal{N}(0, \sigma_a^2)$, $\forall i$.

In this context, we adapted the model described above to our particular half-diallel design (presented in Chapter 2) that includes the diagonal with parental inbred strains from two species. Thus we included in our model intra- and inter-specific additive effects, inbreeding effects and intra- and inter-specific heterosis effects.

Formally, let y_{ijk} be the observed phenotype for the cross between parents i and j in replica k. Our model reads:

$$\begin{aligned} y_{ijk} = \mu + I_{s(i)=s(j)}(A_{w_i} + A_{w_j}) + I_{s(i)\neq s(j)}(A_{b_i} + A_{b_j}) + \\ + I_{i\neq j}(I_{s(i)=s(j)}H_{w_{ij}} + I_{s(i)\neq s(j)}H_{b_{ij}}) + \\ + I_{i=j}(\beta_{s(i)} + B_i) + \epsilon_{ijk}, \end{aligned} \tag{1.56}$$

where:

- μ is the overall mean;
- $s(i)$ associates to each parental strain i the specie it belongs to:

$$s(i) \in \{S.\ cerevisiae, S.\ uvarum\}$$

- A_{w_i} and A_{b_i} denote, respectively, the additive contributions of strain i in intra-specific (within species, *i.e.* $s(i) = s(j)$), and inter-specific (between species, *i.e.* $s(i) \neq s(j)$) crosses;
- $H_{w_{ij}}$ and $H_{b_{ij}}$ denote the interaction effect between parents (i, j) in intra-specific (within species) and inter-specific (between species) crosses, respectively. In our half-diallel design with no reciprocal crosses, they are assumed to be symmetric, *i.e.* $H_{w_{ij}} = H_{w_{ji}}$ and $H_{b_{ij}} = H_{b_{ji}}$. Hereafter we will refer to these effects as intra- and inter-specific heterosis effects, respectively;
- $\beta_{s(i)}$ and B_i are, respectively, the deviation from the fixed overall effect for the species $s(i)$ and the associated strain-specific contribution of strain i in the case of inbred lines. Hereafter we will refer to B_i as inbreeding effect;
- ϵ_{ijk} is the residual, the specific deviation of individual ijk;
- $I_{condition}$ is an indicator variable. Its value is equal to 1 if the condition is satisfied and 0 otherwise.

Therefore, for the parental lines we have:

$$y^p_{iik} = \mu + 2A_{w_i} + \beta_{s(i)} + B_i + \epsilon_{iik}, \tag{1.57}$$

for the intra-specific hybrids:

$$y^{intra}_{ijk} = \mu + A_{w_i} + A_{w_j} + H_{w_{ij}} + \epsilon_{ijk}, \tag{1.58}$$

and for the inter-specific hybrids:

$$y^{inter}_{ijk} = \mu + A_{b_i} + A_{b_j} + H_{b_{ij}} + \epsilon_{ijk}. \tag{1.59}$$

All genetic effects were considered as random variables drawn from a normal distribution. Formally, letting $\boldsymbol{q} \in \{\boldsymbol{A_w}, \boldsymbol{A_b}, \boldsymbol{B}, \boldsymbol{H_w}, \boldsymbol{H_b}\}$ denote the genetic effect under consideration:

$$\forall i \quad q_i \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{q}}). \tag{1.60}$$

The full mixed-effect genetic model is thus defined by three fixed effects (the intercept μ and the inbreeding effects β_{Su} and β_{Sc}) and five genetic random effect variances ($\sigma^2_{\boldsymbol{A_w}}$, $\sigma^2_{\boldsymbol{A_b}}$, $\sigma^2_{\boldsymbol{B}}$, $\sigma^2_{\boldsymbol{H_w}}$, $\sigma^2_{\boldsymbol{H_b}}$).

In Chapter 3, I present the detailed description of our findings.

Remarks

- Quantitative traits are described in populations by variance components.
- Genetic variance components reflect both allele frequency and genetic effects at the underlying loci.
- All genetic variance components are important in determining the population response to evolutionary pressures.
- They can be estimated using dedicated cross designs.

Bibliography

Andersson, L. (2001). Genetic dissection of phenotypic diversity in farm animals, *Nature Reviews Genetics* **2**(2): 130–138.

Bulmer, M. G. (1971). The Effect of Selection on Genetic Variability, *The American Naturalist* **105**(943): 201–211.

Bulmer, M. G. (1980). *The mathematical theory of quantitative genetics*, Clarendon Press ; New York : Oxford University Press Oxford.

Charlesworth, B., Morgan, M. T. and Charlesworth, D. (1991). Multilocus models of inbreeding depression with synergistic selection and partial self-fertilization, *Genetics Research* **57**(2): 177–194.

Chevin, L.-M. and Lande, R. (2015). Evolution of environmental cues for phenotypic plasticity, *Evolution* **69**(10): 2767–2775.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/evo.12755*

Cochran, W. and Cox, G. M. (1950). *Experimental designs, 2nd ed*, Experimental designs, 2nd ed, Wiley, Oxford, England.

Cockerham, C. C. and Weir, B. S. (1977). Quadratic analyses of reciprocal crosses, *Biometrics* **33**(1): 187–203.

Crow, J. F. (1960). The Genetic Basis of Selection. I. Michael Lerner. Wiley, New York, 1958. xvi + 298 pp. Illus. $8, *Science* **131**(3405): 979–980.

Crow, J. F. (2001). Genetic Load, *in* S. Brenner and J. H. Miller (eds), *Encyclopedia of Genetics*, Academic Press, New York, pp. 838–839.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.*, John Murray, London, England.

Eberhart, S. A. and Gardner, C. O. (1966). A General Model for Genetic Effects, *Biometrics* **22**(4): 864–881.

Fisher, R. A. (1919). Xv.—the correlation between relatives on the supposition of mendelian inheritance., *Transactions of the Royal Society of Edinburgh* **52**(2): 399–433.

Fiévet, J. B., Nidelet, T., Dillmann, C. and de Vienne, D. (2018). Heterosis Is a Systemic Property Emerging From Non-linear Genotype-Phenotype Relationships: Evidence From in Vitro Genetics and Computer Simulations, *Frontiers in Genetics* **9**.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5968397/*

Forsman, A. (2014). Effects of genotypic and phenotypic variation on establishment are important for conservation, invasion, and infection biology, *Proceedings of the National Academy of Sciences* **111**(1): 302–307.

Gillespie, J. H. (2004). *Population Genetics: A Concise Guide*, JHU Press. Google-Books-ID: KAcAfiyHpcoC.

Greenberg, A. J., Hackett, S. R., Harshman, L. G. and Clark, A. G. (2010). A hierarchical bayesian model for a novel sparse partial diallel crossing design, *Genetics* .

Griffing, B. (1956). Concept of general and specific combining ability in relation to diallel crossing systems, *Australian Journal of Biological Sciences* **9**: 463–493.

Haag, C. R. and Roze, D. (2007). Genetic Load in Sexual and Asexual Diploids: Segregation, Dominance and Genetic Drift, *Genetics* **176**(3): 1663–1678.

Haldane, J. B. S. (1937). The Effect of Variation on Fitness, *The American Naturalist* **71**(735): 337–349.

Haldane, J. B. S. and Jayakar, S. D. (1963). Polymorphism due to selection of varying direction, *Journal of Genetics* **58**(2): 237–242.

Hallauer, A. R. and Filho, J. B. M. (1988). *Quantitative Genetics in Maize Breeding*, Iowa State University Press.

Harrison, S. and Taylor, A. D. (1997). Empirical evidence for metapopulation dynamics, *Metapopulation biology*, Elsevier, pp. 27–42.

Hartl, D. L. and Clark, A. G. (1997). *Principles of Population Genetics, Third Edition*, Sinauer Associates.

Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection, *Genetical Research* **8**(3): 269–294.

Iserbyt, A., Bots, J., Van Gossum, H. and Sherratt, T. N. (2013). Negative frequency-dependent selection or alternative reproductive tactics: maintenance of female polymorphism in natural populations, *BMC Evolutionary Biology* **13**: 139.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3704290/*

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution, *Evolution* **30**(2): 314–334.

Lande, R. (1979). Quantitative Genetic Analysis of Multivariate Evolution, Applied to Brain: Body Size Allometry, *Evolution* **33**(1): 402–416.

Lande, R. and Schemske, D. W. (1985). The Evolution of Self-Fertilization and Inbreeding Depression in Plants. I. Genetic Models, *Evolution* **39**(1): 24–40.

Lenarcic, A. B., Svenson, K. L., Churchill, G. A. and Valdar, W. (2012). A general bayesian approach to analyzing diallel crosses of inbred strains, *Genetics* **190**(2): 413–435.

Lerner, I. M. (1954). *Genetic homeostasis*, New York: Wiley.

Muller, H. J. (1964). The relation of recombination to mutational advance, *Mutation Research/-Fundamental and Molecular Mechanisms of Mutagenesis* **1**(1): 2–9.

Otto, S. P., Scott, M. F. and Immler, S. (2015). Evolution of haploid selection in predominantly diploid organisms, *Proceedings of the National Academy of Sciences of the United States of America* **112**(52): 15952–15957.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4703000/*

Roze, D. (2014). Selection for sex in finite populations, *Journal of Evolutionary Biology* **27**(7): 1304–1322.

Roze, D. and Rousset, F. (2004). Joint effects of self-fertilization and population structure on mutation load, inbreeding depression and heterosis, *Genetics* **167**(2): 1001–1015.

Shull, G. H. (1908). The Composition of a Field of Maize, *Journal of Heredity* **os-4**(1): 296–301.

Sprague, G. F. and Tatum, L. A. (1942). General vs. Specific Combining Ability in Single Crosses of Corn 1, *Agronomy Journal* **34**(10): 923–932.

Stith, B. M., Fitzpatrick, J. W., Woolfenden, G. E. and Pranty, B. (1996). Classification and conservation of metapopulations: a case study of the Florida scrub jay, *Metapopulations and wildlife conservation. Island Press, Washington, DC, USA* pp. 187–215.

Vanhoenacker, E., Sandell, L. and Roze, D. (2018). Stabilizing selection, mutational bias, and the evolution of sex*, *Evolution* **72**(9): 1740–1758.

Zhu, J. and Weir, B. S. (1996). Mixed model approaches for diallel analysis based on a bio-model, *Genetical Research* **68**(3): 233–240.

Chapter 2

Chapter 2

The yeast model and the HeterosYeast Project

Yeast is part of a large group of unicellular fungi widespread in nature. It is a powerful model system to address core issues in evolutionary biology such us the architecture of the genome and its evolution, the ecological and genetic structure of natural populations, the mechanisms of selection that lead to adaptation and the evolution of sex and mating systems (Gu and Oliver, 2009).

There are many advantages working with yeast, in particular with *Saccharomyces cerevisiae*. It is easy to grow in the laboratory, genetically tractable, and has been used as a model system for studying eukaryotic cellular processes for over 50 years. These studies have provided insights into fundamental eukaryotic processes, including transcription, translation, RNA processing, cell signaling, cytoskeletal dynamics and vesicle trafficking. Presently, over 75% of yeast ORFs have known or predicted functions, and much of this information is easily accessible in a variety of databases on the world wide web (Chervitz et al., 1999; Payne and Garrels, 1997; Güldener et al., 2005; Bader et al., 2003; Habeler et al., 2002).

Beside, yeast is important in many areas, including agriculture, medicine, biotechnology and food industry. Specially, *S. cerevisiae* has been widely used for the production of wine, beer and bread, but also as cell factory for the production of recombinant proteins for use as pharmaceuticals (Nielsen, 2013), of bulk and fine chemicals (Kavšček et al., 2015) and more recently for the production of bio-ethanol (Mohd Azhar et al., 2017). Some processes, such as biofuel production or wine/beer making, require new yeasts to solve specific challenges, especially those associated with sustainability, novel flavors and altered alcohol contents. For instance the development of inter-specific strains, such as *S. cerevisiae* × *S. uvarum*, could be considered for the beer market.

2.1 Evolutionary history and domestication of *Saccharomyces cerevisiae* and *S. uvarum*

2.1.1 Evolutionary history

The first eukaryotic genome fully sequenced was the genome of *S. cerevisiae* (Goffeau et al., 1996). Subsequently, the genomes of about 40 yeast species have been sequenced, which has led to notable advances in our understanding of evolutionary mechanisms and to the construction of robust yeast phylogenies (fig. 2.1 to 2.4). Unexpectedly, extensive sequence divergence have been observed between lineages, reflecting major genomic changes that contrast with the conservation of biological properties of yeast for very long evolutionary times. Bottleneck events of clonal populations may explain this observations. Indeed, under favorable conditions the majority of yeast species can propagate indefinitely by mitotic divisions, *i.e.* without genetic exchange, forming large haploid or diploid clonal populations. For example, *S. cerevisiae* predominantly reproduces asexually, with a rate of sexual to asexual reproduction around 10^{-5} under optimal conditions. Accordingly, analysis of polymorphism at selected loci suggests that in nature genetic exchanges and recombination are limited in this species. Therefore sub-populations tend to form with independent accumulation of sequence variations. The genetic drift resulting from such a mode of propagation is high as it offers the possibility for non-optimized variants to survive and colonize novel niches (Dujon, 2010).

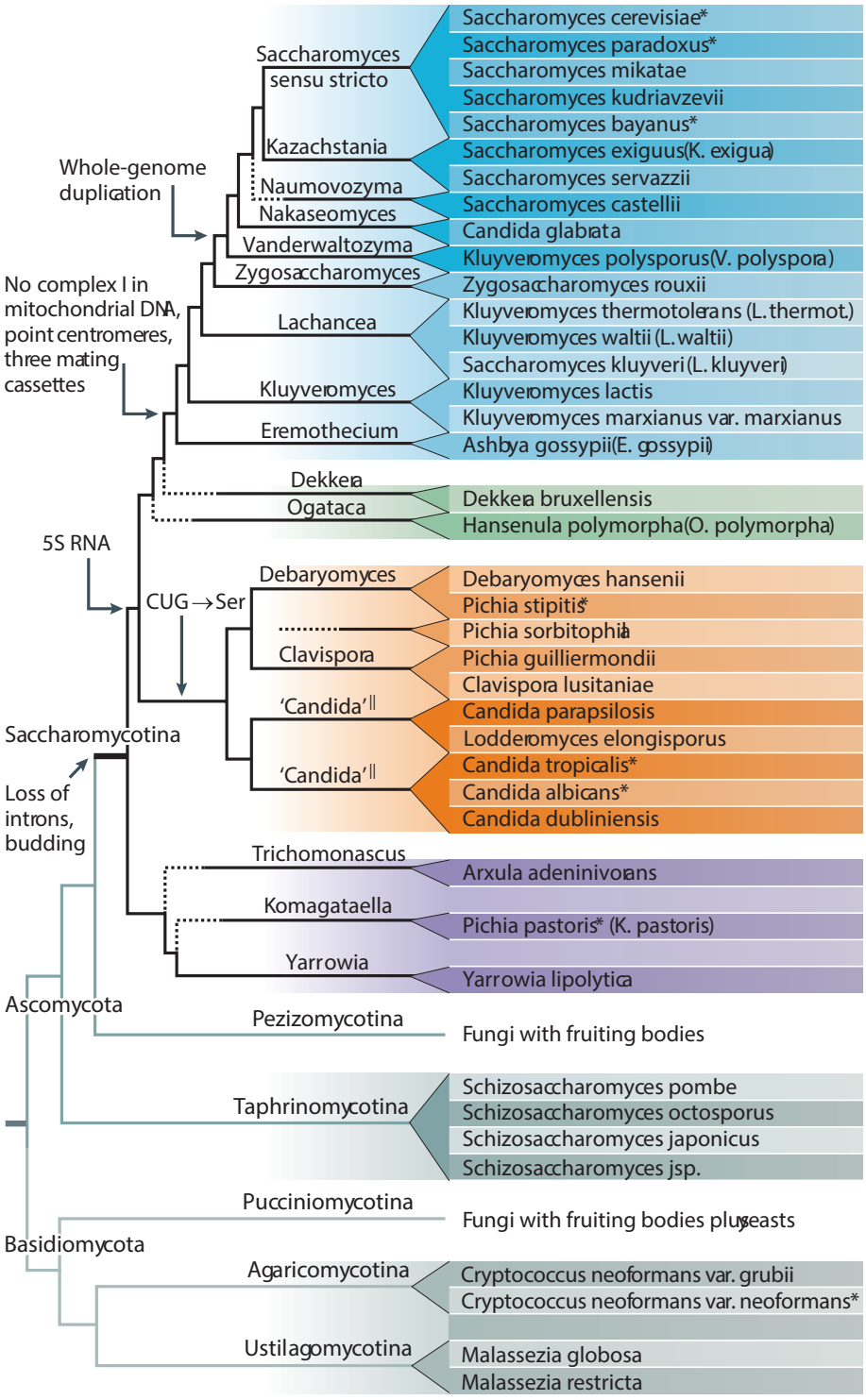


Figure 2.1: Overview of the sequenced yeast genomes (Dujon, 2010). Colored triangles represent clades or genera with their most recent designation (on the left). The dotted lines illustrate uncertainty and/or incongruence between different published phylogenies. Genomic architectures identify three major groups in Saccharomycotina: Saccharomycetacea (blue); CTG (or Candida) clade (orange); Dipodacaceae (purple). The arrows point to major evolutionary events. "*" Species for which several strains have been sequenced.

Nevertheless, inter-specific hybridization is not rare in yeast, and is accelerated by stressful conditions. Recent genomic studies have identified *S. pastorianus* as a hybrid between *S. cerevisiae* and *S. uvarum* (Libkind et al., 2011). However hybridization in *Saccharomyces sensu stricto* is generally accompanied by loss of genes, of chromosomal segments or of complete chromosomes, from which novel lineages could emerge. Those specific gene losses are expected to severely reduce the meiotic fertility of hybrids.

Species that belong to the same genus and share highly conserved gene synteny can exhibit large sequence divergence, as is the case for *S. cerevisiae* and species from the *S. bayanus* group (that includes *S. uvarum*). Experiments to investigate sequence divergence between yeast species from the same clade suggest that, assuming mutations to be neutral, independent and occurring at a rate of about 10^{-10}, yeast species derive from very recent clonal expansion from samples of large populations that had undergone similar successive bottlenecks. This would explain why *Saccharomyces sensu stricto* clade have nearly identical chromosomal maps interrupted by only a few chromosomal translocations.

Finally, the genome of *S. cerevisiae* contain DNA fragments from *S. paradoxus*, *S. kudriavzevii*, *S. uvarum* and *Zygosaccharomyces bailii*, suggesting that recent introgressions have occured. This process can be caused by the final step in nucleus fusing in inter-specific hybrids that allows for transfers of chromosomal fragments from one nucleus to another. The ecological proximity and selective pressures to adapt to high sugar, low-nitrogen and high-ethanol conditions during fermentation may facilitate this phenomenon, explaining the frequent introgressions observed in industrial *S. cerevisiae* strains. Domestication of *S. uvarum* is similarly supported by introgressions of genes from *S. eubayanus*, leading to over-representation of several gene categories involved in wine fermentation.

2.1.2 Domestication of *S. cerevisiae* and *S. uvarum*

Yeast species involved in alcoholic fermentation commonly belong to the clade of *Saccharomyces sensu stricto*, to which *S. cerevisiae* and *S. uvarum* are part. *S. cerevisiae* is extensively used in the food industry, for wine, beer, bread, etc., while *S. uvarum* has a more restricted use, for white wine fermentation, *e.g.* in the northern regions of France, and/or for red wine fermentation in Hungary, Italy and Spain. It is also the major yeast involved in cider making (Naumov et al., 2000).

S. cerevisiae is well known for its capacity of being highly fermentative, osmotolerant, heat resistant and to be able to survive in low pH environments. *S. uvarum* produces less acetic acid and ethanol, more glycerol and succinic acid, and synthesizes malic acid without posterior degradation. Furthermore, *S. uvarum* is well recognized for its ability in producing volatile compounds such as phenyl-ethanol, acetate and thiols, and for being cryotolerant.

In *Saccharomyces*, several cases of genome modifications through hybridization, introgressions and genome rearrangements have been documented. In particular, *S. cerevisiae* lineages used in the food industry have become genetically distinct from their wild relatives (Sicard and Legras, 2011), highlighting the human influence on their evolution (fig. 2.2). Oenological strains show a higher degree of heterozygosity as compared to strains in natural environments, reflecting a higher rate of sexual reproduction and/or an advantage of heterozygotes under oenological conditions (Hittinger, 2013). Moreover, the vast majority of oenological *S. cerevisiae* strains belong to the same genetic group, probably derived from a major domestication event of Mesopotamian origin, where most wine strains "migrated" through two major routes, the Danube valley and the Mediterranean Sea (Legras et al., 2007).

S. uvarum domestication has only recently been investigated (Almeida et al., 2014). *S. uvarum* would come from the super-continent Gondwana in the southern hemisphere. This was suggested by the fact that (i) its main host, *Nothofagus* tree, only lives in the southern hemisphere, and (ii) it displays in the southern hemisphere high genetic diversity and high frequency of isolation. In the northern hemisphere *S. uvarum* is the host of *Quercus* (oaks), which belongs to the same order

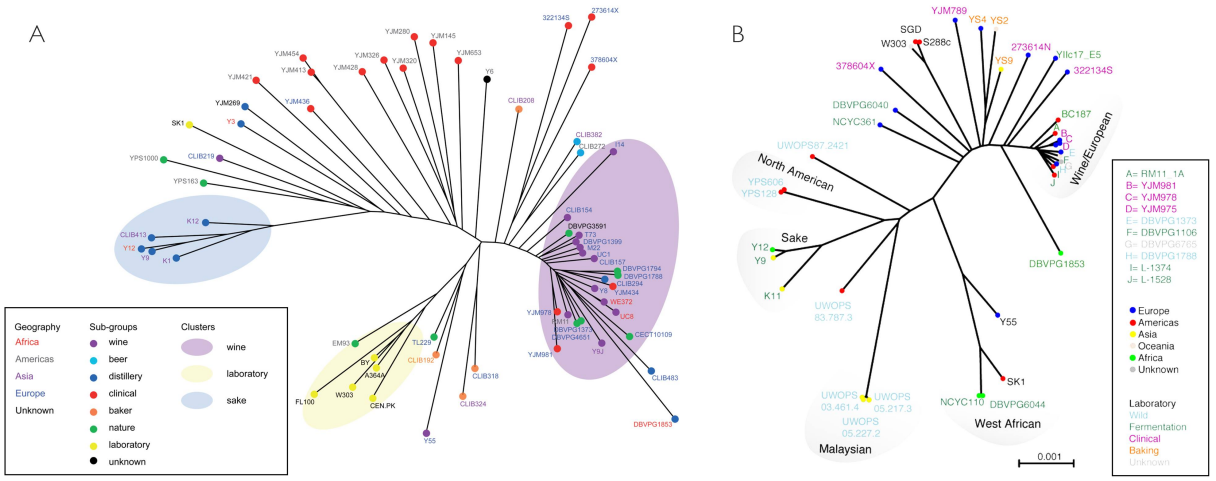


Figure 2.2: Neighbor-joining trees based on SNP differences of *S. cerevisiae* strains: **A**, branch lengths are proportional to the number of segregating sites that differentiate each pair of strains. Font color of strain name denotes geographic origin and circle color denotes ecological niche as specified in the key. (Schacherer et al., 2009). **B**, clean lineages highlighted in grey, with color indicating source (name) and geographic origin (dots) (Liti et al., 2009).

as *Nothofagus* (Fagales), and is also associated with anthropogenic environments such as wine-making and cider-making environments. Phylogenetic analysis resolved the various representatives of *S. uvarum* into three main clades showing high genetic differences (fig. 2.3): a first group composed of Holarctic strains (found in the northern hemisphere), a second group of South American strains and the third of Australasian strains. In the northern group genetic differences between strains are weak, with traces of recent hybridization with *Saccharomyces* strains from industrial environments.

In the clade *Saccharomyces sensu stricto*, inter-specific hybridizations between domesticated strains are common, as attested by the chromosomal introgressions (Sicard and Legras, 2011; Libkind et al., 2011; Giudici et al., 1998). For instance, *S. pastorianus* domesticated species is now known to come from the fusion of a *S. cerevisiae* ale-strain and *S. eubayanus*, a species recently isolated in Patagonia (Libkind et al., 2011), which is itself a hybrid between *S. cerevisiae* and a species related to the genetically complex *S. bayanus* group. The hybridization between *S. cerevisiae* and *S. eubayanus* has resulted in the creation of a hybrid with the strong fermentative ability of *S. cerevisiae* and the cold tolerance of *S. eubayanus* (Gibson and Liti, 2015).

More recently, the "Muri" strain, a unique hybrid between *S. cerevisiae* and *S. uvarum*, has been isolated from Norwegian farmhouse beer (fig. 2.4). The strain possesses a range of industrially desirable phenotypic properties such as broad temperature tolerance, ethanol resistance, efficient carbohydrate use and formation of desirable aroma-active esters (Krogerus et al., 2018). Identifying the mechanisms under selection during domestication process may clarify the emergence of new traits.

? Question

How does human selection targeted the ability to complete fermentation in yeast?

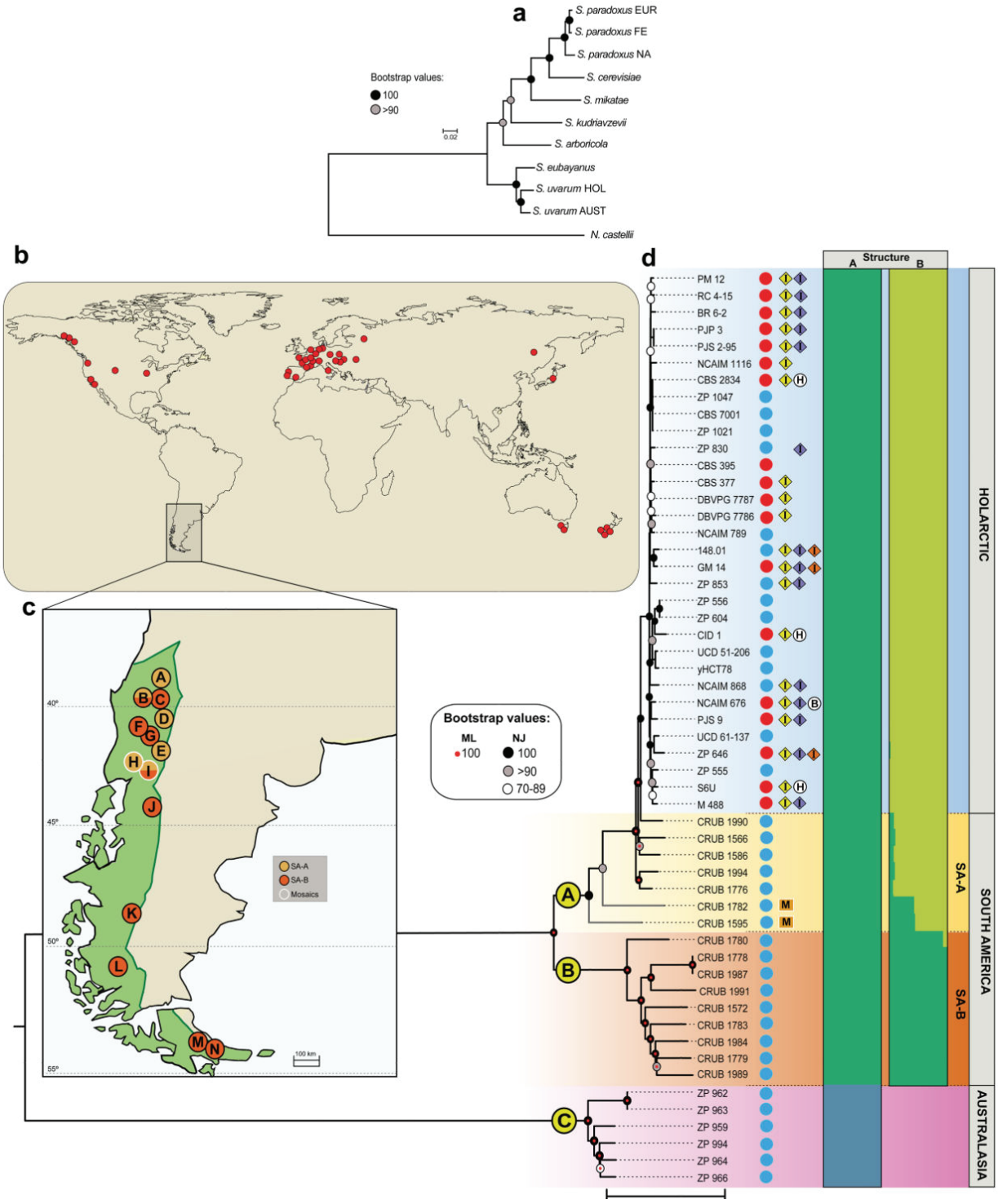


Figure 2.3: Geographic distribution, phylogeny and population structure of *S. uvarum*. **a**, maximum likelihood phylogeny of the genus *Saccharomyces* based on a concatenated alignment of 14 gene sequences; **b-c**, geographic origin of the different strains of *S. uvarum*; **d**, whole genome Neighbor-Joining phylogeny of 54 strains based on 129096 SNPs (Almeida et al., 2014)

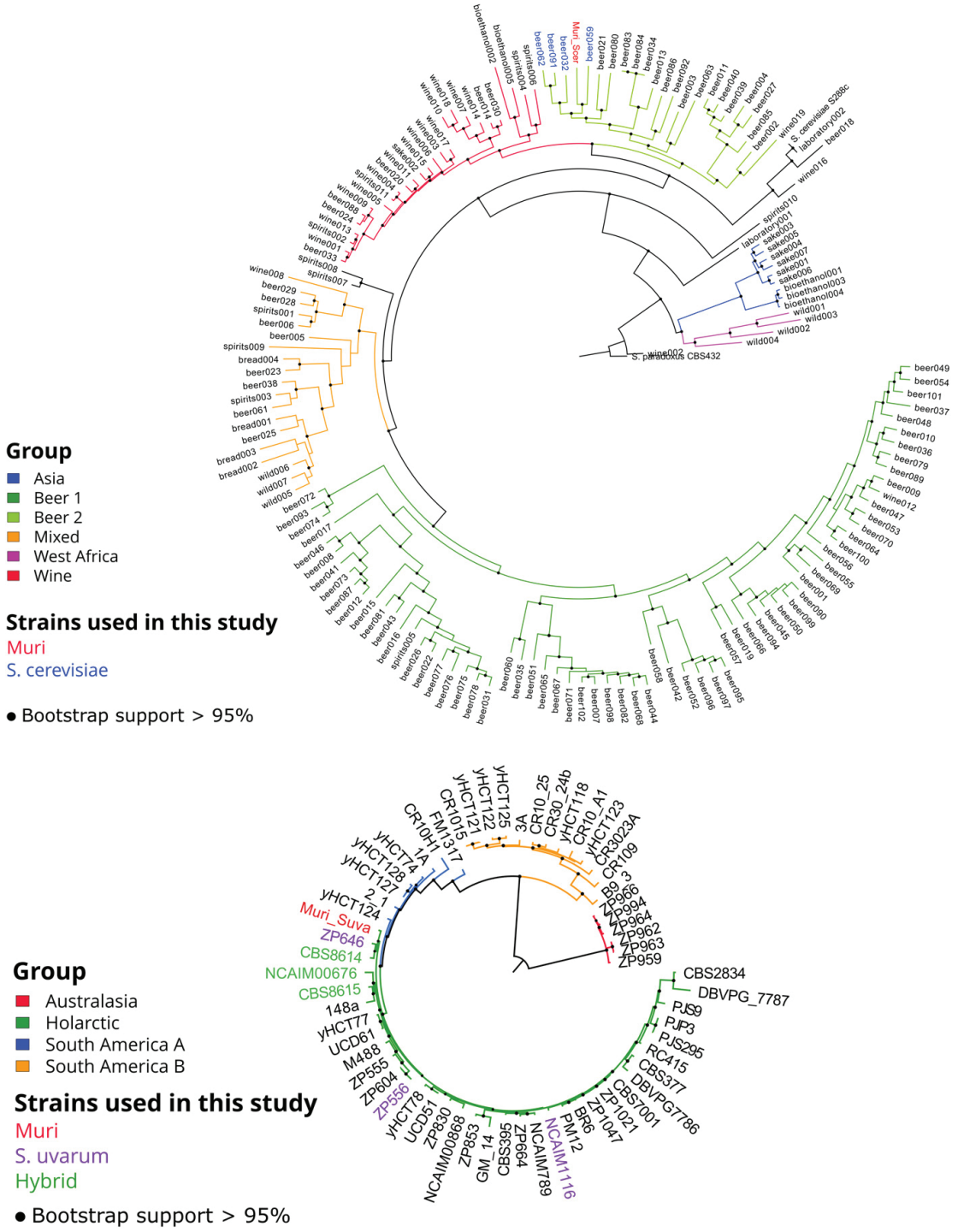


Figure 2.4: **A**, phylogeny of *S. cerevisiae* strains from Gallone et al. (2016) that shows that the Muri strains clusters with *S. cerevisiae* beer strains. **B**, phylogeny of *S. uvarum* and hybrid strains from Almeida et al. (2014) and Krogerus et al. (2018) that shows that Muri and other hybrid strains descent from the Holartic group. Branches are colored according to lineage. Muri strain is highlighted in red. Branch lengths represent the number of substitutions per site. Black dots on nodes indicate bootstrap support value $> 95\%$.

2.2 Variability of life-history and fermentation traits in yeast

2.2.1 Relationships between life-history traits and resource availability

The traits associated to the life-cycle, such as reproduction rate (r), carrying capacity (or maximum population size K) and cell size (S) are generally referred to as life-history traits. These traits are tightly linked to the fitness of the organism (Stearns and Hoekstra, 2005).

As stated above, yeasts have a complex life-cycle, largely studied in *S. cerevisiae*, that responds directly to environmental conditions and includes reproduction in both the haploid and diploid states via budding. Under non limiting conditions, diploid cells reproduce asexually. When nutrients are depleted, the mechanism of mating switches, cells enter in meiosis and sporulate. Spores divide equally into Mat***a*** and Mat***α*** mating-types. Haploid cells can then reproduce vegetatively through budding, or can mate with the opposite mating-type (Bardwell, 2004; Greig and Leu, 2009). Therefore there are tight interrelations between life-history traits, which can be the result of both evolutionary processes and physico-chemical cellular constraints. The balance of energy allocation to reproduction, growth or survival represents a life-history strategy (Schluter Dolph et al., 1991). Spor et al. (2008) have shown in *S. cerevisiae* that there is a continuum of strategies distributed between two extremes: the "ant" and the "grasshopper" strategies. In batch cultures, yeasts first consume glucose through fermentation. When glucose is exhausted, metabolism switches to respiration. The "ant" strategy consists of quick reproduction (high r_f), high carrying capacity (high K), and small cell size (small S) in fermentation, but low reproduction rate r_r in respiration. The "grasshopper" strategy consists of slow reproduction (low r_f), low carrying capacity (low K), large cell size (large S) in fermentation and high reproduction rate in respiration (high r_r). The strategy chosen by *S. cerevisiae* strains depends on the ecological niche. In particular, forest and laboratory strains generally adopt the "ant" strategy, while industrial strains opt for the "grasshopper" strategy (Spor et al., 2008).

The differences in life-history traits reflect differences in habitats of origin: strains from similar habitats (even geographically isolated) have similar life-history strategies, *i.e.* niche-driven evolution had probably led to phenotypic convergence. To investigate this point, Spor et al. (2014) performed an evolutionary experiment with six yeast strains, chosen along the K-cell size gradient, in environments differing for the amount of resources (1% and 15% of glucose) and the time spent in the media (48h and 96h). Experiments were performed independently in batch cultures for the four environments. The authors showed that each ancestral strain evolved different combinations of life-history traits under the different selection regimes, adapting to the local conditions. The strains evolved under the same selection regime developed similar life-history traits. Strains adopted the "ant" strategy in poor media, with low glucose consumption, whereas strains in rich media selected the "grasshopper" strategy with high glucose consumption rate. Therefore, the K-cell size trade-off seems to be explained by resource availability. Phenotypic convergence could be partly accounted for by selection of mutations in genes involved in the same pathways. In particular, Spor et al. (2014) identified mutations at the BMH1 locus with antagonistic phenotypic effects depending on the selection regime.

2.2.2 Fermentation trait variation is linked to life-history traits

Similarly, environment is the main factor shaping alcoholic fermentation. Albertin et al. (2011), studying the ability of nine different strains of *S. cerevisiae* from winery, brewing and distillery origins, have shown that glucose uptake displays plastic and genetic variability. Oenological strains consume all sugar and produced more CO_2 in less time in oenology medium than beer and distillery strains, which displayed slow or incomplete fermentation. In the brewer and bakery mediums the maximum CO_2 release rate (V_{max}) was higher and was reached faster, and fermentation ended faster, than in the oenology medium.

Furthermore, Albertin et al. (2011) have shown that V_{max} is highly correlated with K and not with J_{max}, the maximum CO_2 release rate per cell, suggesting that human selection targeted the ability to complete fermentation by influencing the ability to reproduce rather than the metabolic efficiency. Similarly, K was significantly correlated to nitrogen consumption and biomass, but negatively correlated to the amount of acetic acid and trehalose measured at the end of fermentation. again, K was found to be negatively correlated with cell size, while cell size was positively correlated with trehalose and the reproduction rate in respiration (r_r), and with J_{max}.

2.2.3 Relation *K*-cell size and protein abundance variation

The trade-off between K and cell size is robust: it has been found in yeast isolated in natural populations (Spor et al., 2009), in industrial strains associated to different food processes (Spor et al., 2008; Albertin et al., 2011) and in strains derived from experimental evolution (Spor et al., 2014).

Albertin et al. (2013A) analyzed this evolutionary constraint with quantitative proteomics, focusing on the abundances of the enzymes and isoforms of alcoholic fermentation, using the same nine food-processing strains as those of Albertin et al. (2011). They showed that the enzymatic pool allocated to the fermentation proteome was constant over the culture media and the strains, but there was variability in abundance of individual enzymes and sometimes much more of their post-translationally modified isoforms. This suggests the existence of selective constraints on total protein abundance and trade-offs between isoforms. Interestingly, abundance variation of some isoforms was significantly associated to metabolic traits and growth-related traits. In particular, cell size and K were highly correlated with the degree of N-terminal acetylation of the alcohol dehydrogenase. Thus the fermentation proteome was found to be shaped by human selection, through the differential targeting of a few isoforms for each food-processing origin of strains. These results highlighted the importance of post-translational modifications in the diversity of metabolic and life-history traits.

Remarks

Understanding the mechanisms shaping yeast biodiversity needs a comprehensive study of the different levels of cellular organization and analysis of their relationships, from the molecular and genetic point of view.

2.3 HeterosYeast: Exploitation of the heterosis phenomenon for wine yeast improvement

In the continuity of the previous studies, the ANR interdisciplinary project “HeterosYeast: Exploitation of the heterosis phenomenon for wine yeast improvement”, 2009-2013, coordinated by Dominique de Vienne and Philippe Marullo, provided a large set of heterogeneous data to investigate heterosis for fermentation and life-history trait variation. The HeterosYeast project focused on three tightly related goals: better understanding of the genetic and molecular bases of heterosis, developing predictors of heterosis and, in the long run, derive yeast hybrid strains with high oenological performance. HeterosYeast relied on a diallel design, which is the most comprehensive design to decompose the genetic effect and their variance for quantitative traits, as stated in section 1.3.

2.3.1 Construction of the diallel design

Parental strains

Among the myriad of yeast species, *S. cerevisiae* and *S. uvarum* have been chosen since they are characterized by the ability to achieve grape must fermentation. They differ in their habitat and in a number of phenotypic traits, but natural hybrids between the two species exists. The original strains of the experimental design were seven *S. cerevisiae* and four *S. uvarum* strains associated to various food processes (oenology, brewery, cider fermentation and distillery) or isolated from natural environments (oak exudates) (tab. 2.1).

Table 2.1: Parental yeast strains used for the construction of the diallel design. All strains are diploid. They come from various origins and are associated to different food processes. Homozygous diploid strains are named "W", "D", "E" and "U", for forest, distillery, oenology and *uvarum* strains, respectively.

Original strains

Strains	Genotype	Species	Ploidy	Collection/Supplier	Origin	Reference
YSP128	HO/HO	S. cerevisiae	diploid	SGRP	Forest Oak exudate, Pennsylvania, USA	Liti et al. (2009)
Alcotec 24	ho/ho	S. cerevisiae	diploid	Hambleton Bard	Distillery, UK	Albertin et al. (2011)
CLIB-294	HO/HO	S. cerevisiae	diploid	CIRM-Levures	Distillery, Cognac, France	Albertin et al. (2011)
VL1	HO/HO	S. cerevisiae	diploid	Laffort Oenologie	Enology, Bordeaux, France	Marullo et al. (2006)
F10	HO/HO	S. cerevisiae	diploid	Laffort Oenologie	Enology, Bordeaux, France	Marullo et al. (2009)
VL3c	HO/HO	S. cerevisiae	diploid	Laffort Oenologie	Enology, Bordeaux, France	Marullo et al. (2004)
BO213	HO/HO	S. cerevisiae	diploid	Laffort Oenologie	Enology, Bordeaux, France	Marullo et al. (2006)
PM12	HO/HO	S. uvarum	diploid	ISVV	Grape must fermentation, Jurançon, France	Masneuf-Pomarède et al. (2007)
PJP3	HO/HO	S. uvarum	diploid	ISVV	Grape must fermentation, Sancerre, France	Masneuf-Pomarède et al. (2007)
Br6.2	HO/HO	S. uvarum	diploid	ADRIA Normandie	Cider fermentation, Normandie, France	Albertin et al. (2013A)
RC4-15	HO/HO	S. uvarum	diploid	ISVV	Grape must fermentation, Alsace, France	Masneuf-Pomarède et al. (2007)

Homozygous diploid parental strains

Strains	Genotype	Derivation	Ploidy	Collection/Supplier	Reference
W1	HO/HO	YSP128	diploid	ISVV	Blein-Nicolas et al. (2013)
D2	ho/ho	Alcotec24	diploid	ISVV	Albertin et al. (2011)
D1	HO/HO	CLIB-294	diploid	ISVV	Albertin et al. (2011)
E3	HO/HO	VL1	diploid	ISVV	Albertin et al. (2011)
E4	HO/HO	F10	diploid	ISVV	Albertin et al. (2011)
E5	HO/HO	VL3c	diploid	ISVV	Blein-Nicolas et al. (2013)
E2	HO/HO	BO213	diploid	ISVV	Marullo et al. (2009)
U1	HO/HO	PM12	diploid	ISVV	Blein-Nicolas et al. (2013)
U2	HO/HO	PJP3	diploid	ISVV	Blein-Nicolas et al. (2013)
U3	HO/HO	Br6.2	diploid	ISVV	Blein-Nicolas et al. (2013)
U4	HO/HO	RC4-15	diploid	ISVV	da Silva et al. (2015)

Nine out eleven strains were analyzed previous to the construction of the diallel (Blein-Nicolas et al., 2013; Marullo et al., 2009). The clustering of the lines depended on the type of trait considered (fig. 2.5). The strains of *S. uvarum* and a group of *S. cerevisiae* displayed similar fermentative performances despite strong proteomic and genomic differences. Indeed, the proteomes of the two species were contrasted, which could be related to a differential recruitment of proteins of the glucose pathway encoded by duplicated genes. Altogether, these results indicate that the ability of *S. cerevisae* and *S. uvarum* to complete grape fermentation must arise through different evolutionary roads, involving different metabolic pathways and sets of proteins (Blein-Nicolas et al., 2013).

This set of strains showing a high variability at every level of cellular organization seemed appropriate for the construction of the diallel cross. Nevertheless, they could not be used as such as parents of a diallel design because they were suspected to be heterozygous at many loci. The way the strains were made homozygous is described in details in (da Silva et al., 2015). Briefly, monosporic clones were isolated by tetrad dissection using a micromanipulator. All original strains but D2 were homothallic (HO/HO), therefore fully homozygous diploid strains were spontaneously obtained by fusion of opposite mating type cells. For D2 that was ho/ho, one isolated haploid meiospore was diploidized via transient expression of the HO endonuclease. These strains, called

W1, D1, D2, E2, E3, E4 and E5 for *S. cerevisiae* and U1, U2, U3 and U4 for *S. uvarum*, were used as the parental strains for the construction of a half diallel design with diagonal. All strains were grown at 24°C in YPD medium (da Silva et al., 2015).

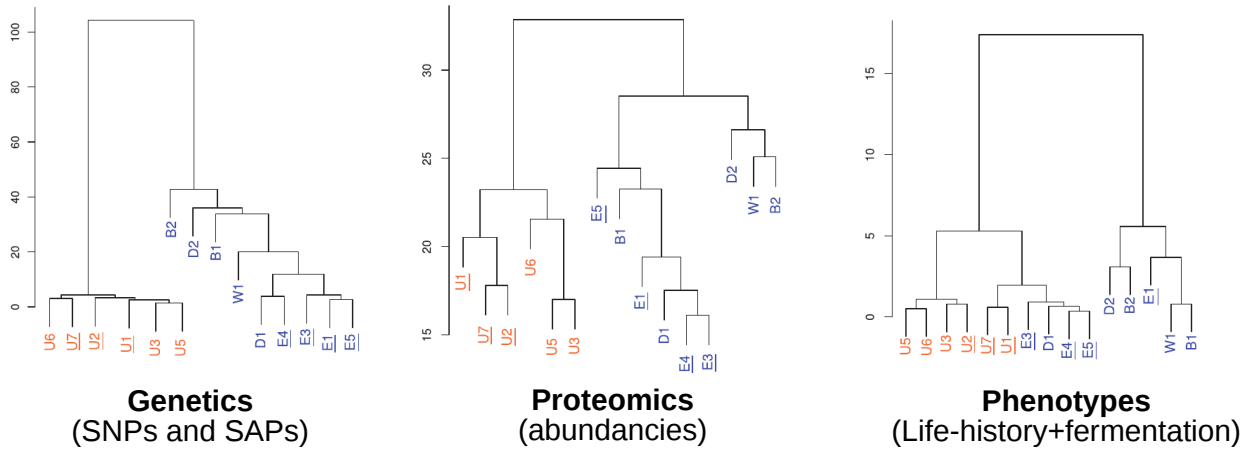


Figure 2.5: Clustering of *S. cerevisiae* and *S. uvarum* (Blein-Nicolas et al., 2013). Among the 15 strains analyzed in this study, nine have been employed as the parental strains in the diallel design of the HeterosYeast Project. Clustering of six strains of *S. uvarum* (orange) and nine strains of *S. cerevisiae* (blue) based on: (i) sequence variability inferred from 498 SNPs and 2681 SAPs (left); (ii) proteome variability assessed from abundances of 401 proteins (center); (iii) lag-phase time, times to complete 30%, 50% and 100% of fermentation, cell size, and population size at 30% of CO_2 release (right).

Hybrid construction

In order to produce intra- and interspecific hybrids, the eleven diploid parental strains were transformed with a cassette containing the HO allele disrupted by a gene of resistance, as previously described in Albertin et al. (2013B). After transformation, monosporic clones were isolated, and the mating-type (Mat***a*** or Mat***α***) of antibiotic-resistant clones was determined using testers of known mating-type. Strain transformation allowed conversion to heterothallism for the homothallic strains (all but D2) and antibiotic resistance allowed easy hybrid selection. For each hybrid construction, parental strains of opposite mating- types were put in contact for 2 to 6 hours in YPD medium at room temperature, and then plated on YPD-agar containing the appropriate antibiotics. The 55 possible hybrids from the 11 parental strains, namely 21 *S. cerevisiae* intraspecific hybrids, 6 *S. uvarum* intraspecific hybrids and 28 interspecific hybrids, were obtained. For each cross, a few independent colonies were collected. After recurrent cultures on YPD-agar corresponding to ∼80 generations, the nuclear chromosomal stability of the hybrids was controlled by pulsed field electrophoresis, as well as homoplasmy (only one parental mitochondrial genome) as detailed in Albertin et al. (2013B).

2.3.2 Phenotypic characterization

This unique biological material was grown in triplicate in fermentors with a medium close to oenological conditions at two temperatures (18° C and 26° C, optimum for *S.u.* and *S.c.*, respectively). Thus a total of 396 alcoholic fermentations were performed. In order to access a multi-level description of the heterosis phenomenon, two types of phenotypic traits were measured or estimated from sophisticated data adjustment models (da Silva et al., 2015; Blein-Nicolas et al., 2015):

- *Protein abundances.* Using high-throughput shotgun LC-MS/MS technique, the intensities of more than 10 000 peptides allowed estimating the abundances of ∼ 1400 proteins, and

as many as 97 360 protein-per-hybrid-per-temperature combinations were analyzed in the Pappso facility (*http://pappso.inra.fr*). The abundances of 615 proteins present in all strains were measured from both shared and proteotypic peptides relying on original Bayesian developments (Blein-Nicolas et al., 2012). Massive variations were found, that clearly differentiated the two species (see above). Heterosis was found for numerous proteins in variable proportions depending on the parental strain and on the temperature considered (from 8.4 % to 61.2 %). In the intra-specific hybrids, this proportion was higher at non-optimal temperature. Unexpectedly, heterosis for protein abundance was strongly biased toward positive values in inter-specific hybrids but not in intra-specific hybrids, and the proportion of hybrids in which a protein was heterotic was positively correlated to the number of putative transcription factors of the encoding gene. Computer simulations assuming concave relationships between protein abundances and their controlling factors accounted quite well for these observations (Blein-Nicolas et al., 2015), which is consistent with the role of non-linear processes in the emergence of heterosis (Fiévet et al., 2018).

- *Fermentative traits.* A total of 35 fermentative traits (~ 13 000 data points) were obtained, which were classified into: kinetics traits (estimated from the CO_2 release curve), population dynamics traits (estimated from cell concentrations over time), basic oenological products (ethanol, residual sugar, acetic acid, etc.), and aromatic traits. Mixed ANOVA models and multivariate analyses showed that, depending on the types of trait, the sources of variation (strain, temperature and strain × temperature effects) differed in a large extent. For instance the kinetics traits and some population traits (temporal variables, growth traits, CO_2 flux) were very sensitive to temperature, unlike key metabolites for oenology. However some of the latter and various population traits (maximum CO_2, carrying capacity, viability, cell size) exhibited large strain per temperature interactions. The global comparison of the three types of hybrids (*S.c.*×*S.c.*, *S.u.*×*S.u.* and *S.c.*×*S.c.*) revealed that hybridization could generate multi-trait phenotypes with improved oenological performances. In addition the inter-specific hybrids displayed better homeostasis with respect to temperature, which could explain why interspecific hybridization is so common in natural and domesticated yeasts, and open the way to applications for wine-making (da Silva et al., 2015).

Figure 2.6 summarizes the experimental protocol of the HeterosYeast project.

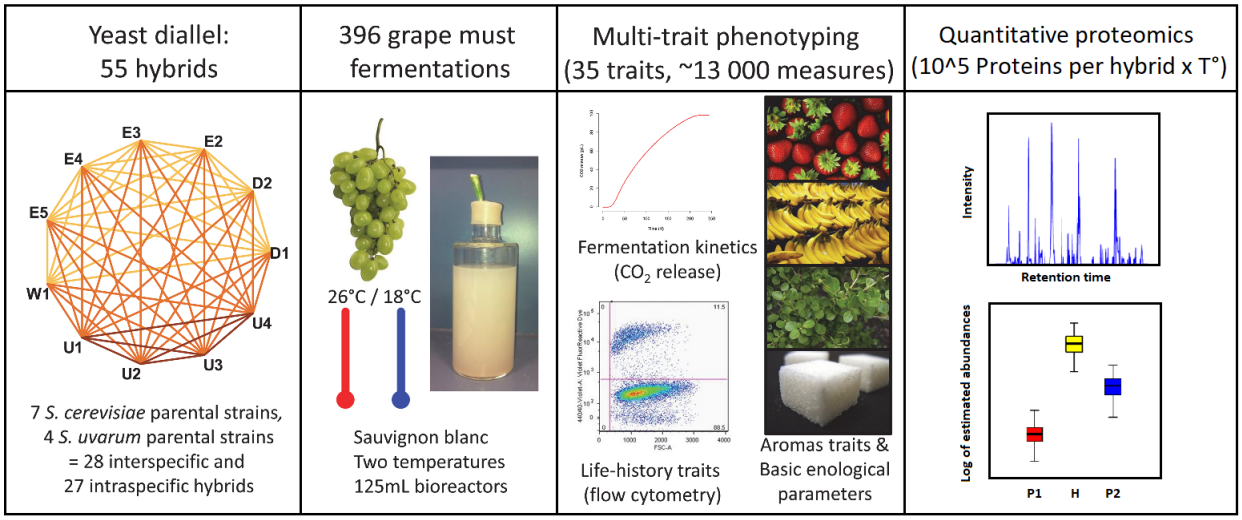


Figure 2.6: Experimental protocol. Fully homozygous diploid strains were used as parental strains in a half-diallel design. W1, D1, D2, E2, E3, E4 and E5 are *S. cerevisiae* strains, U1, U2, U3 and U4 *S. uvarum* strains. Fermentations were carried out in Sauvignon blanc grape juice and run at 18°C and 26°C in triplicate in fermentors for a total of 396 experiments. Thirty-five traits were collected and grouped into four classes (Fermentation Kinetics Traits, Life-history traits, Basic Oenological Parameters and Aromatic Traits). Protein abundances have been quantified for each strain × temperature combination (da Silva et al., 2015).

2.4 Aim of the thesis

The exceptional dataset produced in the HerosYeast project was far from being fully exploited. In particular such a set of heterogeneous data, which corresponds to different levels of cellular organization, was ideally convenient for multi-scale modelling and testing models for predicting the variation of integrated phenotypes from protein and metabolic traits, taking into account the dependence structures between variables, but also between observations.

The aim of my thesis was to develop original mathematical and statistical models in systems biology to investigate the molecular and genetic bases of phenotypic variation in yeast and to integrate different types of data measured at different scales.

I have adopted two main approaches to address these issues.

Analysis of the diallel design. A first goal was to characterize phenotypic variation at each level of cellular organization by means of genetic variance components. To this end, I exploited the particular half-diallel cross design to infer the parts of variance attributed to additive, inbreeding and heterosis effects for each trait, distinguishing intra and interspecific additive and heterosis effects. Then the integration of the different levels of cellular organization has been performed by clustering traits displaying similar partition of variance components, to search for parallel behaviour between proteins and life history/fermentation traits that could suggest functional links. A major finding of this first part of my thesis work is the possible decoupling between the heterosis and inbreeding variances (Chapter 3, article: **Decoupling the Variances of Heterosis and Inbreeding Effects Is Evidenced in Yeast's Life-History and Proteomic Traits** published in *Genetics*; Petrizzelli et al. (2019)).

Search for predictors of fermentation and life-history traits. The second part of the thesis work consisted in finding predictors of fermentation and life-history traits. To this end, I predicted an additional phenotypic level, the metabolic fluxes, which result from the metabolic network

functioning and integrate the activities of possibly many proteins. I proposed a novel method to introduce protein abundance data into constraint-based models and predicted steady-state fluxes for each strain separately. Finally, I used statistical approaches to integrate the three different levels of cellular organization to gain information on the metabolic and molecular predictors of the integrated traits. This constitutes the Chapter 5, **Data integration uncovers the metabolic bases of phenotypic variation in yeast**, which will be submitted soon to *Molecular Systems Biology*.

Bibliography

Albertin, W., da Silva, T., Rigoulet, M., Salin, B., Masneuf-Pomarede, I., de Vienne, D., Sicard, D., Bely, M. and Marullo, P. (2013B). The mitochondrial genome impacts respiration but not fermentation in interspecific saccharomyces hybrids, *PLOS ONE* **8**(9): 1–14.

Albertin, W., Marullo, P., Aigle, M., Dillmann, C., Vienne, D. d., Bely, M. and Sicard, D. (2011). Population Size Drives Industrial Saccharomyces cerevisiae Alcoholic Fermentation and Is under Genetic Control, *Appl. Environ. Microbiol.* **77**(8): 2772–2784.
URL: *https://aem.asm.org/content/77/8/2772*

Albertin, W., Marullo, P., Bely, M., Aigle, M., Bourgais, A., Langella, O., Balliau, T., Chevret, D., Valot, B., Silva, T. d., Dillmann, C., Vienne, D. d. and Sicard, D. (2013A). Linking Post-Translational Modifications and Variation of Phenotypic Traits, *Molecular & Cellular Proteomics* **12**(3): 720–735.
URL: *http://www.mcponline.org/content/12/3/720*

Almeida, P., Gonçalves, C., Teixeira, S., Libkind, D., Bontrager, M., Masneuf-Pomarède, I., Albertin, W., Durrens, P., Sherman, D., Marullo, P., Hittinger, C. T., Gonçalves, P. and Sampaio, J. P. (2014). A Gondwanan Imprint on Global Diversity and Domestication of Wine and Cider Yeast Saccharomyces uvarum, *Nature communications* **5**: 4044.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5081218/*

Bader, G. D., Betel, D. and Hogue, C. W. V. (2003). BIND: the Biomolecular Interaction Network Database, *Nucleic Acids Research* **31**(1): 248–250.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165503/*

Bardwell, L. (2004). A walk-through of the yeast mating pheromone response pathway, *Peptides* **25**(9): 1465–1476.

Blein-Nicolas, M., Albertin, W., da Silva, T., Valot, B., Balliau, T., Masneuf-Pomarède, I., Bely, M., Marullo, P., Sicard, D., Dillmann, C., de Vienne, D. and Zivy, M. (2015). A Systems Approach to Elucidate Heterosis of Protein Abundances in Yeast, *Molecular & cellular proteomics: MCP* **14**(8): 2056–2071.

Blein-Nicolas, M., Albertin, W., Valot, B., Marullo, P., Sicard, D., Giraud, C., Huet, S., Bourgais, A., Dillmann, C., de Vienne, D. and Zivy, M. (2013). Yeast Proteome Variations Reveal Different Adaptive Responses to Grape Must Fermentation, *Molecular Biology and Evolution* **30**(6): 1368–1383.
URL: *https://doi.org/10.1093/molbev/mst050*

Blein-Nicolas, M., Xu, H., de Vienne, D., Giraud, C., Huet, S. and Zivy, M. (2012). Including shared peptides for estimating protein abundances: a significant improvement for quantitative proteomics, *Proteomics* **12**(18): 2797–2801.

Chervitz, S. A., Hester, E. T., Ball, C. A., Dolinski, K., Dwight, S. S., Harris, M. A., Juvik, G., Malekian, A., Roberts, S., Roe, T., Scafe, C., Schroeder, M., Sherlock, G., Weng, S., Zhu, Y., Cherry, J. M. and Botstein, D. (1999). Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure, *Nucleic Acids Research* **27**(1): 74–78.
URL: *https://academic.oup.com/nar/article/27/1/74/1243398*

da Silva, T., Albertin, W., Dillmann, C., Bely, M., la Guerche, S., Giraud, C., Huet, S., Sicard, D., Masneuf-Pomarede, I., de Vienne, D. and Marullo, P. (2015). Hybridization within Saccharomyces Genus Results in Homoeostasis and Phenotypic Novelty in Winemaking Conditions, *PLoS ONE* **10**(5).
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4422614/*

Dujon, B. (2010). Yeast evolutionary genomics, *Nature Reviews. Genetics* **11**(7): 512–524.

Fiévet, J. B., Nidelet, T., Dillmann, C. and de Vienne, D. (2018). Heterosis is a systemic property emerging from nonlinear genotype-phenotype relationships: evidence from in vitro genetics and computer simulations, *Frontiers in Genetics* **9**.
URL: *https://www.frontiersin.org/articles/10.3389/fgene.2018.00159/abstract*

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., Teiling, C., Steffy, B., Taylor, M., Schwartz, A., Richardson, T., White, C., Baele, G., Maere, S. and Verstrepen, K. J. (2016). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts, *Cell* **166**(6): 1397–1410.e16.

Gibson, B. and Liti, G. (2015). Saccharomyces pastorianus: genomic insights inspiring innovation for industry, *Yeast (Chichester, England)* **32**(1): 17–27.

Giudici, P., Caggia, C., Pulvirenti, A. and Rainieri, S. (1998). Karyotyping of Saccharomyces strains with different temperature profiles, *Journal of Applied Microbiology* **84**(5): 811–819.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996). Life with 6000 Genes, *Science* **274**(5287): 546–567.
URL: *https://science.sciencemag.org/content/274/5287/546*

Greig, D. and Leu, J.-Y. (2009). Natural history of budding yeast, *Current Biology* **19**(19): R886–R890.
URL: *https://www.cell.com/current-biology/abstract/S0960-9822(09)01461-4*

Gu, Z. and Oliver, S. (2009). Yeasts as models in evolutionary biology, *Genome Biology* **10**(3): 304.

Güldener, U., Münsterkötter, M., Kastenmüller, G., Strack, N., van Helden, J., Lemer, C., Richelles, J., Wodak, S. J., García-Martínez, J., Pérez-Ortín, J. E., Michael, H., Kaps, A., Talla, E., Dujon, B., André, B., Souciet, J. L., De Montigny, J., Bon, E., Gaillardin, C. and Mewes, H. W. (2005). CYGD: the Comprehensive Yeast Genome Database, *Nucleic Acids Research* **33**(Database issue): D364–368.

Habeler, G., Natter, K., Thallinger, G. G., Crawford, M. E., Kohlwein, S. D. and Trajanoski, Z. (2002). YPL.db: the Yeast Protein Localization database, *Nucleic Acids Research* **30**(1): 80–83.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC99114/*

Hittinger, C. T. (2013). Saccharomyces diversity and evolution: a budding model genus, *Trends in genetics: TIG* **29**(5): 309–317.

Kavšček, M., Stražar, M., Curk, T., Natter, K. and Petrovič, U. (2015). Yeast as a cell factory: current state and perspectives, *Microbial Cell Factories* **14**(1).
URL: *http://www.microbialcellfactories.com/content/14/1/94*

Krogerus, K., Preiss, R. and Gibson, B. (2018). A Unique Saccharomyces cerevisiae × Saccharomyces uvarum Hybrid Isolated From Norwegian Farmhouse Beer: Characterization and Reconstruction, *Frontiers in Microbiology* **9**.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6165869/*

Legras, J.-L., Merdinoglu, D., Cornuet, J.-M. and Karst, F. (2007). Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history, *Molecular Ecology* **16**(10): 2091–2102.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2007.03266.x*

Libkind, D., Hittinger, C. T., Valério, E., Gonçalves, C., Dover, J., Johnston, M., Gonçalves, P. and Sampaio, J. P. (2011). Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast, *Proceedings of the National Academy of Sciences of the United States of America* **108**(35): 14539–14544.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3167505/*

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen Ba, A. N., Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R. and Louis, E. J. (2009). Population genomics of domestic and wild yeasts, *Nature* **458**(7236): 337–341.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2659681/*

Marullo, P., Bely, M., Masneuf-Pomarede, I., Aigle, M. and Dubourdieu, D. (2004). Inheritable nature of enological quantitative traits is demonstrated by meiotic segregation of industrial wine yeast strains, *FEMS yeast research* **4**(7): 711–719.

Marullo, P., Bely, M., Masneuf-Pomarède, I., Pons, M., Aigle, M. and Dubourdieu, D. (2006). Breeding strategies for combining fermentative qualities and reducing off-flavor production in a wine yeast model, *FEMS Yeast Research* **6**(2): 268–279.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1567-1364.2006.00034.x*

Marullo, P., Mansour, C., Dufour, M., Albertin, W., Sicard, D., Bely, M. and Dubourdieu, D. (2009). Genetic improvement of thermo-tolerance in wine Saccharomyces cerevisiae strains by a backcross approach, *FEMS yeast research* **9**(8): 1148–1160.

Masneuf-Pomarède, I., Le Jeune, C., Durrens, P., Lollier, M., Aigle, M. and Dubourdieu, D. (2007). Molecular typing of wine yeast strains Saccharomyces bayanus var. uvarum using microsatellite markers, *Systematic and Applied Microbiology* **30**(1): 75–82.

Mohd Azhar, S. H., Abdulla, R., Jambo, S. A., Marbawi, H., Gansau, J. A., Mohd Faik, A. A. and Rodrigues, K. F. (2017). Yeasts in sustainable bioethanol production: A review, *Biochemistry and Biophysics Reports* **10**: 52–61.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5637245/*

Naumov, G. I., Masneuf, I., Naumova, E. S., Aigle, M. and Dubourdieu, D. (2000). Association of Saccharomyces bayanus var. uvarum with some French wines: genetic analysis of yeast populations, *Research in Microbiology* **151**(8): 683–691.

Nielsen, J. (2013). Production of biopharmaceutical proteins by yeast, *Bioengineered* **4**(4): 207–211.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3728191/*

Payne, W. E. and Garrels, J. I. (1997). Yeast Protein Database (YPD): a database for the complete proteome of Saccharomyces cerevisiae, *Nucleic Acids Research* **25**(1): 57–62.
URL: *https://academic.oup.com/nar/article/25/1/57/1092483*

Petrizzelli, M., Vienne, D. d. and Dillmann, C. (2019). Decoupling the Variances of Heterosis and Inbreeding Effects Is Evidenced in Yeast's Life-History and Proteomic Traits, *Genetics* **211**(2): 741–756.
URL: *https://www.genetics.org/content/211/2/741*

Schacherer, J., Shapiro, J. A., Ruderfer, D. M. and Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of S. cerevisiae, *Nature* **458**(7236): 342–345.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2782482/*

Schluter Dolph, Price Trevor D., Rowe Locke and Grant Peter Raymond (1991). Conflicting selection pressures and life history trade-offs, *Proceedings of the Royal Society of London. Series B: Biological Sciences* **246**(1315): 11–17.
URL: *https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1991.0118*

Sicard, D. and Legras, J.-L. (2011). Bread, beer and wine: yeast domestication in the Saccharomyces sensu stricto complex, *Comptes Rendus Biologies* **334**(3): 229–236.

Spor, A., Kvitek, D. J., Nidelet, T., Martin, J., Legrand, J., Dillmann, C., Bourgais, A., Vienne, D. d., Sherlock, G. and Sicard, D. (2014). Phenotypic and Genotypic Convergences Are Influenced by Historical Contingency and Environment in Yeast, *Evolution* **68**(3): 772–790.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/evo.12302*

Spor, A., Nidelet, T., Simon, J., Bourgais, A., de Vienne, D. and Sicard, D. (2009). Niche-driven evolution of metabolic and life-history strategies in natural and domesticated populations of Saccharomyces cerevisiae, *BMC Evolutionary Biology* **9**(1): 296.
URL: *https://doi.org/10.1186/1471-2148-9-296*

Spor, A., Wang, S., Dillmann, C., Vienne, D. d. and Sicard, D. (2008). "Ant" and "Grasshopper" Life-History Strategies in Saccharomyces cerevisiae, *PLOS ONE* **3**(2): e1579.
URL: *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0001579*

Stearns, S. and Hoekstra, R. (2005). *Evolution*, second edition edn, Oxford University Press, Oxford, New York.

Chapter 3

Decoupling the variances of heterosis and inbreeding effects is evidenced in yeast's life-history and proteomic traits

Marianyela Petrizzelli[*], **Dominique de Vienne**[*] **and Christine Dillmann**[*,1]
[*]Génétique Quantitative et Evolution – Le Moulon, INRA, Université Paris-Saclay, Université Paris-Sud, CNRS, AgroParisTech, 91190 Gif-sur-Yvette, France

ABSTRACT Heterosis (hybrid vigor) and inbreeding depression, commonly considered as corollary phenomena, could nevertheless be decoupled under certain assumptions according to theoretical population genetics works. In order to explore this issue on real data, we analyzed the components of genetic variation in a population derived from a half-diallel cross between strains from *Saccharomyces cerevisiae* and *S. uvarum*, two related yeast species involved in alcoholic fermentation. A large number of phenotypic traits, either molecular (coming from quantitative proteomics) or related to fermentation and life-history, were measured during alcoholic fermentation. Because the parental strains were included in the design, we were able to distinguish between inbreeding effects, which measures phenotypic differences between inbred and hybrids, and heterosis, which measures phenotypic differences between a specific hybrid and the other hybrids sharing a common parent. The sources of phenotypic variation differed depending on the temperature, indicating the predominance of genotype by environment interactions. Decomposing the total genetic variance into variances of additive (intra- and inter-specific) effects, of inbreeding effects and of heterosis (intra- and inter-specific) effects, we showed that the distribution of variance components defined clear-cut groups of proteins and traits. Moreover, it was possible to cluster fermentation and life-history traits into most proteomic groups. Within groups, we observed positive, negative or null correlations between the variances of heterosis and inbreeding effects. To our knowledge, such a decoupling had never been experimentally demonstrated. This result suggests that, despite a common evolutionary history of individuals within a species, the different types of traits have been subject to different selective pressures.

KEYWORDS Hybrid vigor; inbreeding depression; diallel crossing; mixed effect genetic model

Heterosis, or hybrid vigor, refers to the common superiority of hybrids over their parents for quantitative traits. This phenomenon has been observed for virtually any quantitative trait, from mRNA abundances to fitness, and in a large diversity of species, including microorganisms. For decades it has been extensively studied and exploited for plant and animal breeding, since it affects traits of high economical interest such as biomass, fertility, growth rate, disease resistance etc. (Gowen 1952; Schnable and Springer 2013).

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: 1st February 2019
[1]To whom correspondence should be addressed: UMR Génétique Quantitative et Évolution - Le Moulon, Ferme du Moulon, 91190 Gif-sur-Yvette, France. E-mail: christine.dillmann@inra.fr

There are three classical, non exclusive genetic models to account for hybrid vigor: dominance, overdominance and epistasis. In the dominance model, the hybrid superiority results from the masking of the deleterious alleles of one parent by the non deleterious ones of the other parent (Davenport 1908). In the overdominance model, the hybrid superiority is due to the advantage *per se* of the heterozygous state at a given locus (Hull 1946). Actually, more common is pseudo-overdominance, which is due to dominance at two loci linked in repulsion, *e.g.* in maize (Graham *et al.* 1997; Lariepe *et al.* 2012) or yeast (Martì-Raga *et al.* 2017). Lastly, the epistasis model postulates favorable intergenic interactions created in the hybrids (Powers 1944). In particular, "less-than-additive" (antagonistic) epistasis, which is quite common in plant and animal species (Redden 1991; Shao *et al.* 2008) can account for best-parent heterosis (Fievet *et al.* 2010). In this

last paper, it is theoretically shown that epistasis can result in best-parent heterosis even if there is no dominance at any locus. The respective parts of the various genetics effects in heterosis depends on the trait, the species and the genetic material (Xiao *et al.* 1995; Huang *et al.* 2016; Seymour *et al.* 2016). Altogether, heterosis appears to be a pervasive phenomenon, accounted for by the common non-linearity of the genotype-phenotype map (Wright 1934; Omholt *et al.* 2000; Fiévet *et al.* 2018).

Because heterosis is associated with heterozygosity, heterosis for life-history traits is associated with genetic load: the average population fitness can never exceed the maximum fitness. Genetic load drives the evolution of sexual reproduction, of mating systems as well as the fate of small populations. Indeed, high levels of homozygosity in outcrossing species is generally associated with decreased growth rate, survival or fertility (discussed in Charlesworth and Willis (2009)). In population genetics, inbreeding depression is defined as the fitness of self-fertilized progenies as compared with fitness of outcrossing progenies. In sexual species, the balance between selfing and outcrossing is driven by the genetic load due to inbreeding depression relative to the cost of sexual reproduction (twice as expensive as clonal reproduction): selfing can evolve whenever inbreeding depression is less costly than the sexual reproduction, or after purging deleterious mutations as can arise in small populations (Lande and Schemske 1985). However, heterosis due to less-than-additive epistasis could explain the large number of predominantly (but not fully) selfing species exhibiting a persistent amount of inbreeding depression and heterosis (Charlesworth *et al.* 1991). Considering a metapopulation, Roze and Rousset (2004) defined inbreeding depression as the fitness reduction of selfed progeny relative to outcrossed progeny within populations, and heterosis as the difference between the fitness of the outcrossed progeny within population and the outcrossed progeny over the whole metapopulation. They showed that while selfing reduced both inbreeding depression and heterosis, inbreeding depression decreased and heterosis increased with the degree of subdivision of the metapopulation. Hence, from a population genetics point of view, heterosis is expected even in predominantly selfing species.

In a breeding perspective, the pioneer work of Shull (1908) in maize predicted that given the large amounts of heterosis within the species, the best way to maximize yield was to create inbreds from existing population varieties in order to seek for the best hybrid combinations. Diallel designs were popularized as the most comprehensive designs for estimating genetic effects, predicting hybrid values and generating breeding populations to be used as basis for selection and development of elite varieties (i.e. Hallauer and Filho (1988)). The simplest and most popular analytic decomposition of genetic effects in diallel designs is that of Griffing (1956), in which the mean phenotypic value, y_{ij}, of the cross between lines i and j is modeled as:

$$y_{ij} = \mu + GCA_i + GCA_j + SCA_{ij}, \quad (1)$$

where μ is the mean phenotypic value of the population, GCA_i (resp. GCA_j) is the *General Combining Ability* of line i (resp. j), *i.e.* the average performance of line i (resp. j) in hybrid combinations expressed as a deviation from the mean value of all crosses, and SCA_{ij} is the *Specific Combining Ability* of hybrid $i \times j$. It is defined as the difference between the mean phenotypic value of the progeny and the sum of the combining abilities of the parental lines (Sprague and Tatum 1942). Therefore, superior individuals can be selected from their GCA and/or SCA. Numerous extensions of the Griffing's model have been proposed to extract other effects, such as maternal and paternal effects or sex-linked variations (Cockerham and Weir 1977; Bulmer 1980; Zhu and Weir 1996; Greenberg *et al.* 2010). In many crop species, combining ability groups have been identified, with lines from the same group characterized by high specific combining ability with other groups (Hallauer *et al.* 1988). Generally, combining ability groups are redundant with population structure within a species (Melchinger and Gumber 1998; Ramya *et al.* 2018), which is consistent with the population genetics predictions of Roze and Rousset (2004).

When parental lines are included in the analysis, GCA and SCA effects can be decomposed in more suitable genetic effects. Indeed, the value of a particular hybrid can be compared either to the average value of its inbred parents, or to the average value of the other hybrids sharing either parent. Heterosis can be split into *average heterosis* (average difference between inbreds and outbreds), *variety heterosis* (average difference between one inbred parent and all crosses sharing the same parents), and *specific heterosis* (difference between the hybrid and all hybrids sharing at least one parent) (Eberhart and Gardner 1966). A modern version of this model have been proposed by Lenarcic *et al.* (2012) along with a Bayesian framework to estimate the genetic effects.

In this work, we study a half-diallel design with diagonal constructed from the crosses between 11 yeast strains belonging to two close species, *Saccharomyces cerevisiae* and *S. uvarum*. The design included both intra- and inter-specific crosses. Two categories of phenotypic traits were considered: (i) protein abundances measured at one time point of alcoholic fermentation (Blein-Nicolas *et al.* 2013, 2015); (ii) a set of fermentation traits measured during and/or at the end of fermentation, which were divided into kinetic parameters, basic enological parameters, aromas and life-history traits (da Silva *et al.* 2015). All traits were independently measured at two temperatures.

We propose a decomposition of the genetic effects based on Lenarcic *et al.* (2012) that takes into account the presence of two species in the diallel design and that distinguishes heterosis and inbreeding effects. We could characterize every trait by the set of its variance components and we could clearly cluster the traits from this criterion, which suggests that traits sharing a similar pattern of variance components could share common life-history. We were able to assign each fermentation trait to one group of protein traits, which shows that integrated phenotypes and proteins can share similar life-history. Finally, our results show a poor correlation between the variances of heterosis and inbreeding effects within groups. This confirms the importance of epistatic interactions in determining the components of phenotypic variation both within and between close species. Altogether, our results suggest that despite a common demographic history of individuals within a species, the genetic variance components of the traits can be used to trace back other trait-specific evolutionary pressures, like selection.

Materials and Methods

Materials

The genetic material of the experimental design consisted in 7 strains of *S. cerevisiae* and 4 strains of *S. uvarum* associated to various food-processes (enology, brewery, cider fermentation and distillery) or isolated from natural environment (oak exudates). These strains, called W1, D1, D2, E2, E3, E4, E5 for *S. cerevisiae* and U1, U2, U3, U4 for *S. uvarum* could not be used

as such as parents of a diallel design because they were suspected to be heterozygous at many loci. Monosporic clones were isolated from each of these strains using a micromanipulator (Singer MSM Manual; Singer Instrument, Somerset, United Kingdom), as indicated in da Silva *et al.* (2015). All strains but D2 were homothallic (HO/HO), therefore fully homozygous diploid strains were spontaneously obtained by fusion of opposite mating type cells. For D2 (ho/ho), the isolated haploid meiospore were diploidized via transient expression of the HO endonuclease (Albertin *et al.* 2009). The derived fully homozygous and diploid strains were used as the parental strains of a half-diallel design with diagonal, *i.e.* including the inbred lines. The parental lines were selfed and pairwise crossed, which resulted in a total of 66 strains: 11 inbred lines, 27 intra-specific hybrids (21 for *S. cerevisiae*, noted *S. c.*, and 6 for *S. uvarum*, noted *S. u.*) and 28 inter-specific (noted *S. u.* × *S. c*). For each hybrid construction, parental strains of opposite mating type were put in contact for 2 to 6 hours in YPD medium at room temperature, and then plated on YPD-agar containing the appropriate antibiotics. The nuclear and mitochondrial stability of the hybrids was checked after recurrent cultures on YPD-agar corresponding to ≈ 80 generations (see details in Albertin *et al.* (2013a)). In addition, for each of the 28 interspecific hybrids, both parental sets of more than 600 proteins were detected in a proteomic approach Blein-Nicolas *et al.* (2015), with no evidence of hybrid instability.

The 66 strains were grown in triplicate in fermentors at two temperatures, 26° and 18°, in a medium close to enological conditions (Sauvignon blanc grape juice) (da Silva *et al.* 2015). From a total of 396 alcoholic fermentations (66 strains × 2 temperatures × 3 replicas), 31 failed due to poor fermenting abilities of some strains. The design was implemented considering a block as two sets of 27 fermentations (26 plus a control without yeast to check for contamination), one carried out at 26° and the other at 18°. The distribution of the strains in the block design was randomized to minimize the residual variance of the estimators of the strain and temperature effects, as described in Albertin *et al.* (2013b).

For each alcoholic fermentation, two types of phenotypic traits were measured or estimated from sophisticated data adjustment models: 35 fermentation traits and 615 protein abundances.

The fermentation traits were classified into four categories (da Silva *et al.* 2015):

- *Kinetics parameters*, computed from the CO_2 release curve modeled as a Weibull function fitted on CO_2 release quantification monitored by weight loss of bioreactors: the fermentation lag-phase, *t-lag* (h); the time to reach the inflection point out of the fermentation lag-phase, *t-*V_{max} (h); the fermentation time at which 45 gL^{-1} and 75 gL^{-1} of CO_2 was released, out of the fermentation lag-phase, *t-45* (h) and *t-75* (h) respectively; the time between *t-lag* and the time at which the CO_2 emission rate became less than, or equal to, $0.05 gL^{-1}h^{-1}$, *AFtime* (h); the maximum CO_2 release rate, $V_{\max}$ ($gL^{-1}h^{-1}$); and the total amount of CO_2 released at the end of the fermentation, $CO_{2\max}$ (gL^{-1}).
- *Life history traits*, estimated and computed from the cell concentration curves over time, modeled from population growth, cell size and viability quantified by flow cytometry analysis: the growth lag-phase, *t-*$N_0(h)$; the carrying capacity, *K* (log[cells/mL]); the time at which the carrying capacity was reached, *t-*N_{max} (h); the intrinsic growth rate, *r* (log[cell division/mL/h]); the maximum value of the estimated CO_2 production rate divided by the estimated cell concentration, J_{max} ($gh^{-1}10^{-8}cell^{-1}$); the average cell size at *t-*N_{max}, *Size-t-*N_{max}(μm); the percentage of living cells at *t-*N_{max}, *Viability-t-*N_{max} (%); and the percentage of living cells at *t-75*, *Viability-t-75* (%).
- *Basic enological parameters*, quantified at the end of fermentation: *Residual Sugar* (gL^{-1}); *Ethanol* (%vol); the ratio between the amount of metabolized sugar and the amount of released ethanol, *Sugar.Ethanol.Yield* (gL^{-1}%vol^{-1}); *Acetic acid* (gL^{-1} of H_2SO_4); *Total* SO_2 (mgL^{-1}) and *Free* SO_2 (mgL^{-1}).
- *Aromatic traits*, mainly volatile compounds measured at the end of alcoholic fermentation by GC-MS: two higher alcohols (*Phenyl-2-ethanol* and *Hexanol*, mgL^{-1}); seven esters (*Phenyl-2-ethanol acetate*, *Isoamyl acetate*, *Ethyl-propanoate*, *Ethyl-butanoate*, *Ethyl-hexanoate*, *Ethyl-octanoate* and *Ethyl-decanoate*, mgL^{-1}); three medium chain fatty acids (*Hexanoic acid*, *Octanoic acid* and *Decanoic acid*, mgL^{-1}); one thiol *4-methyl-4-mercaptopentan-2-one*, *X4MMP*(mgL^{-1}) and the acetylation rate of higher alcohols, *Acetate ratio*.

For proteomic analyses the samples were harvested at 40 % of CO_2 release, corresponding to the maximum rate of CO_2 release. Protein abundances were measured by LC-MS/MS techniques from both shared and proteotypic peptides relying on original Bayesian developments (Blein-Nicolas *et al.* 2012). A total of 615 proteins were quantified in more than 122 strains × temperature combinations as explained in details in Blein-Nicolas *et al.* (2015).

Cross-referencing MIPS micro-organism protein classification (Ruepp *et al.* 2004), KEGG pathway classification (Kanehisa and Goto 2000; Kanehisa *et al.* 2016, 2017) and Saccharomyces Genome database (Cherry *et al.* 2012), we attributed each protein to a single functional category based on our expert knowledge (Table ST1). Considering the genes encoding the proteins, we also assigned to each protein a number of putative transcription factors (TFs). A total of 313 TFs with a consensus DNA-binding sequence were retrieved from the Yeastrack database (Teixeira *et al.* 2014; Abdulrehman *et al.* 2011; Monteiro *et al.* 2008; Teixeira *et al.* 2006).

Statistical Methods

In order to estimate the genetic variance components for the different phenotypic traits, we adapted the model described in Lenarcic *et al.* (2012) to our particular half-diallel design that includes the diagonal with parental inbred strains from two species. Thus we included in our model intra- and inter-specific additive effects, inbreeding effects and intra- and inter-specific heterosis effects.

Formally, let y_{ijk} be the observed phenotype for the cross between parents i and j in replica k. Our model reads:

$$\begin{aligned} y_{ijk} = \mu + I_{s(i)=s(j)}(A_{w_i} + A_{w_j}) + I_{s(i)\neq s(j)}(A_{b_i} + A_{b_j}) + \\ + I_{i\neq j}(I_{s(i)=s(j)}H_{w_{ij}} + I_{s(i)\neq s(j)}H_{b_{ij}}) + \\ + I_{i=j}(\beta_{s(i)} + B_i) + \epsilon_{ijk}, \end{aligned} \quad (2)$$

where:

- μ is the overall mean;
- $s(i)$ associates to each parental strain i the specie it belongs to:
$$s(i) \in \{S.\ cerevisiae, S.\ uvarum\}$$
- A_{w_i} and A_{b_i} denote, respectively, the additive contributions of strain i in intra-specific (within species, *i.e.* $s(i) = s(j)$), and inter-specific (between species, *i.e.* $s(i) \neq s(j)$) crosses;

- $H_{w_{ij}}$ and $H_{b_{ij}}$ denote the interaction effect between parents (i, j) in intra-specific (within species) and inter-specific (between species) crosses, respectively. Due to our half-diallel design (no reciprocal crosses), they are assumed to be symmetric, *i.e.* $H_{w_{ij}} = H_{w_{ji}}$ and $H_{b_{ij}} = H_{b_{ji}}$. Hereafter we will refer to these effects as intra- and inter-specific heterosis effects, respectively;
- $\beta_{s(i)}$ and B_i are, respectively, the deviation from the fixed overall effect for the species $s(i)$ and the associated strain-specific contribution of strain i in the case of inbred lines. Hereafter we will refer to B_i as inbreeding effect;
- ϵ_{ijk} is the residual, the specific deviation of individual ijk;
- $I_{condition}$ is an indicator variable. Its value is equal to 1 if the condition is satisfied and 0 otherwise.

Therefore, for the parental lines we have:

$$y^{p}_{iik} = \mu + 2A_{w_i} + \beta_{s(i)} + B_i + \epsilon_{iik}, \quad (3)$$

for the intra-specific hybrids:

$$y^{intra}_{ijk} = \mu + A_{w_i} + A_{w_j} + H_{w_{ij}} + \epsilon_{ijk}, \quad (4)$$

and for the inter-specific hybrids:

$$y^{inter}_{ijk} = \mu + A_{b_i} + A_{b_j} + H_{b_{ij}} + \epsilon_{ijk}. \quad (5)$$

All genetic effects were considered as random variables drawn from a normal distribution. Formally, letting $\boldsymbol{q} \in \{\boldsymbol{A_w}, \boldsymbol{A_b}, \boldsymbol{B}, \boldsymbol{H_w}, \boldsymbol{H_b}\}$ denote the genetic effect under consideration:

$$\forall i \quad q_i \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{q}}). \quad (6)$$

The full mixed-effect genetic model is thus defined by three fixed effects (the intercept μ and the inbreeding effects β_{Su} and β_{Sc}) and five genetic random effect variances ($\sigma^2_{\boldsymbol{A_w}}$, $\sigma^2_{\boldsymbol{A_b}}$, $\sigma^2_{\boldsymbol{B}}$, $\sigma^2_{\boldsymbol{H_w}}$, $\sigma^2_{\boldsymbol{H_b}}$).

We did not declare mitochondrial effects because many genes encoding mitochondrial proteins are repressed under fermentation conditions, and because inter-specific hybrids harbor similar fermentation features for most fermentation kinetics and enological parameters whatever their mitochondrial genotype (Albertin *et al.* 2013a). In addition, we did not know the mitochondrial inheritance for most of the intra-specific crosses (table ST3).

The fitting algorithm

Fixed effects, variance components of the genetic effects as well as their Best Linear Unbiased Predictors (*BLUPs*) were estimated using the *hglm* package in R (Ronnegard *et al.* 2010) that implements the estimation algorithm for hierarchical generalized linear models and allows fitting correlated random effects as well as random regression models by explicitly specifying the design matrices both for the fixed and random effects. The model, based on a maximum likelihood estimation, is deemed to produce unbiased statistics (Gumedze and Dunne 2011).

A separate analysis was conducted for each trait at each temperature, considering the vector of observations for the trait/temperature combination of interest, $\mathbf{y}$, and re-writing model (eq. (2)) in matrix form:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon}, \quad (7)$$

where X is the design matrix for the fixed effects, Z the design matrix for the random effects, $\boldsymbol{\beta} = (\mu, \beta_{S.u.}, \beta_{S.c.})$ and $\boldsymbol{u} = (\boldsymbol{A_w}, \boldsymbol{A_b}, \boldsymbol{B}, \boldsymbol{H_w}, \boldsymbol{H_b})$ are respectively the vectors of fixed effect parameters and random effect parameters, and $\boldsymbol{\epsilon}$ is the vector of residual errors. With this notation, the construction of the model is straightforward from the data (for details see The fitting algorithm in Supplementary Materials).

Whenever the full model (eq. 2) failed to converge, we considered the subsequent model obtained by removing one effect at a time following the hierarchy imposed by the order of the fitting algorithm, *i.e.* first heterosis, second inbreeding effects and finally additive effects. The full model converged for all proteomic data. For the fermentation traits, the model did not converge for most of the *Ethyl* esters (*Ethyl-propanoate*, *Ethyl-butanoate*, *Ethyl-hexanoate*, *Ethyl-octanoate* and *Ethyl-decanoate*), as well as for *Acetate Ratio* and for *Acetic acid* that were removed from the analysis. For all other fermentation traits, the full model converged, except for *t.lag* at 18°, for which the additive model applied. For this trait, other genetic variance components were set to zero.

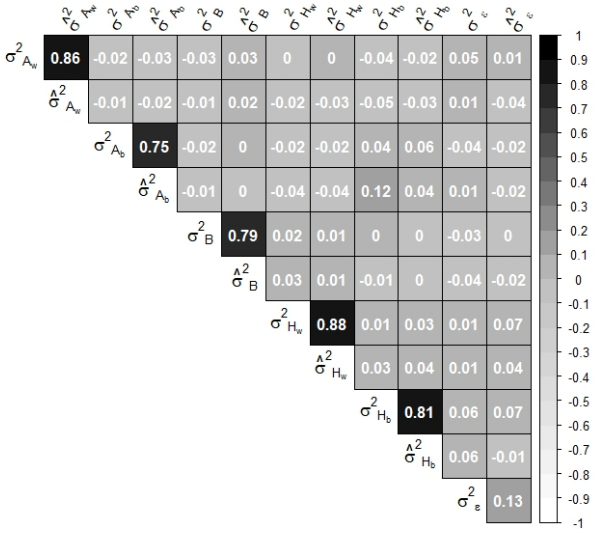


Figure 1 Correlation between estimated variance components and their true value. Variances have been estimated on a simulated half-diallel between 11 parental strain (seven belonging to a specie, four to the other). Phenotypic values have been computed as detailed in section Testing for the reliability of the model.

In order to test the robustness of the results, a bootstrap analysis was performed by sampling the 55 hybrids with replacement, conditionally to the 11 parental strains. Each bootstrap sample was submitted to the same analysis as described above. For each variance component, we checked that the estimations in the experimental sample were close to the median of the estimations in the bootstrap samples.

Testing for the reliability of the model

Computer simulations were performed to test the statistical power of the *hglm* algorithm in predicting the values of the observables while producing unbiased estimations of the model parameters. We simulated a half-diallel between 11 strains, seven belonging to a species, four to the other. We computed the phenotypic values of each simulated cross by first drawing μ, $\beta_{specie1}$, $\beta_{specie2}$, $\sigma^2_{\boldsymbol{A_w}}$, $\sigma^2_{\boldsymbol{A_b}}$, $\sigma^2_{\boldsymbol{B}}$, $\sigma^2_{\boldsymbol{H_w}}$, $\sigma^2_{\boldsymbol{H_b}}$ and $\sigma^2_{\boldsymbol{\epsilon}}$ from a Gamma distribution fitted from the values estimated by the

model on our dataset (see fig. SF1). Second, for each random effect $\boldsymbol{q} \in \{\boldsymbol{A_w}, \boldsymbol{A_b}, \boldsymbol{B}, \boldsymbol{H_w}, \boldsymbol{H_b}, \boldsymbol{\epsilon}\}$ we drew

$$\forall i \quad q_i \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{q}}) \tag{8}$$

and computed the phenotypic values as in eq. 2, generating three replicas per cross.

We repeated the simulation 1000 times. We fitted the model and checked that the estimation of the random effects, the predicted phenotypic values as well as their variance components were close enough to the true values (fig. 1) and we noticed that inbreeding parameters were the most variable (fig. SF2 in *Supplementary figures*).

In addition, since we were interested in the correlation structure between the variance components of the genetic effects, we checked that possible correlations between random effects were not a statistical artifact of the model. Therefore, we simulated uncorrelated variances of random effects and we checked that no correlation structure was found between the estimated variance components, as can be seen in fig. 1. Simulations performed with different numbers of parental lines led to similar results (not shown).

Fermentation traits

Before fitting our model, we updated eq. 2 in order to account for a block effect:

$$y_{ijkl} = y_{ijk} + block_l + \epsilon_{ijkl}, \tag{9}$$

assuming that

$$\forall l \quad block_l \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{block}}). \tag{10}$$

Many fermentation traits, mostly aromatic, were *log*-transformed in order to deal with the variable mean of the residuals. So as to handle the null values in the observations, we chose to consider the following transformation:

$$y_{ijk} = log(max(y_{ijk}, \delta)) \tag{11}$$

where $\delta \sim \mathcal{U}(0, min(\boldsymbol{y}))$. In this situation, as we introduced a random term in our analysis, which may skew parameter estimation, we decided to: (i) perform the *log*-transformation, (ii) compute the fitting algorithm, (iii) record the parameter's estimation, then after having computed it a hundred times, (iv) consider the median of the estimators in order to achieve a more robust statistics.

Protein abundances

For each cross, protein abundances have been quantified on average. Yet, to perform a diallel analysis at the proteomic level, replicas are critical for quantifying genetic variation. Therefore, we generated pseudo replicas using the residual variance estimated when quantifying protein abundances (Blein-Nicolas *et al.* 2013). Formally, let y_{ij} be the average protein abundance of the cross between parents i and j. We generated three replicas as follows:

$$y_{ijk} = y_{ij} + \epsilon_k \tag{12}$$

$$\epsilon_k \sim \mathcal{N}(0, \hat{\sigma}^2_\epsilon) \text{ for } k = 1, 2, 3 \tag{13}$$

where $\hat{\sigma}^2_\epsilon$ is the residual variance. Simulations of pseudo replicas and parameter estimations were performed 100 times. The final value of the parameters was the median of its estimation.

Variance component analysis

For each trait, our mixed model generates a vector of variance components

$$\boldsymbol{v} = (\hat{\sigma}^2_{\boldsymbol{A_w}}, \hat{\sigma}^2_{\boldsymbol{A_b}}, \hat{\sigma}^2_{\boldsymbol{B}}, \hat{\sigma}^2_{\boldsymbol{H_w}}, \hat{\sigma}^2_{\boldsymbol{H_b}}) \tag{14}$$

and the results were summarized in a matrix with rows being the different trait by temperature combinations, and columns the relative contribution of each component to the total genetic variance of the trait. We chose to perform unsupervised classification to compare the distributions of variance components between traits. Following the recommendations of Kurtz *et al.* (2015), percentages of variance components were transformed into real numbers using the following *clr*-transformation:

$$clr(\hat{\sigma}^2_{\boldsymbol{q}}) = log(\frac{\hat{\sigma}^2_{\boldsymbol{q}}}{(\prod_{k \in \boldsymbol{Q}} \hat{\sigma}^2_{\boldsymbol{k}})^{1/N_q}}) \tag{15}$$

where N_q is the total number of random effects and $\boldsymbol{Q}$ is the set of random variables fitted by the model. For fermentation traits, $N_q = 7$ (accounting for block and residual variances, eq. 9), while $N_q = 6$ for proteomic traits (eq. 2). We chose the *clr*-transformation because it satisfies *scale invariance*, *sub-compositional dominance* and *perturbation invariance* properties (Tsagris *et al.* 2011). Therefore the distance relationship between the original profiles is preserved by the selected sub-vectors thanks to the sub-compositional dominance property of the *clr*-transformation (see section Subcompositional dominance and distances in Supplementary Materials). The *clr*-transformation allowed us to test finite Gaussian mixture models using model-based clustering proposed in the *Mclust* package in R (Scrucca *et al.* 2016). Percentage of good assignments were computed by separating the data into training and validation sets.

This procedure was first applied separately for proteomic and fermentation traits (see Structuration of genetic variability at the fermentation trait level in Supplementary Materials). Protein groups were tested for enrichment in either Kegg pathways, transcription factors and heterotic proteins. Fermentation traits were tested for enrichment in the different trait categories (kinetic parameters, life-history, basic enological parameters, aromatic traits). For each cluster, Pearson's chi-square test of enrichment was computed on protein functional category frequencies taking as prior probability the expected categorical frequency found in the MIPS database.

Further, fermentation traits were assigned to clusters identified on protein abundances profiles based on their membership probability computed through Gaussian finite mixture models.

Data Availability

The data that support the findings of the current study are available at figshare DOI:10.6084/m9.figshare.7378412. Supplementary materials contain:

- Demonstration of the relationship between the *subcompositional dominance* property and distances in the Euclidean space;
- Detailed description of the fitting algorithm;
- Description of the construction of the simulated values on a half-diallel design based on the genetic models supposed to explain heterosis and inbreeding;
- Demonstration of the equality between the variances of heterosis and inbreeding effects in three parents half-diallel designs with no maternal effects;
- Clustering analysis for the fermentation and life-history traits;

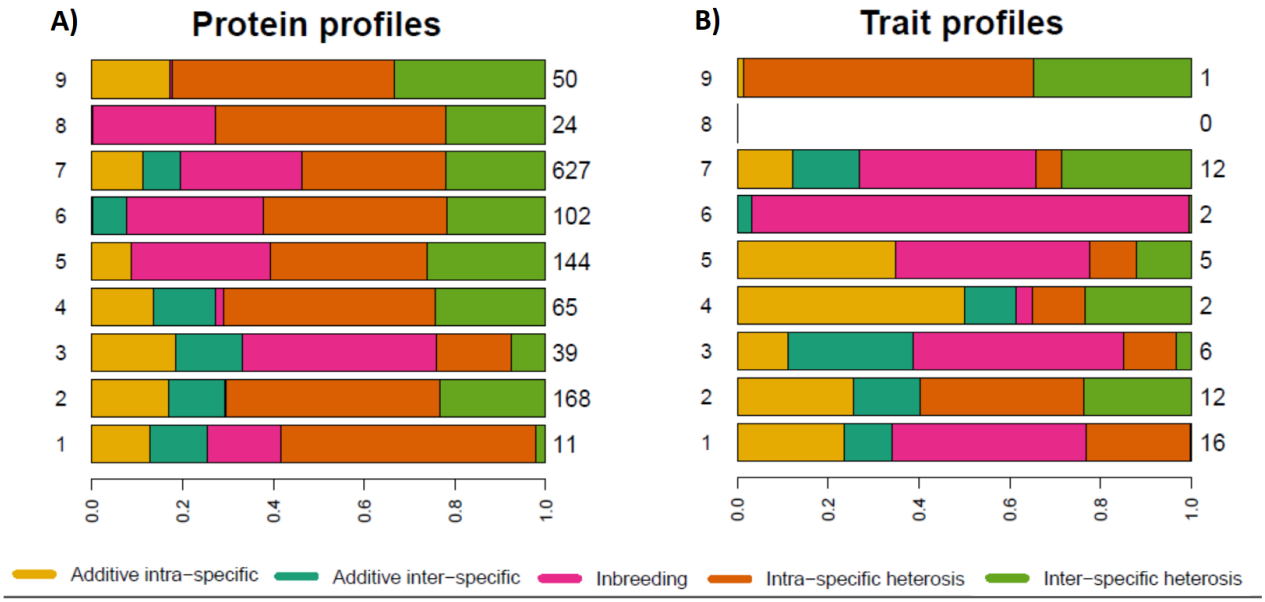


Figure 2 Clustering profiles of genetic variance components for protein abundances (A) against profiles of fermentation traits (B) predicted in each cluster. Cluster numbers are reported on the left, on the right the number of proteins or traits found in each cluster.

- Strains characterization based on the estimated BLUP of their genetic effects;
- Table ST1: Protein functional category classification (available at figshare DOI:10.6084/m9.figshare.6683666);
- Table ST2: Raw values of genetic variances and broad sense heritability (BSH) estimated and analyzed in this study for protein abundances, and fermentation and life-history traits (available at figshare DOI:10.6084/m9.figshare.7128152);
- Table ST3: Mitochondrial inheritance of the phenotyped crosses of our study;
- Table ST4: Table of results from the Pearson's chi-square test of cluster enrichment in proteins with a particular functional category;
- Figure SF1: Density distribution of the genetic variances estimated by the model;
- Figure SF2: Predicted BLUPs and phenotypic values versus their prior value used to compute the values of simulated diallels;
- Figure SF3: Clustering profiles of fermentation and life-history traits;
- Figure SF4: Global correlations of the genetic variance components for both protein abundances and the more integrated traits;
- Figure SF5: Representation of the standardized Pearson's chi-square residuals of each cluster computed at 18° versus those at 26° estimated for the analysis of cluster enrichment in proteins with a particular functional category;
- Figure SF6: Correlation plot between genetic effects of fermentation and life-history trait profiles;
- Figure SF7: Intra-cluster correlations of variance components profiles for fermentation and life-history traits;
- Figure SF8: Variance components of fermentation and life-history traits at the two temperatures;
- Figure SF9: Summary example of the density distribution of a genetic variance estimation through bootstrap analysis;
- Figure SF10: Representation of the relationship between the variances of heterosis and inbreeding effects simulated through different genetic models;
- Figure SF11: For each trait and for each genetic effect are shown the strains with highest and lowest contribution at both temperatures;
- Figure SF12: For each trait are shown the estimated BLUPs of each genetic parameter.

Results

In order to estimate genetic variance components from a diallel cross involving two yeast species, we proposed a decomposition of genetic effects based on the model of Lenarcic *et al.* (2012) that allowed to split the classical General (GCA) and Specific (SCA) Combining Abilities into intra- and inter-specific additive and heterosis effects, and to take into account inbreeding effects, defined as the difference between the inbred line value and the average value of all the crosses that have this inbred as parent.

Simulations showed that despite the small number of parents in the diallel, our model led to unbiased estimations of variance components, and that correlations between variance components did not arise from unidentifiability of some model's parameter (fig. 1). Significance of variance components was assessed by bootstrap sampling. We found that whenever the fitting algorithm converged, variance component estimations were significant. For some traits and some variance components, the bootstrap distributions of the estimated variances were bimodal, suggesting a strong influence of a particular hybrid combination. However, the estimates were globally closed to the median of the bootstrap distribution (see example fig. SF9). Therefore, we are confident with our estimations, conditionally to the parents of the diallel.

Table 1 Pearson's chi-square test for count data: comparison between the number of heterotic proteins in each cluster and group membership probability. The statistics clearly highlight clusters enriched of heterotic proteins (p-value<0.05).

Cluster	1	2	3	4	5	6	7	8	9
Number of proteins	11	168	39	65	144	102	627	24	50
Number of heterotic proteins	7	35	3	22	13	13	72	5	2
Proportion of heterotic proteins	0.64	0.21	0.08	0.34	0.09	0.13	0.11	0.21	0.04
Chi-square standardized residuals	4.42	2.56	–1.07	4.40	–1.69	–0.35	–2.39	0.91	1.93

Because temperature has a major effect on many traits and because, in previous work, numerous strain × temperature effects have been detected (da Silva *et al.* 2015; Blein-Nicolas *et al.* 2015), the model was applied to each trait separately at the two temperatures. We obtained estimations of fixed and random effect parameters, their corresponding variances, residuals and residual variances. For each trait, normality of residuals and homogeneity of variances was checked. Broad sense heritability was measured as the ratio of the sum of genetic variance components to the total phenotypic variance. It varied between 0.05 to 0.98 for protein abundances and between 0.04 to 0.95 for fermentation traits. Altogether, protein abundance measurements were highly repeatable (median heritability of 0.53), while fermentation traits were more variable. Median broad sense heritability was 0.77 for fermentation kinetic trait, 0.49 for life-history traits, 0.36 for basic enological products and 0.32 for aromatic traits. Whatever the amount of residual variance, all genetic variance components were significant for all traits, except for *t.lag* at 18°, for which only the variances of additive effects were significant. We found that variances associated to each genetic effect differ in a large extent between the two temperatures (shown for fermentation traits in fig. SF12).

Because of their potential interest for wine-making, BLUPs of fermentation traits are presented in section Strain characterization of Supplementary Materials. In the following, we focus on genetic variance components.

Structuration of genetic variance components at the proteomic level

A Gaussian mixture model was used to classify the proteins according to their genetic variance components. The best model clearly identified nine clusters, each characterized by a particular profile of genetic variance components (fig. 2). Cluster 1 (88.4% of good assignments) consists of 11 proteins that have high variance of intra-specific heterosis effects and the smallest variance of inter-specific heterosis effects. Clusters 2, 4 and 9 have a very small variance of inbreeding effects. Clusters 2 and 4 differ from cluster 9 by their significant variance of inter-specific additive effects. 6.4% of proteins from cluster 2 (composed of 168 proteins with 93.2% of good assignments) can be attributed to cluster 4 and 10.4% of proteins from cluster 4 (65 proteins, 80.5% good assignments) to cluster 2. Proteins from clusters 3 (80.5% of good assignments) and 7 (93.3% of good assignments) have similar profiles.

Indeed, 19.5% of the proteins from cluster 3 can be attributed to cluster 7 and 4% of the proteins from cluster 7 can be attributed to cluster 3. Cluster 3 consists in 39 proteins with relatively higher variance of additive and inbreeding effects. Cluster 7 has 627 proteins with higher variance of heterosis effects. Proteins from cluster 5 (144 proteins, 96% of good assignments) have significant variance of intra-specific additive effects but null variance of inter-specific additive effects and high heterosis and inbreeding effects variances. On the contrary, cluster 6 (102 proteins, 96.2% of good assignments) has null variance of intra-specific additive effects, small variance of additive inter-specific effects, and high variance of heterosis and inbreeding effects. Cluster 8 (96.9% of good assignments) consists of 24 proteins that have null variances of additive effects and high variances of heterosis and inbreeding effects. Finally, the 50 proteins in cluster 9 (95.4% of good assignments) are characterized by a null variance of additive inter-specific and inbreeding effects and high variance of intra-specific and inter-specific heterosis effects. Overall the same protein is generally found in two different clusters at the two temperatures (only 37% of proteins belong to the same cluster at the two temperatures).

The nine clusters were also clearly distinguishable from each other from their pattern of correlation between variance components (fig. 3). Globally, all variance components are negatively correlated, except for the variances of heterosis effects, $\sigma^2_{H_w}$ and $\sigma^2_{H_b}$, that are positively correlated ($r = 0.47$, fig. SF4).

Therefore, we can state that the 615 proteins at 18° and 26° form highly structured and well defined clusters according to their genetic variance component profiles.

Proteins sharing a similar variance component structure share functional properties

In each protein cluster we tested for enrichment in functional categories at the two temperatures separately. Clusters were split into two groups of proteins, those measured at 18° and those measured at 26°, and the enrichment analysis was performed for each group. The statistical tests were significant for each cluster, except for cluster 1 at 18° and cluster 6 at 26° (tab. ST4). Even though one protein generally falls into two different clusters at two different temperatures, functional enrichments were globally the same at the two temperatures. Indeed, we found a high correlation between Pearson's chi-squared residuals at both temperatures, except for clusters 3 and 9 (fig. SF5). Whenever a functional category was enriched/depleted at one temperature, it also tended to be enriched/depleted at the other temperature.

Cluster 1 is enriched with proteins quantified at 26° linked to response to stress, mating and transcription, and depleted with proteins related to cell fate and protein synthesis. Cluster 3 is enriched with proteins measured at 18° linked to amino-acid and nucleotide metabolism, and at 26° to cell fate and response to stress. Cluster 6 is enriched with proteins quantified at 18° linked to protein synthesis and nucleotide metabolism, and depleted in proteins linked to metabolism, other than amino acid, nucleotide and carbon metabolism. Cluster 9 is enriched in proteins linked to transcription at both temperatures, it is enriched in proteins measured at 18° linked to response to stress and

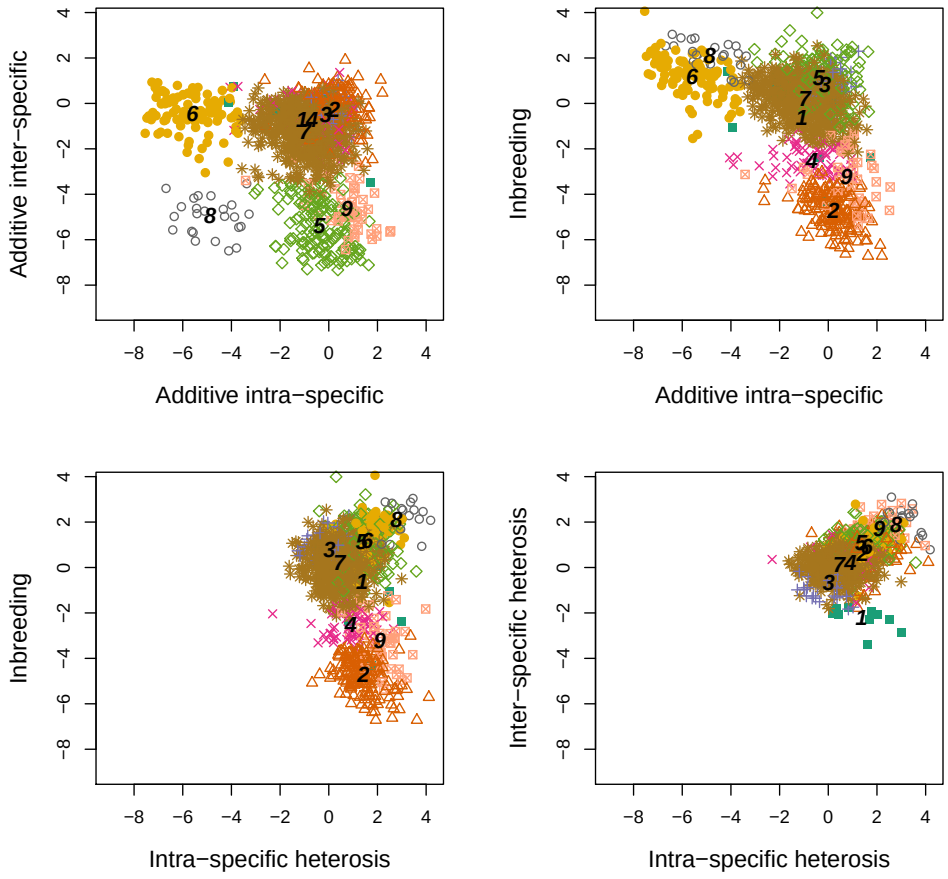


Figure 3 Patterns of correlations between genetic variance components of protein abundances. Points correspond to proteins, type and color combinations identify the clusters obtained by their classification based on a Gaussian Mixture model. Numbers from 1 to 9 identify class centers for each cluster.

mating, and depleted in proteins linked to protein synthesis and cell fate; at 26° it is enriched in proteins linked to nucleotide metabolism and transport. The other protein clusters have the same profile at both temperatures. Cluster 2 is enriched with proteins linked to amino-acids and carbon metabolism, cell fate and response to stress, and depleted in proteins linked to transport and mating. Cluster 4 is enriched in proteins linked to amino-acid metabolism, and to stress response at 26°. Cluster 5 is enriched in proteins linked to protein synthesis, amino-acid, nucleotide and other but not carbon metabolism, and depleted in proteins linked to transcription. Cluster 7 is enriched in proteins linked to amino-acids and carbon metabolism, and depleted in proteins linked to transcription, transport and signal. Cluster 8 is enriched in proteins linked to cell fate, stress response, nucleotide metabolism and mating, and depleted in proteins linked to other metabolisms, transport and protein synthesis. Hence, genetic variance components tend to cluster proteins having similar functions at both temperatures.

Concerning the number of transcription factors, we found no correlation between the number of transcription factors and the components of genetic variation of protein abundances.

Finally, Pearson's chi-square test have been performed in order to investigate if there were differences between clusters regarding the proportion of heterotic proteins quantified in Blein-Nicolas *et al.* (2015). Results are shown in tab. 1: cluster 1, 2, 4 are enriched with heterotic proteins while in clusters 5, 7, 9 heterotic proteins are scarce ($\chi^2 = 54.29$, p-value<0.05). Hence, heterotic proteins are preferably found in clusters characterized by low variance of inbreeding effects and high variances of intra-specific and inter-specific heterosis effects.

Briefly, despite poor correlations between variance components measured for the same protein at two temperatures, the nine clusters of proteins identified from the distribution of variance components group together proteins of similar function, based on their functional annotation. Heterotic proteins that show non-additive inheritance between parents and hybrids are mostly found in protein clusters with high variances of intra-specific and inter-specific heterosis effects and low variance of inbreeding effects.

Variance components of fermentation traits fall into the proteomic landscape

Using for the fermentation/life history traits the same clustering approach as for the proteins, we clearly identified three profiles of genetic variance components (fig. SF3; see description in the section Structuration of genetic variability at the fermentation

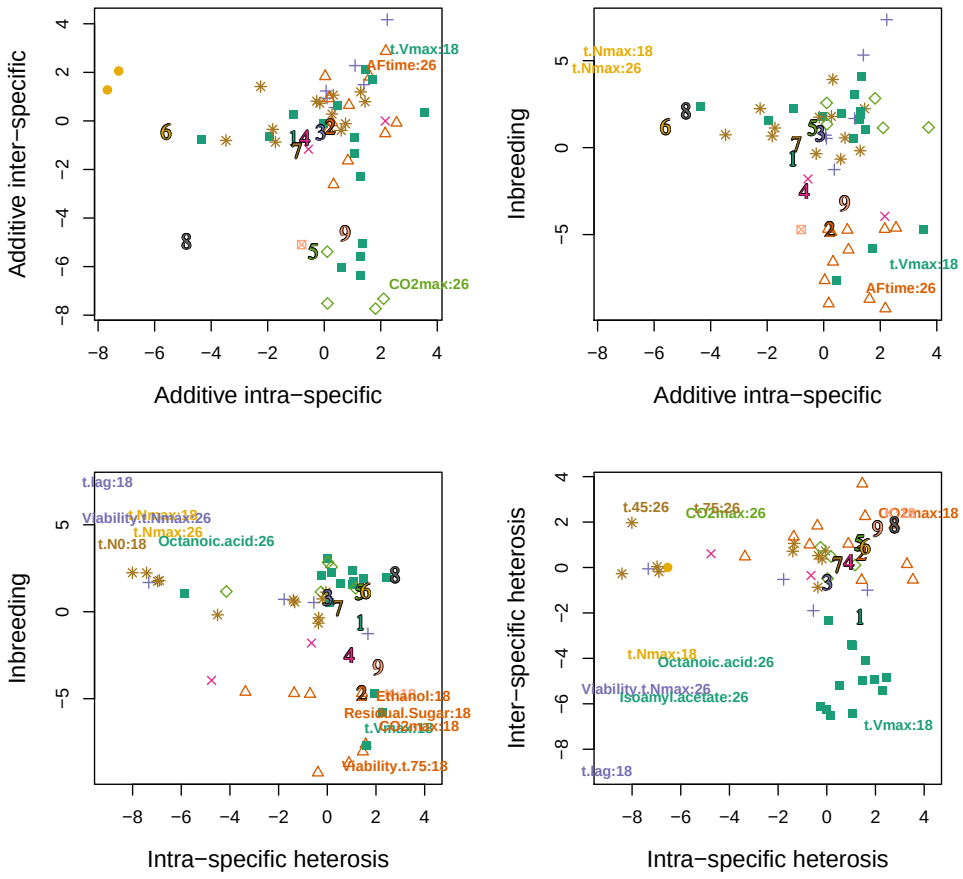


Figure 4 Variance components of fermentation traits. Fermentation traits are assigned to clusters identified at the proteomic level based on their membership probability computed through Gaussian finite mixture models. They are identified by the type and color combination of the cluster to which they are assigned. Numbers 1 to 9 identify class centers for each protein cluster. Labels are only given for outlier traits, *i.e.* those that do not belong to the 95% confidence interval of the genetic variance estimates of protein abundances on the plotted direction.

trait level of Supplementary Materials).

In order to compare the patterns of genetic variation of protein abundances and fermentation traits, we tried to assign fermentation traits to proteomic clusters based on the Gaussian Mixture model fitted on protein abundances profiles, as explained in section Variance component analysis of Materials and Methods. We chose for each fermentation trait the cluster of maximal membership probability. Most traits were assigned to a single protein cluster with a probability higher than 80%. The exceptions were *Sugar/EthanolYield* (26°), *X4MPP* (26°), *t.75* (26°), *t.lag* (26°) and *t.lag* at both temperatures. Average variance components for each cluster are represented in fig. 2. Altogether, the 56 fermentation traits fall into eight proteomic clusters, most of them being assigned to clusters 1 (16 traits), 2 (12 traits), 7 (12 traits), 3 (6 traits), 5 (5 traits). Note that no trait was assigned to cluster 8, which corresponds to the cluster with the lowest variances of additive effects. Despite similarities with protein abundance traits, fermentation traits are characterized by higher variance of additive and inbreeding effects and globally higher contrasts in genetic variance components (fig. 4). Overall, 8 traits were attributed to the same cluster at the two temperatures: J_{max}, r, $t\text{-}N_{max}$, *Viability-t-75*, *X4MMP*, *Hexanoic acid*, *Hexanol*, *Ethanol*.

In addition, we investigated, for each temperature, the link between protein category in each cluster and type of fermentation trait. We see that at 18°, most Basic Enological Parameters (BEP) fall in cluster 2 where we found proteins involved in metabolism and stress response. Life History Traits fall in cluster 7 (amino-acid and carbon metabolism) and carrying capacity K falls in cluster 9 (cell growth) while $t\text{-}N_{max}$ is found in cluster 6 (nucleotide metabolism and protein synthesis). At 26°, most Aromatic Traits fall in cluster 1 (cell fate, stress response), most Fermentation Kinetics traits are found in cluster 7 (amino-acid and carbon metabolism), and BEP are in cluster 4 (stress response).

In conclusion, traits are generally attributed to different clusters at the two temperatures, based on the underlying components of genetic variation. Those clusters are characterized by the enrichment in proteins with a certain functional category, that may vary between temperatures. Interestingly, we found an association between traits linked to different metabolic processes and proteins involved in such processes just by taking into account their genetic variance decomposition.

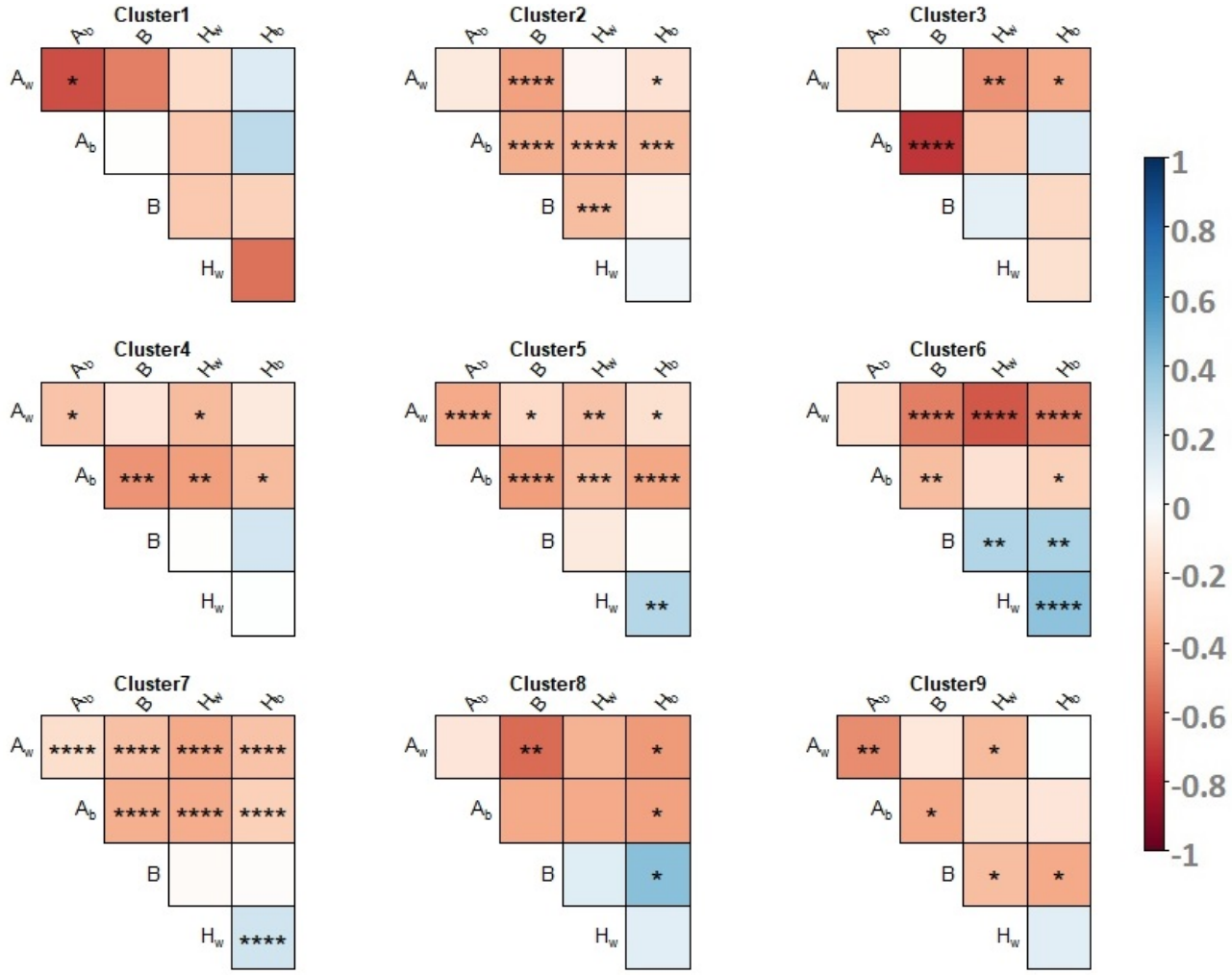


Figure 5 Pearson's correlation test performed to investigate the intra-cluster correlations on proteomic data. For each cluster, correlation between variances of the genetic effects are indicated by a color-code. Warm colors stand for negative correlations and cold colors for positive correlations. * significant at $p < 0.05$; ** significant at $p < 5 \cdot 10^{-3}$; *** significant at $p < 5 \cdot 10^{-4}$; **** significant at $p < 5 \cdot 10^{-5}$. No symbol: not significant.

Intra-cluster correlations between variance components

Pearson's correlation coefficients were computed for each pair of variance components within each cluster of proteins. Results clearly show different correlation structures between groups, particularly concerning correlation between the variances of heterosis and inbreeding effects (fig. 5). In cluster 1, variances of additive effects strongly and negatively correlate with each other. In cluster 3, there is a slightly negative correlation between $\sigma^2_{A_w}$ and the variances of heterosis effects, and there is a strong correlation between $\sigma^2_{A_b}$ and variance of inbreeding effects. Cluster 4 is characterized by a weak negative correlation between $\sigma^2_{A_w}$, $\sigma^2_{A_b}$, $\sigma^2_{H_w}$ variances, and between $\sigma^2_{A_b}$, and the variances of heterosis and inbreeding effects. Clusters 5 and 7 preserve the global correlation structure. In cluster 2, the variances of intra-specific heterosis and inbreeding effects are negatively correlated, in cluster 6 the variances of heterosis and inbreeding effects are positively correlated, in cluster 8 the variances of inter-specific heterosis and inbreeding effects are positively correlated, and in cluster 9 the variances of heterosis and inbreeding effects are negatively correlated. Altogether, when a statistical significant correlation between the variances of additive, heterosis and inbreeding effects is found, it is negative.

Variances of additive effects tend to be negatively correlated to variances of heterosis and inbreeding effects, and there is no straightforward relationship between the variances of heterosis and inbreeding effects: σ^2_B can be either negatively (cluster 9) or positively (cluster 6) correlated to both $\sigma^2_{H_b}$ and $\sigma^2_{H_w}$, negatively correlated to $\sigma^2_{H_w}$ (cluster 2), positively correlated to $\sigma^2_{H_b}$ (cluster 8). However, σ^2_B can also be independent from either $\sigma^2_{H_w}$ or $\sigma^2_{H_b}$ (clusters 1, 2, 3, 4, 5, 7, 8).

Discussion

In this paper, we focused on the comparative analysis of genetic variance components estimated through the decomposition of traits value quantified in a half-diallel cross during or at the end

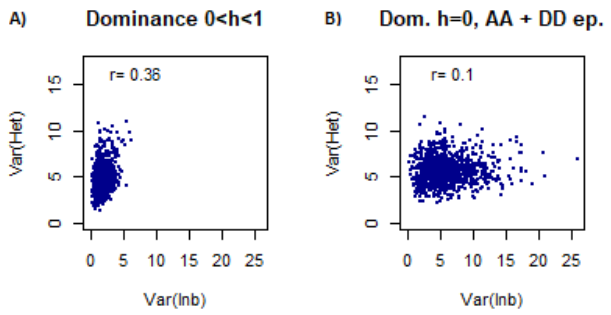


Figure 6 Correlation between the variances of heterosis and inbreeding effects for: A) Additive model with symmetrical dominance (no epistasis), B) Additive model with dominance of the strongest allele, additive × additive and dominance × dominance epistasis. The simulated half-diallel consisted of 11 parental lines. Phenotypic values were supposed to depend on 10 loci, and the number of alleles per loci was imposed to 11. Allele values were drawn from a gamma distribution (k=10, θ=20) and epistatic effects from a normal distribution ($\mathcal{N}(0,3)$).

of alcoholic fermentation. The cross design involved 11 yeast strains from two related species naturally associated with wine fermentations, *S. cerevisiae* and *S. uvarum*, and the set of traits quantified spanned from protein abundances to fermentation and life-history.

Genetic variances have been estimated through a comprehensive genetic model that allowed us to decompose the phenotypic value of a cross, including the parental inbred strains, in terms of additive and interaction effects. This decomposition can be described in the following way. The parental inbred lines have two identical haploid genomes, while the hybrids have two different haploid genomes, each inherited by one parent. Additive effects refer to the average value conferred by a single haploid genome with respect to any other haploid genome, and interaction effects refer to the non-additive effect of a particular genotype computed as the difference between the particular diploid value and the average additive effect of its haploid genomes. The presence of the parental inbreds in the experimental design permits a decomposition of those effects into heterosis and inbreeding effects. Inbreeding effect is defined as the difference between the value of the inbred strain (with the same haploid genome twice) and the average of all the crosses having at least one copy of the haploid parental genome. Heterosis effect is defined as the difference between a single pairwise genome combination and the average value of hybrids having one or the other haploid genome. Thanks to the presence of two different yeast species in our experimental design, we could distinguish intra-specific and inter-specific genetic effects. Indeed, the additive effect of a strain and the heterosis effect of a hybrid between two strains may differ depending on whether the strains belong to the same species or not. Therefore, intra-specific (respectively inter-specific) additive effect refers to the average value conferred by a single haploid genome with respect to any other haploid genome from the same specie (respectively from another species), and intra-specific (respectively inter-specific) heterosis effect refers to the difference between a single pairwise genome combination from the same specie (respectively from the two species) and the average value of the intra-specific (respectively inter-specific) hybrids having one or the other haploid genome.

This general model could be adapted to consider mitochondrial effects, which we did not declare for biological and technical reasons given in Materials and Methods. If such effects do exist in our genetic material they are expected to be weak and confounded with other effects.

The variance components of the genetic effects defined above have been estimated using the linear mixed model (*LMM*) described in eq. 2. Whenever a variance component was significant, it meant that genetic differences were found between strains. We checked the ability of the *LMM* to estimate genetic parameters by means of computer simulations and the robustness of the estimations through bootstrap analysis. In the simulations, despite residual variances that were not well correlated to their true value, estimated genetic variances were found to highly correlate with their true value (fig. 1). However, residuals quantified on the proteomic data highly correlate with their true value (see section Protein abundances). Bootstrap analysis, performed by sampling the 55 hybrids with replacement, conditionally to the 11 parental strains, revealed that for each variance component the estimations in the experimental sample were close to the median of the estimations in the bootstrap samples. For some traits and some variance components, the distribution of the bootstrap estimated variances were bimodal, suggesting a strong influence from a particular hybrid combination. However, it was never flat or smooth, in agreement with the non arbitrary choice of the parameters. Therefore, we are confident about the estimations of the genetic variances, conditionally to the parents of the diallel.

We were able to characterize the 615 proteins and the 28 fermentation and life-history traits quantified at 18° and 26° by a particular profile of genetic variance components despite the small number of parental inbred strains from which the half-diallel was built. We found that variances of intra- and inter-specific effects differed in a large extent, pointing out that the genetic effects are highly influenced by crossing strains from the same species or not. The degree of intra- and inter-specific genetic variation captures the evolutionary history the two species have undergone for the different traits. For instance, traits with a low variance of intra-specific additive effects but high variance of inter-specific additive effects have a high potential to evolve in inter- but not intra-specific crosses.

Each trait has been treated at each temperature separately, considering trait × temperature as independent characters. Indeed, genotype-by-environment interactions affect very commonly phenotypic variation. In particular, it is well documented that the genetic architecture of a trait is not stable under varying environments, highlighting the fact that evolutionary processes may depend largely upon ecological conditions (Falconer 1960; Lynch and Walsh 1998; Hermisson and Wagner 2004; Robinson *et al.* 2009; Malosetti *et al.* 2013). Accordingly we found a weak correlation between genetic variances at the two temperatures.

The molecular phenotypes (protein abundances) reflect the underlying genetic factors involved in the cellular processes regulating the most integrated traits. So we investigated the distribution of the components of genetic variation of protein abundances in relation to fermentation and life-history trait variance components. We found nine clear-cut clusters of protein variance components, and we were able to assign traits to these clusters based on their genetic variance components. Overall, the profiles of the fermentation and life-history traits associated to each cluster were close to that of the proteomic level, but they were characterized by higher variance of additive effects; further, we could not assign any trait to cluster 8, which has null

variance of additive effects, *i.e.* which is the group with the less heritable proteins. Altogether these results reveal that the most integrated traits have a higher evolutionary potential compared to protein abundances.

We tested for cluster enrichment in protein functions, based on the functional annotation of the proteins. Clusters were found to group together proteins of similar functions. Despite the fact that 63% of the proteins were found in different clusters at the two temperatures, the metabolic functions were preserved. This suggests temperature-specific regulatory changes that achieve the maintenance of cell functions. At the trait level, 16 over 28 fermentation/life-history traits (57%) fell into the same cluster at the two temperatures (fig. SF8). For the 12 remaining traits, changes in the distribution of variance components between the two temperatures can be explained by $G \times E$ interactions.

Beside, we have shown that the clusters were characterized by a particular profile of genetic variance components, which suggests that traits that group together share a similar evolutionary history. If all traits were neutral, they would have shown the same equilibrium level of total genetic variance of approximately $2NV_m$ (N the effective population size and V_m the mutational variance (Lynch and Hill 1986)) with a similar partition of genetic variance components. The existence of different profiles of variance components probably reflects that the different types of traits have been subject to particular selective pressures.

Beyond, the nine clusters were clearly distinguishable from each other from their pattern of correlation between variance components. Overall, the variances of intra- and inter-specific additive effects were negatively correlated to the variances of heterosis and inbreeding effects. This may reveal differences in the patterns of allele frequencies at the underlying loci. In a biallelic case, additive genetic variance is always maximum for intermediate allele frequencies, while dominance and epistatic variances (which are components of the variances of heterosis and inbreeding effects) are maximum for more extreme allele frequencies (Hill *et al.* (2008)). A trait with a high variance of additive effects is therefore expected to have lower dominance or epistatic variances. Conversely, a trait with low variance of additive effects may exhibit high dominance and epistatic variances.

In the common view, heterosis and inbreeding are corollary effects. However, we have shown that the variances of heterosis and inbreeding effects could be negatively, positively or not correlated to each other. For a better understanding of such a decoupling, we simulated a half-diallel design between N parental strains (for details see section Half-diallel simulation construction in Supplementary Materials). We computed the phenotypic values of the parental lines and hybrids starting with a simple additive model (neither dominance at any locus nor epistasis), then we added dominance and/or epistasis effects. We considered different degrees of dominance for each couple of alleles (including dominance of the strongest allele, h=0) and *additive* $\times$ *additive* and *dominance* $\times$ *dominance* epistasis, and we let the number of alleles per locus to vary. We considered all possible combinations of these effects. Finally we decomposed the values of the simulated traits into additive, heterosis and inbreeding effects.

Not surprisingly, the variances of heterosis and inbreeding effects are both null when there is neither dominance nor epistasis. If there is *additive* $\times$ *additive* epistasis with no dominance, the variances of heterosis and inbreeding effects are strictly correlated, with very low variance of heterosis effects. In the other conditions, the results depend on the number of parental lines. With three parents, the variance of heterosis and inbreeding effects are strictly equal, as it can be shown analytically (see section Inbreeding depression and heterosis variances are equal in three-parent diallel in Supplementary Materials). Otherwise the correlation between the variances of heterosis and inbreeding effects varies in function of the number of loci affecting the trait of interest, on the frequency of alleles in the population and on the presence of dominance and epistatic effects. In general, the correlation between the variances of heterosis and inbreeding effects tends to become null when the number of parental lines, the number of alleles per locus and the number of loci increase. Given these parameters, whether there is dominance or not, and whatever the type of dominance, the lowest correlations between the variances of heterosis and inbreeding effects are observed when there are both types of epistasis together (fig. 6 and fig. SF10). However in no case we get negative correlations between the two variances. Further, we decided to consider the data obtained on all the different cases together and we run as previously a Gaussian Mixture Model to cluster genetic variances components. We computed intra-cluster correlations varying the number of alleles per locus, the number of loci and the distribution in which we drew allele values. Those correlations did not show profiles similar to those obtained with real data (correlations between genetic effects are commonly positive or null).

Classical genetic studies and modern molecular evolutionary approaches now suggest that inbreeding effects and heterosis are predominantly caused by the presence of recessive deleterious mutations in the population (Charlesworth and Charlesworth 1999; Charlesworth and Willis 2009). Therefore understanding the effects of selection against deleterious alleles is crucial. Population structure also plays a key role in this framework. Indeed, population subdivision increases homozygosity through inbreeding, an effective process for purging deleterious alleles, but it also decreases selection efficiency by decreasing the genetic diversity. Allele frequency changes also modify the genetic variance components (Hill *et al.* 2008; Barton 2017). A more complex model, which takes into account selection, allele frequency, population structure and the presence of deleterious mutations is thus needed to explain our observations. Glémin *et al.* (2003) have discussed about the patterns of correlation between inbreeding effects and heterosis in a structured population assuming low frequencies of deleterious mutations, only present in the heterozygous state. They defined within- and between-demes inbreeding depression as the decline in mean fitness of selfed individuals relative to out-crossed individuals within the demes and as the decline in mean fitness of selfed individuals relative to out-crossed individuals between demes, respectively; and heterosis as the excess in mean fitness of individuals produced by out-crosses between demes relative to mean fitness of individuals produced by out-crosses within the demes. They stated that population structure decreases within-demes inbreeding depression while it increases between-deme inbreeding depression, and that increasing the inbreeding coefficient reduces within- and between-deme inbreeding depression and heterosis. A similar result was obtained by Roze and Rousset (2004) who considered a diffusion model in a population of partially selfing individuals subdivided according to an island model, with a large but finite number of demes. They found that generally within-deme inbreeding depression and heterosis are positively correlated upon selfing and, when the degree of pop-

ulation subdivision is high, inbreeding depression and heterosis are negatively correlated. To our knowledge, the present study reports the first experimental example of such a decoupling.

In conclusion, our findings have special relevance in three main directions: *(i) Detection of Quantitative Trait Loci (QTL).* Variances of additive effects are crucial for the detection of genes with significant quantitative effect, and variances of heterosis/inbreeding effects for the detection of gene-gene interactions when the part of genetic variance they explain is large; *(ii) Integration of proteomic data into Genome Scale Metabolic (GSM) model:* we assigned fermentation traits to clusters obtained on the components of genetic variation of protein abundances. Traits associated to a metabolic process were linked to proteins involved to such process, therefore we are confident that integrating proteins related to the most integrated traits into a GSM could improve their prediction, with particular attention to the prediction of heterosis; *(iii) Model heterosis and inbreeding variation:* we have highlighted various patterns of variation between the variances of heterosis and inbreeding effects that cannot be explained with simple quantitative genetics models. It would be interesting to construct *in silico* experiments to search for the key parameters that drive these patterns.

Acknowledgments

We thank very much Dr. Arnaud Le Rouzic for exciting discussions and its material to purse the preliminary analysis of the diallel design. We thank very much Dr. Monique Bolotin for her help in the functional annotation of the proteins, and Dr. Warren Albertin and Dr. Philippe Marullo for their advice regarding yeast genetic material. This work was supported by a public PhD grant of the French National research Agency (ANR) as part of the "Investissement d'Avenir" program, through the "IDI 2015" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

Literature Cited

Abdulrehman, D., P. T. Monteiro, M. C. Teixeira, N. P. Mira, A. B. Lourenço, *et al.*, 2011 Yeastract: providing a programmatic access to curated transcriptional regulatory associations in saccharomyces cerevisiae through a web services interface. Nucleic Acids Research **39**: D136–D140.

Albertin, W., T. da Silva, M. Rigoulet, B. Salin, I. Masneuf-Pomarede, *et al.*, 2013a The mitochondrial genome impacts respiration but not fermentation in interspecific saccharomyces hybrids. PLOS ONE **8**: 1–14.

Albertin, W., P. Marullo, M. Aigle, A. Bourgais, M. Bely, *et al.*, 2009 Evidence for autotetraploidy associated with reproductive isolation in saccharomyces cerevisiae: towards a new domesticated species. Journal of Evolutionary Biology **22**: 2157–2170.

Albertin, W., P. Marullo, M. Bely, M. Aigle, A. Bourgais, *et al.*, 2013b Linking Post-Translational Modifications and Variation of Phenotypic Traits. Molecular & Cellular Proteomics **12**: 720–735.

Barton, N. H., 2017 How does epistasis influence the response to selection? Heredity (Edinb) **118**.

Blein-Nicolas, M., W. Albertin, T. da Silva, B. Valot, T. Balliau, *et al.*, 2015 A systems approach to elucidate heterosis of protein abundances in yeast. Mol Cell Proteomics **14**: 2056–71.

Blein-Nicolas, M., W. Albertin, B. Valot, P. Marullo, D. Sicard, *et al.*, 2013 Yeast proteome variations reveal different adaptive responses to grape must fermentation. Molecular Biology and Evolution **30**: 1368.

Blein-Nicolas, M., H. Xu, D. de Vienne, C. Giraud, S. Huet, *et al.*, 2012 Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics. PROTEOMICS **12**: 2797–2801.

Bulmer, M. G., 1980 *The mathematical theory of quantitative genetics /M.G. Bulmer*. Clarendon Press ; New York : Oxford University Press Oxford.

Charlesworth, B. and D. Charlesworth, 1999 The genetic basis of inbreeding depression. Genetical Research **74**: 329–340.

Charlesworth, B., M. T. Morgan, and D. Charlesworth, 1991 Multilocus models of inbreeding depression with synergistic selection and partial self-fertilization. Genetics Research **57**: 177–194.

Charlesworth, D. and J. H. Willis, 2009 The genetics of inbreeding depression. Nature Reviews Genetics **10**: 783—-796.

Cherry, J. M., E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, *et al.*, 2012 Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research **40**: D700–705.

Cockerham, C. C. and B. S. Weir, 1977 Quadratic analyses of reciprocal crosses. Biometrics **33**: 187–203.

da Silva, T., W. Albertin, C. Dillmann, M. Bely, S. la Guerche, *et al.*, 2015 Hybridization within saccharomyces genus results in homoeostasis and phenotypic novelty in winemaking conditions. PLOS ONE **10**: 1–24.

Davenport, C. B., 1908 Degeneration, albinism and inbreeding. Science **28**: 454–455, WOS:000201859500057.

Eberhart, S. A. and C. O. Gardner, 1966 A General Model for Genetic Effects. Biometrics **22**: 864–881.

Falconer, D. S., 1960 *Introduction to quantitative genetics*. New York,: Ronald Press Co.

Fievet, J. B., C. Dillmann, and D. de Vienne, 2010 Systemic properties of metabolic networks lead to an epistasis-based model for heterosis. Theoretical and Applied Genetics **120**: 463–473, WOS:000272803700025.

Fiévet, J. B., T. Nidelet, C. Dillmann, and D. de Vienne, 2018 Heterosis is a systemic property emerging from nonlinear genotype-phenotype relationships: evidence from in vitro genetics and computer simulations. Frontiers in Genetics **9**.

Glémin, S., J. Ronfort, and T. Bataillon, 2003 Patterns of inbreeding depression and architecture of the load in subdivided populations. Genetics **165**: 2193–2212.

Gowen, J. W., 1952 *Heterosis*. Iowa state pr edition.

Graham, G. I., D. W. Wolff, and C. W. Stuber, 1997 Characterization of a yield quantitative trait locus on chromosome five of maize by fine mapping. Crop Sci. **37**: 1601–1610, WOS:A1997XZ35500033.

Greenberg, A. J., S. R. Hackett, L. G. Harshman, and A. G. Clark, 2010 A hierarchical bayesian model for a novel sparse partial diallel crossing design. Genetics .

Griffing, B., 1956 Concept of general and specific combining ability in relation to diallel crossing systems. Australian Journal of Biological Sciences **9**: 463–493.

Gumedze, F. and T. Dunne, 2011 Parameter estimation and inference in the linear mixed model. Linear Algebra and its Applications **435**: 1920 – 1944.

Hallauer, A., W. Russell, and K. LAMKEY, 1988 Corn breeding: 463-564. Sprague, GF and. IW Dudley: Corn and corn improvement. Agron. Monogr.(third edition). ASA, CSSA and SSSA Madison, WI .

Hallauer, A. R. and J. B. M. Filho, 1988 *Quantitative Genetics in Maize Breeding*. Iowa State University Press.

Hermisson, J. and G. P. Wagner, 2004 The Population Genetic Theory of Hidden Variation and Genetic Robustness. Genetics **168**: 2271–2284.

Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and theory point to mainly additive genetic variance for complex traits. PLOS Genetics **4**: 1–10.

Huang, X., S. Yang, J. Gong, Q. Zhao, Q. Feng, *et al.*, 2016 Genomic architecture of heterosis for yield traits in rice. Nature **537**: 629–633.

Hull, F., 1946 Overdominance and Corn Breeding Where Hybrid Seed Is Not Feasible. Journal of the American Society of Agronomy **38**: 1100–1103, WOS:A1946UC58500007.

Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, 2017 KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research **45**: D353–D361.

Kanehisa, M. and S. Goto, 2000 KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**: 27–30.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, 2016 KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research **44**: D457–462.

Kurtz, Z. D., C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, *et al.*, 2015 Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology **11**: e1004226.

Lande, R. and D. W. Schemske, 1985 The evolution of self-fertilization and inbreeding depression in plants. i. genetic models. Evolution **39**: 24–40.

Lariepe, A., B. Mangin, S. Jasson, V. Combes, F. Dumas, *et al.*, 2012 The Genetic Basis of Heterosis: Multiparental Quantitative Trait Loci Mapping Reveals Contrasted Levels of Apparent Overdominance Among Traits of Agronomical Interest in Maize (Zea mays L.). Genetics **190**: 795–U835, WOS:000300621200037.

Lenarcic, A. B., K. L. Svenson, G. A. Churchill, and W. Valdar, 2012 A general bayesian approach to analyzing diallel crosses of inbred strains. Genetics **190**: 413–435.

Lynch, M. and W. G. Hill, 1986 PHENOTYPIC EVOLUTION BY NEUTRAL MUTATION. Evolution; International Journal of Organic Evolution **40**: 915–935.

Lynch, M. and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer, Google-Books-ID: UhCCQgAACAAJ.

Malosetti, M., J.-M. Ribaut, and F. A. van Eeuwijk, 2013 The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. Frontiers in Physiology **4**.

Martì-Raga, M., E. Peltier, A. Mas, G. Beltran, and P. Marullo, 2017 Genetic Causes of Phenotypic Adaptation to the Second Fermentation of Sparkling Wines in Saccharomyces cerevisiae. G3: Genes, Genomes, Genetics **7**: 399–412.

Melchinger, A. E. and R. K. Gumber, 1998 Overview of Heterosis and Heterotic Groups in Agronomic Crops. In *Concepts and Breeding of Heterosis in Crop Plants*, edited by K. R. Larnkey and J. E. Staub, pp. 29–44, Crop Science Society of America, USA.

Monteiro, P. T., N. D. Mendes, M. C. Teixeira, S. d'Orey, S. Tenreiro, *et al.*, 2008 Yeastract-discoverer: new tools to improve the analysis of transcriptional regulatory associations in saccharomyces cerevisiae. Nucleic Acids Research **36**: D132–D136.

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

Powers, L., 1944 An expansion of Jones's theory for the explanation of heterosis. The American Naturalist **78**: 275–280.

Ramya, A. R., L. Ahamed M, C. T. Satyavathi, A. Rathore, P. Katiyar, *et al.*, 2018 Towards Defining Heterotic Gene Pools in Pearl Millet [Pennisetum glaucum (L.) R. Br.]. Frontiers in Plant Science **8**.

Redden, R., 1991 The effect of epistasis on chromosome mapping of quantitative characters in wheat. I. Time to spike emergence. Australian Journal of Agricultural Research **42**: 1.

Robinson, M. R., A. J. Wilson, J. G. Pilkington, T. H. Clutton-Brock, J. M. Pemberton, *et al.*, 2009 The Impact of Environmental Heterogeneity on Genetic Architecture in a Wild Population of Soay Sheep. Genetics **181**: 1639–1648.

Ronnegard, L., X. Shen, and M. Alam, 2010 hglm: A package for fitting hierarchical generalized linear models. The R Journal **2**: 20–28.

Roze, D. and F. Rousset, 2004 Joint effects of self-fertilization and population structure on mutation load, inbreeding depression and heterosis. Genetics **167**: 1001–1015.

Ruepp, A., A. Zollner, D. Maier, K. Albermann, J. Hani, *et al.*, 2004 The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research **32**: 5539–5545.

Schnable, P. S. and N. M. Springer, 2013 Progress toward understanding heterosis in crop plants. Annu Rev Plant Biol **64**: 71–88.

Scrucca, L., M. Fop, T. B. Murphy, and A. E. Raftery, 2016 mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal **8**: 205–233.

Seymour, D. K., E. Chae, D. G. Grimm, C. Martín Pizarro, A. Habring-Müller, *et al.*, 2016 Genetic architecture of nonadditive inheritance in Arabidopsis thaliana hybrids. Proc. Natl. Acad. Sci. U.S.A. **113**: E7317–E7326.

Shao, H., L. C. Burrage, D. S. Sinasac, A. E. Hill, S. R. Ernest, *et al.*, 2008 Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. Proceedings of the National Academy of Sciences **105**: 19910–19914.

Shull, G. H., 1908 The Composition of a Field of Maize. Journal of Heredity **os-4**: 296–301.

Sprague, G. F. and E. L. Tatum, 1942 General vs. specific combining ability in single crosses of corn. PROTEOMICS **34**: 923–932.

Teixeira, M. C., P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes, *et al.*, 2006 The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. Nucleic Acids Research **34**: D446–D451.

Teixeira, M. C., P. T. Monteiro, J. F. Guerreiro, J. P. Gonçalves, N. P. Mira, *et al.*, 2014 The yeastract database: an upgraded information system for the analysis of gene and genomic transcription regulation in saccharomyces cerevisiae. Nucleic Acids Research **42**: D161–D166.

Tsagris, M. T., S. Preston, and A. T. A. Wood, 2011 A data-based power transformation for compositional data. ArXiv e-prints .

Wright, S., 1934 Physiological and evolutionary theories of dominance. American Naturalist **68**: 24–53, WOS:000200907000002.

Xiao, J., J. Li, L. Yuan, and S. Tanksley, 1995 Dominance Is the Major Genetic-Basis of Heterosis in Rice as Revealed by Qtl Analysis Using Molecular Markers. Genetics **140**: 745–754, WOS:A1995RA36600028.

Zhu, J. and B. S. Weir, 1996 Mixed model approaches for diallel analysis based on a bio-model. Genetical Research **68**: 233–240.

3.5 Conclusions

In this analysis, I have characterize phenotypic variation at each level of cellular organization by means of genetic and residual variance components contributing to each trait through the decomposition of the particular diallel-cross design.

Traits have been treated at each temperature as independent characters and the portion of variance attributed to genetic effects was further decomposed into additive, inbreeding and heterosis effects, distinguishing intra- and inter-specific additive and heterosis effects.

The analysis of variance components in the population have allowed to identify:

- the presence of genotype-by-environment interaction at every level of cellular organization;
- the independence of heterosis variances on the type of cross at the proteomic level;
- a buffering mechanism towards genetic interaction for life-history and fermentation traits;
- groups of protein abundances and fermentation and life-history traits that have possibly been submitted to the same selective pressures;

Along, the most striking result was the possible decoupling between heterosis and inbreeding depression that can be explained by simple genetic models with epistatic interactions.

Beside, integration of the two different levels of cellular organization have been performed through association of proteins and fermentation/life-history traits sharing a similar partition of genetic variance components: groups identified at the proteomic level shared functional properties, and it was possible to associate fermentation and life-history traits to proteomic groups.

In the following, I focus in the characterization of the more integrated traits (life-history and fermentation) by means of the underlying metabolic fluxes in order to investigate the main mechanisms underlying multi-trait variation. Indeed, metabolic fluxes result from network functioning and integrate the activities of possibly many proteins. To this end, I have introduced protein abundance data into constraint-based models and predicted steady-state fluxes for each strain per temperature separately.

Chapter 4

Chapter 4

Metabolism modeling

Life-history traits are the observable results of unobservable processes that occur at a cellular scale. During the last decades, novel profiling technologies and high-throughput techniques have made possible the inventory of a majority of biological components underlying phenotypic variation along with genome-scale characterization of genomic sequences. This included transcriptomic, metabolomic and proteomic data at individual level. Quantification of omic data have enabled biologists to view and study cell as a system of interacting components. The metabolism of a cell can be seen as a network in which compounds are transformed through a series of steps into other compounds. This process is governed by enzymes, which are catalysts allowing reactions to proceed more rapidly and which tune the rate of the metabolic reactions, for example in response to changes in the cell's environment or to signals from other cells.

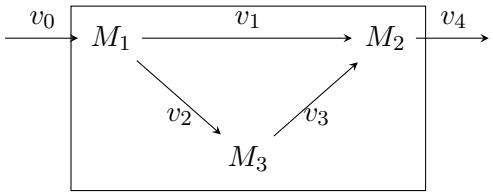
Based on genome annotation and biochemical knowledge, genome-scale metabolic models have been proposed for the description of cell metabolism. They can be used to study genotype-phenotype relationships, and their application to microbial strain engineering is increasing in popularity. To this end, the determination of flux distributions is essential, for a better understanding of the interplay between different metabolic pathways, for investigating the genetic and molecular bases of the multi-trait variation and, lastly, for the prediction of the integrated phenotypes. Nevertheless, metabolic fluxes are difficult to measure. Metabolic Flux Analysis is powerful (Antoniewicz, 2015), but it is based on RMN and differential usage of radioactive isotopes. It remains low-throughput and cannot be applied on numerous individuals. Technical developments in mass spectrometry popularized metabolomics (Nicholson and Lindon, 2008), which allows to characterize in some extent the metabolome, *i.e.* the set of metabolites in a cell, tissue, organ or organism. However, the technique still suffers from standardization procedures and does not allow for high-throughput quantitative comparisons (Riekeberg and Powers, 2017).

Sophisticated methods for the analysis of the global organization of cellular behavior have been proposed, one of them being constraint-based reconstruction and analysis applied to genome-scale metabolic networks (Bordbar et al., 2014). In this chapter, I briefly review the main approaches developed for constraint-based modeling of cellular metabolism, and I present the main properties of the yeast central carbon metabolism model that I use in the Chapter 5.

Metabolism
Metabolism is the set of chemical reactions that take place within each cell of a living organism and that provides energy for vital processes and for synthesizing new organic material.

4.1 Constraint-Based Modeling (CBM)

The metabolism of a cell can be described by the complex network of chemical reactions between M metabolites and N reactions. In the toy model below, metabolite $M1$ enters the cell and is transformed into $M2$ and $M3$, while $M2$ can also be produced by an isomerization of $M3$. $M2$ is exported outside the cell at a rate v_4. Fluxes are reaction rates v_i that depend on enzyme activities.



The dynamics of the process results in temporal changes of metabolite concentrations m_i of M_i. Here, m_1 changes following:

$$\dot{m}_1 = v_0 - v_1 - v_2$$

4.1.1 Mathematical formalism

In a mathematically consistent framework, it is possible to define an $M \times N$ stoichiometric matrix S in which rows correspond to the stoichiometric coefficients of the corresponding metabolites of all the reactions. In the toy model above, the stoichiometric coefficients are 1, -1 or 0. Assuming mass-balance and limited interval of variation for the different reactions, the problem consists in finding the set of fluxes $\boldsymbol{v} \in \mathbb{R}^N$ such that

$$S\boldsymbol{v} = \dot{\boldsymbol{m}} \tag{4.1}$$

$$\boldsymbol{v}^{inf} \leq \boldsymbol{v} \leq \boldsymbol{v}^{sup} \tag{4.2}$$

where $\dot{\boldsymbol{m}} \in \mathbb{R}^M$ is the set of metabolite input/output rates, and the vectors $\boldsymbol{v}^{inf}$, $\boldsymbol{v}^{sup}$ are the extremes of variability of the fluxes. In general, $M \leq N$ and the system of stoichiometric equations is typically under-determined. Rather than deriving a single solution, constraint-based models have an associated solution space:

$$L = \{\boldsymbol{v} | S\boldsymbol{v} = \dot{\boldsymbol{m}}, \boldsymbol{v}^{inf} \leq \boldsymbol{v} \leq \boldsymbol{v}^{sup}\}$$

in which all feasible $\boldsymbol{v}$ exist given the imposed constraints that account for the different processes acting on and in cells.

Eq. 4.1 describes metabolic fluxes that are constrained by network topology. In general, it is assumed that cells consume and produce metabolites at a constant rate in a mass-balance manner, that is cells are under steady-state and $\dot{\boldsymbol{m}} = 0$ represents a further constraint. Fluxes are also constrained by upper and lower bounds, generally known from the literature and used to model a specific cellular process (ineq. 4.2). Further constraints such as physiologically relevant fluxes can be introduced to reduce L. Different techniques have been proposed to deduce network behavior by dimensional reduction of L, most of them based on two key components: the method of analysis to predict fluxes and observed/known constraints on the biological system.

Flux Balance Analysis (FBA). The first constraint-based method for biological predictions was Flux Balance Analysis (Fell and Small, 1986; Varma and Palsson, 1994). In FBA, an objective function is introduced and is assumed to be maximized/minimized by the cell, such as the consumption/production of metabolites or of biomass. It requires experimental inputs to establish the metabolite composition of cell biomass. Notice that the optimal solution to the flux-balance problem is rarely unique with many possible, and equally optimal, solutions.

Flux Variability Analysis (FVA). The method consists in the identification of lower and upper values of fluxes through each reaction iteratively when the flux of the objective is typically constrained to its maximum/minimum value (Gudmundsson and Thiele, 2010). Reactions that support a low variability of fluxes are likely to be of a higher importance to an organism.

Both methods, FBA and FVA, require identification of objective functions given the experimental data. However, objective functions may change under changing environments and under different conditions.

Markov Chain Monte Carlo techniques (MCMC). This approach does not require to assume any objective. It consists in sampling in L to provide a probability distribution for the feasible fluxes. The imposition of constraints in the model defines the associated solution space of the CBM, *i.e.*

$$L \equiv L(\dot{\boldsymbol{m}}; \boldsymbol{v}^{inf}; \boldsymbol{v}^{sup}) \tag{4.3}$$

Simple constraints include input and output ranges on the basis of uptake/secretion of metabolites and genetic knockouts by setting reactions to zero. More advanced techniques include setting metabolite rates or flux bounds to experimentally measured values (reviewed in section 4.2).

4.1.2 Exploring the space of possible solutions

MCMC techniques have been proposed to approximately compute the posterior distribution of fluxes in L, such as the *Hit and Run* (HR) algorithm (Bélisle et al., 1993). Recently, a novel method, which combines statistical physics and Bayesian approaches, and which does not require sampling in L, has been proposed by Braunstein et al. (2017), the *Expectation Propagation algorithm* (EP algorithm).

Hit and Run algorithm

The problem consists in efficiently generate samples in $L \subset \mathbb{R}^N$ with polygonal constraints imposed by the lower and upper bounds of fluxes. The algorithm proposed by Bélisle et al. (1993) consists in iteratively exploring the solution space by increasing the dimensionality (k):

- *Step 0.* Choose a starting point $\boldsymbol{v}_0 \in L$, with $k = 0$;
- *Step 1.* Generate a random direction $\boldsymbol{e}^k \in \mathbb{R}^N$, with $\left\|\boldsymbol{e}^k\right\| = 1$;
- *Step 2.* Choose $\lambda_k \in \Lambda_k$, where $\Lambda_k = \{\lambda \in \mathbb{R} : \boldsymbol{v}_k + \lambda \boldsymbol{e}^k \in L\}$ from the density distribution

$$f_k(\lambda) = \frac{f(\boldsymbol{v}_k + \lambda \boldsymbol{e}^k)}{\int f(\boldsymbol{v}_k + r\boldsymbol{e}^k)dr} \tag{4.4}$$

  and where $f(\boldsymbol{v})$ is the prior density distribution of $\boldsymbol{v} \in L$, assuming a multinomial distribution.

- *Step 3.* Set $\boldsymbol{v}_{k+1} = \boldsymbol{v}_k + \lambda_k \boldsymbol{e}^k$ and $k = k + 1$;
- *Step 4.* Return to *Step 1.*

The accuracy obtained with HR depends of course on the number of samples, and sampling accurately can be very time consuming.

Expectation propagation algorithm

Braunstein et al. (2017) formulated the problem as follows: consider the set of fluxes $\boldsymbol{v}$ compatible with eq. 4.1 and ineq. 4.2. It is possible to define a quadratic energy function $\mathcal{E}(\boldsymbol{v})$ whose minimum(s) lies on the assignment of variables $\boldsymbol{v}$ satisfying the stoichiometric constraints in equation 4.1:

$$\mathcal{E}(\boldsymbol{v}) = \frac{1}{2}(S\boldsymbol{v} - \dot{\boldsymbol{m}})^\top (S\boldsymbol{v} - \dot{\boldsymbol{m}}) \tag{4.5}$$

It is easy to see that if $\boldsymbol{v}$ satisfies eq. 4.1, $\mathcal{E}(\boldsymbol{v})$ will be at a/the minimum. Therefore, the likelihood of observing $\dot{\boldsymbol{m}}$ given a set of fluxes $\boldsymbol{v}$ can be expressed as a Boltzmann distribution:

$$P(\dot{\boldsymbol{m}}|\boldsymbol{v}) = \left(\frac{\beta}{2\pi}\right)^{\frac{M}{2}} e^{-\frac{\beta}{2}(S\boldsymbol{v}-\dot{\boldsymbol{m}})^{\mathrm{T}}(S\boldsymbol{v}-\dot{\boldsymbol{m}})} \tag{4.6}$$

where β is a positive parameter, the inverse of temperature in statistical physics jargon, that governs the penalty of whose configurations of fluxes that are far from the minimum of the energy. Using Bayes formula, the posterior probability of observing the set of fluxes $\boldsymbol{v}$ given $\dot{\boldsymbol{m}}$ is:

$$P(\boldsymbol{v}|\dot{\boldsymbol{m}}) = \frac{P(\dot{\boldsymbol{m}}|\boldsymbol{v})P(\boldsymbol{v})}{P(\dot{\boldsymbol{m}})} \tag{4.7}$$

where the prior

$$P(\boldsymbol{v}) = \prod_{n=1}^{N} \psi_n(v_n) = \prod_{n=1}^{N} \frac{\mathbb{1}(v_n \in [v_n^{inf}, v_n^{sup}])}{v_n^{sup} - v_n^{inf}} \tag{4.8}$$

The function $\mathbb{1}(v_n \in [v_n^{inf}, v_n^{sup}])$ is an indicator function that takes values 1 if $v_n \in [v_n^{inf}, v_n^{sup}]$ and 0 otherwise. It constraints flux values to verify the inequality imposed in eq. 4.2. An expression for the posterior distribution is:

$$P(\boldsymbol{v}|\dot{\boldsymbol{m}}) = \frac{1}{P(\dot{\boldsymbol{m}})}\left(\frac{\beta}{2\pi}\right)^{\frac{M}{2}} e^{-\frac{\beta}{2}(S\boldsymbol{v}-\dot{\boldsymbol{m}})^{\mathrm{T}}(S\boldsymbol{v}-\dot{\boldsymbol{m}})} \prod_{n=1}^{N} \psi_n(v_n) \tag{4.9}$$

Computation of the marginal distribution $P(v_n|\dot{\boldsymbol{m}})$ for each $n \in \{1, 2, \ldots, N\}$ requires calculation of multiple integrals, which is computationally very expensive and cannot be performed analytically in a efficient way. Therefore, the EP technique suggests to replace the prior distribution of fluxes, but not the n-th flux, by a Gaussian distribution

$$\phi_m(v_m; a_m, d_m) = \frac{e^{-\frac{(v_m-a_m)^2}{2d_m}}}{\sqrt{2\pi d_m}} \tag{4.10}$$

whose mean and variance are constrained to be equal to the one of $\psi_m(v_m)$. To this end, consider the n-th flux, its corresponding approximate prior $\phi_n(v_n; a_n, d_n)$ and define a tilted distribution $Q^{(n)}$ as

$$Q^{(n)}(\boldsymbol{v}|\dot{\boldsymbol{m}}) \equiv \frac{1}{\boldsymbol{Z}_{Q^{(n)}}} e^{-\frac{\beta}{2}(S\boldsymbol{v}-\dot{\boldsymbol{m}})^{\mathrm{T}}(S\boldsymbol{v}-\dot{\boldsymbol{m}})} \psi_n(v_n) \prod_{m\neq n} \phi_m(v_m) \tag{4.11}$$

where $\boldsymbol{Z}_{Q^{(n)}}$ is the normalization constant:

$$\boldsymbol{Z}_{Q^{(n)}} = \int d^n\boldsymbol{v} e^{-\frac{\beta}{2}(S\boldsymbol{v}-\dot{\boldsymbol{m}})^{\mathrm{T}}(S\boldsymbol{v}-\dot{\boldsymbol{m}})} \psi_n(v_n) \prod_{m\neq n} \phi_m(v_m) \tag{4.12}$$

The problem consists in finding the unknown parameters a_n and d_n of $\phi_n(v_n; a_n, d_n)$ such that the multivariate-truncated Gaussian distribution

$$Q(\boldsymbol{v}|\dot{\boldsymbol{m}}) \equiv \frac{1}{\boldsymbol{Z}_Q} e^{-\frac{\beta}{2}(S\boldsymbol{v}-\dot{\boldsymbol{m}})^{\mathrm{T}}(S\boldsymbol{v}-\dot{\boldsymbol{m}})} \prod_{n=1}^{N} \phi_n(v_n) \tag{4.13}$$

is as close as possible to $Q^{(n)}$. This process can be performed by matching the first two moments of the distribution

$$\begin{cases} \langle v_n \rangle_{Q^{(n)}} = \langle v_n \rangle_Q \\ \langle v_n^2 \rangle_{Q^{(n)}} = \langle v_n^2 \rangle_Q \end{cases}$$

from which a relation for the parameters a_n and d_n can be found through sequentially repeating the update step for all fluxes and iterate until a numerical convergence is reached.

4.2 Integration of experimental data

The advent of high-throughput techniques have allowed quantification of omic data that have encouraged scientists to propose novel methods for the integration into CBM. They can be used to add an additional layer of constraints for reaction fluxes (Patil and Nielsen, 2005), to determine context specific flux distributions (Lobel et al., 2012) or to compare and validate FBA predictions (Schuetz et al., 2012). Indeed, experimental data, even incomplete, provides information about the intra-cellular processes in the organisms. Different approaches have been proposed, with different rationales and advantages:

GIMME (Gene Inactivity Moderated by Metabolism and Expression) uses quantitative gene expression data and one or more presupposed metabolic objectives to produce the context-specific reconstruction that is most consistent with the available data (Becker and Palsson, 2008). Under the assumption that environmental changes determine metabolic pathway usage, enzymes associated to metabolic pathways that are not used are assumed to be not synthesized. Therefore, the method searches for sub-models by constraining to zero the fluxes to which no associated gene expression data is observed. In this way, each condition may be characterized by a different combination of fluxes.

Eflux. A variation of GIMME was proposed (Colijn et al., 2009) that used transcriptomic expression data to model the maximum possible flux through metabolic reactions. When the expression for a particular enzyme-coding gene is low (relative to some reference), a tight constraint is posed. When expression is high the constraint is looser. Then FBA is performed with the applied constraints and an appropriate objective function. The method was successfully applied to study light and temperature acclimation in *Arabidopsis thaliana* (Töpfer et al., 2013). Instead of a single objective function, the authors considered the maximization of a collection of metabolic functions that were characterized under different environmental conditions. The method allowed to determine which metabolic pathways, from both primary and secondary metabolism, were significantly affected in the experiments.

IOMA or Integrative Omics-Metabolic Analysis method is formulated as a quadratic programming problem that seeks a steady-state flux distribution, in which flux through reactions with measured proteomic and metabolomic data are as consistent as possible with kinetically derived flux estimations (Yizhak et al., 2010). It assumes that protein abundances are proportional to kinetic fluxes:

$$v_i^{kin} = E_i(k_{E_i} + \epsilon_i) \tag{4.14}$$

where v_i^{kin} denotes the flux of the i-th kinetic reaction, E_i the abundance of the enzyme associated to the i-th reaction, k_{E_i} the kinetic constant associated to reaction i and ϵ_i is a residual. The problem turns in finding the set of fluxes and of kinetic constants that satisfy the stoichiometric and constrains on fluxes while minimizing the variance of the residuals. In addition, metabolomic data can be exploited to estimate kinetic constants.

Profile comparison. The method proposed by Lee et al. (2012) is based on maximization of the correlation between experimentally measured absolute gene expression data or protein abundances and predicted internal reaction fluxes. It assumes that the likely solution in L minimizes the distance between absolute gene expression profile and that of fluxes. The problem is formulated as

a linear programming problem, and turns in finding the set of fluxes for which:

$$Z = \sum_{j=1}^{k} \frac{1}{\sigma_i} |v_i - E_i| \tag{4.15}$$

is minimum.

In their work, Lee et al. (2012) have shown that the method proposed outperformed with respect to traditional methods in predicting exchange fluxes (Table 4.1), using quantitative transcriptomic data acquired from *S. cerevisiae* cultures under two growth conditions. This approach improved prediction and did not require knowledge of the biomass composition of the organism under the conditions of interest. For these reasons, I have chosen to adopt a similar approach for predicting metabolic fluxes in the HeterosYeast dataset.

Flux	Observed	Predicted		
		Profile comparison	FBA	GIMME
Ethanol	23.8	25.7	0	0
CO_2	22.7	31.5	37.6	31.5
Glycerol	3.54	0	0	0
Acetate	0.311	0.016	0	0
Trehalose	0.0356	0.0301	0	0
Lactate	0.00873	0.0301	0	0

Table 4.1: Observed and predicted exchange fluxes from different data-integration methods (Lee et al., 2012). The profile comparison method results in a better prediction of fluxes.

4.3 The DynamoYeast model

The DynamoYeast is a previously developed constraint-based model of central carbon metabolism of *S. cerevisiae* (Celton et al., 2012). This model comprises the cytosol, mitochondria and extracellular medium and includes upper and lower glycolysis, the PPP (Pentose Phosphate Pathway), the synthesis of glycerol, the synthesis of ethanol, and the reductive and oxidative branches of the TCA as the main metabolic pathways. It consisted of 70 reactions and 60 metabolites. Figure 4.1 shows a representation of this model. In red are indicated flux constraints for exchange metabolic fluxes.

4.3.1 Sampling the solution space

The feasible space of solution L of fluxes from the DynamosYeast model was first characterized by the posterior distribution of fluxes obtained through the HR sampling method (implemented in *R* by Meersche et al. (2009)). We compared the efficiency of the HR algorithm to the predictions obtained through the EP algorithm (Braunstein et al., 2017).

The posterior density distribution obtained by HR and EP algorithms were compared after running the HR with a burning length equal to 10^6 and a jump of 0.5, for a number of iteration from 10^6 to 10^7, and the EP algorithm with a high β parameter (Boltzmann inverse temperature parameter). Figure 4.2 shows the sampled space of solution through the HR (histograms) and the EP estimate (red curve). Even though the results were not exactly the same, the two distributions were similar.

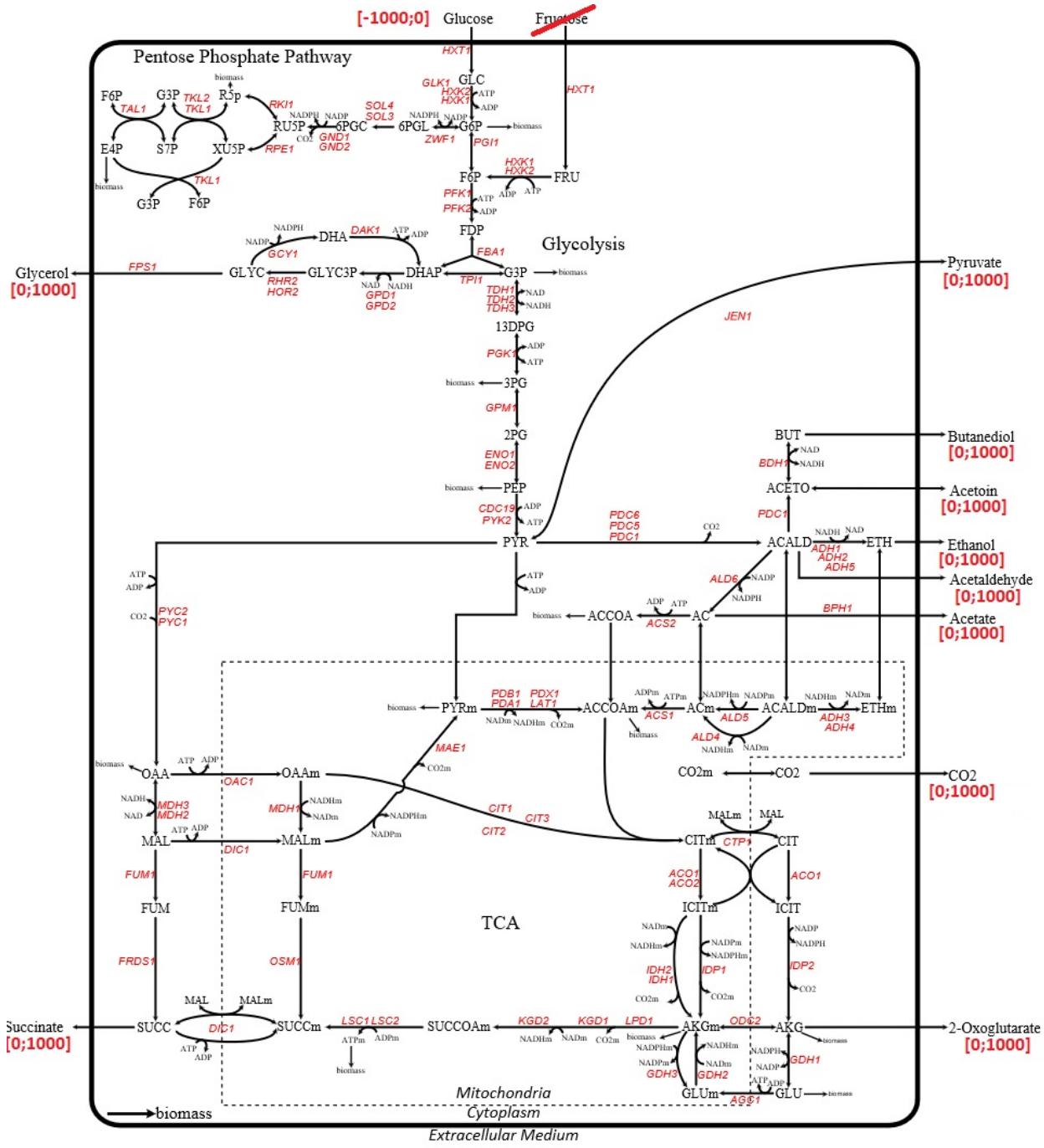


Figure 4.1: Representation of the DynamosYeast model of central carbon metabolism of *S. cerevisiae*. In red are indicated flux constraints for the exchange fluxes. Proteins associated to the reactions are in red capital letters

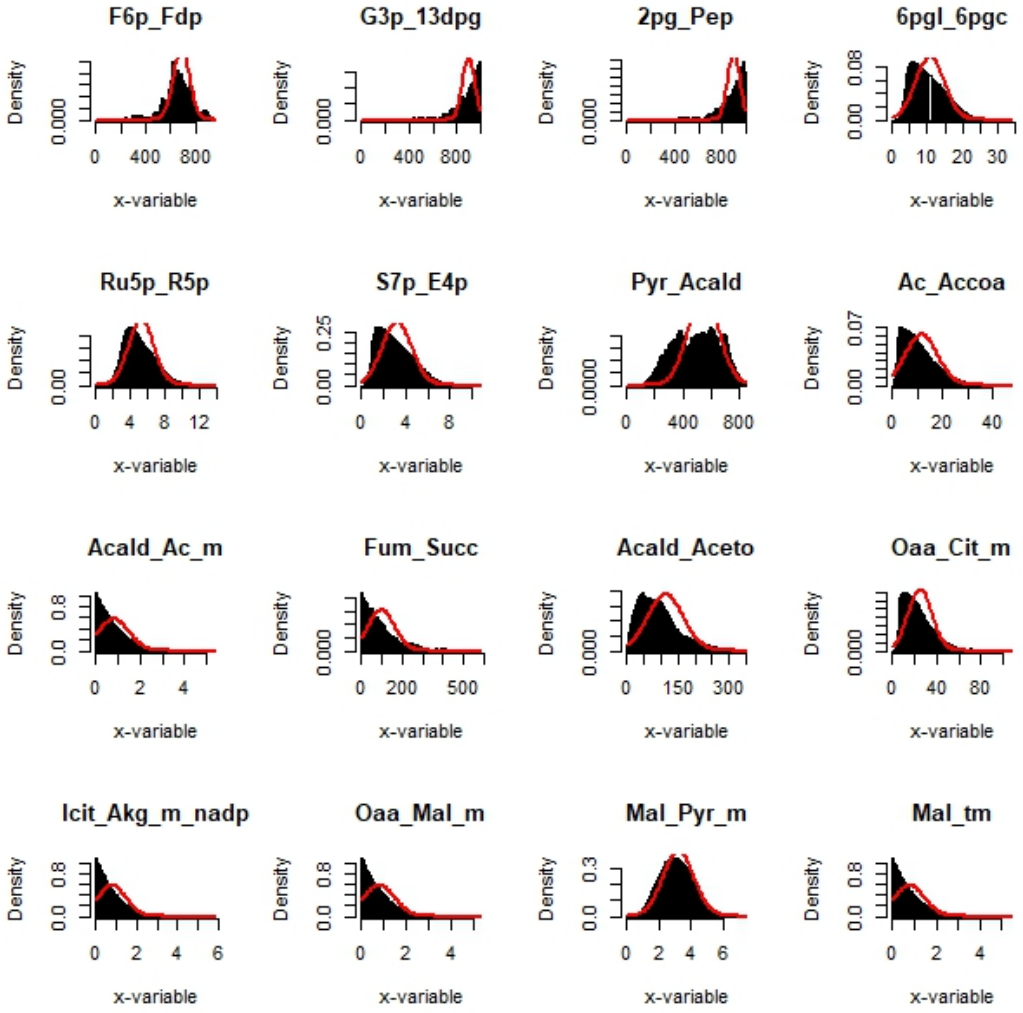


Figure 4.2: Marginal probability densities of sixteen fluxes of the yeast carbon metabolism, randomly chosen. The histograms represent the result of the HR for $T \sim 10^7$ sampling points. The red line is the result of the EP estimate.

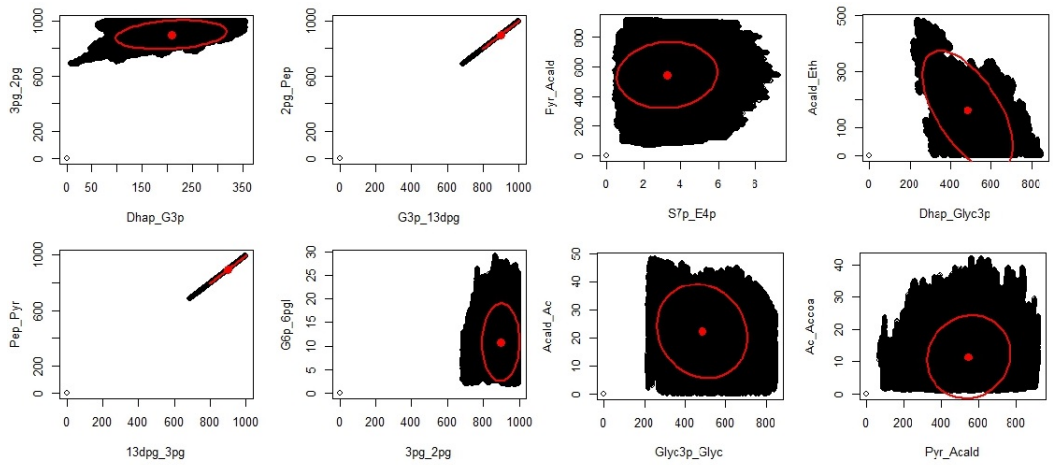


Figure 4.3: Comparison of the results of HR *versus* EP. The plot shows the relation between eight pairwise fluxes. Correlation ellipses, computed by the EP algorithm, are drawn in red. Dot points represent the mean value of fluxes computed through EP. HR sampling points: $T \sim 5 \cdot 10^6$.

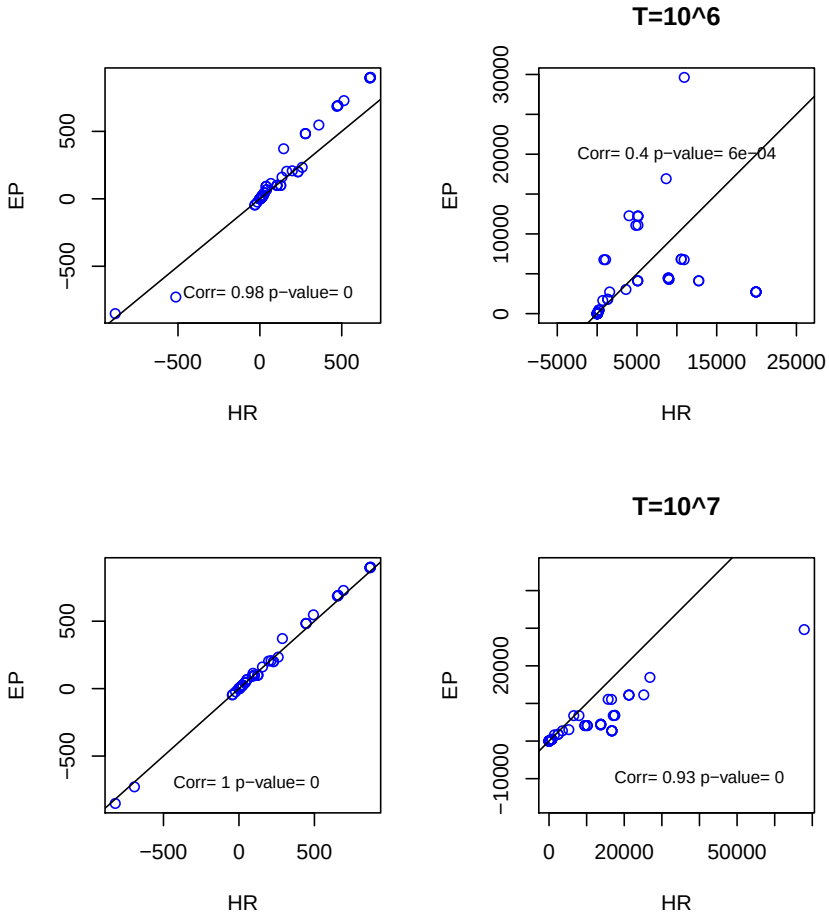


Figure 4.4: Comparison of the results of HR *versus* EP. The plots on the right are scatter plots of the means and on the left variances of the approximated marginals computed via EP against the ones estimated via HR for an increasing number of explored configurations T, top $T \sim 10^6$, bottom $T \sim 10^7$.

Figure 4.4 shows the relation between means and variances estimated through EP and HR for different number of iterations sampled points in L. It shows that the correlation between means and variances estimated through the two methods increase as the number of the HR samples increases. Assuming that the HR algorithm returns the true distribution of fluxes, it is easy to see that means are well predicted by the EP algorithm, although variances are underestimated.

We further investigated if the EP algorithm well predicted the variance-covariance matrix between the DynamoYeast fluxes. Figure 4.3 shows the relation between eight pairwise fluxes randomly chosen. Correlation ellipses (red curve) have been obtained through the EP algorithm. As can be seen, the EP algorithm well predicts the variance-covariance matrix between fluxes satisfying eq. 4.1 and 4.2, on the basis of the HR predictions.

4.3.2 Constraining the solution space with experimental data

Nidelet et al. (2016) have analyzed the diversity of metabolic fluxes of 43 yeast strains from *S. cerevisiae* from six different ecological origins, grown in wine fermentation conditions. Typical wine fermentation comprises a lag phase, a growth phase of approximately 24–36 h followed by a stationary phase, during which most of the sugar is fermented. In the study, production of biomass and metabolites, including ethanol, glycerol, acetate, succinate, pyruvate and alpha-ketoglutarate

were measured during the growth phase (at 11 g/L CO_2 released), which can be considered as steady state (Table 4.2).

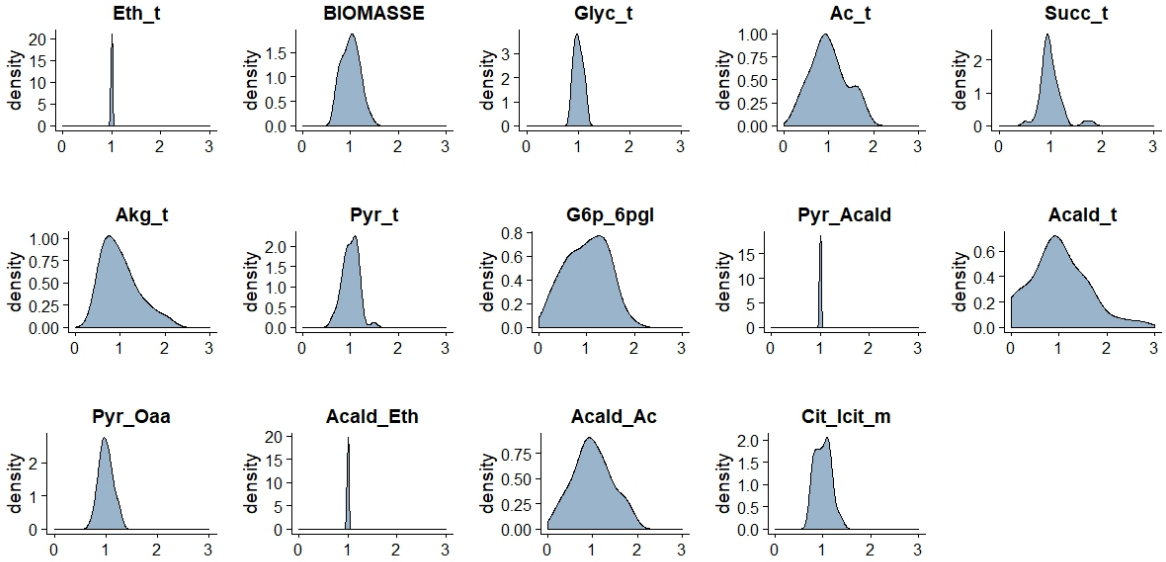


Figure 4.5: Between-strain variations for 14 fluxes from central carbon metabolism in yeast. For each of 47 strains, the fluxes were predicted by minimizing glucose uptake rate and constraining the observed exchange fluxes around their experimental observation. Fluxes are normalized by the average flux of each reaction, and represented by a value between 0 and 3, where 1 is the average flux. Reactions with the subscript "_t" correspond to transport reactions.

Measured exchange fluxes have been introduced as additional constraints of the constraint-based DynamosYeast model, and fluxes known to be irreversible in the context of fermentation have been bounded in just one direction. Further, butanediol and acetoin formation fluxes, *Aceto_But* and *Acald_Aceto*, were set to 0. Finally, under mass balance and steady state assumptions ($\dot{\boldsymbol{m}} = \mathbf{0}$), fluxes have been predicted through minimization of glucose uptake rate.

In order to check the DynamosYeast model, we reproduced the data from Nidelet et al. (2016) and predicted internal fluxes for each of the 43 yeast strains, using as additional constraints the observations. Figure 4.5 shows a schematic representation of the variability of predicted fluxes between the 43 strains. Most fluxes, including the biomass pseudo-flux, show a wide range of variation among strains, except for the glycolysis and ethanol synthesis pathways (Nidelet et al., 2016).

We also reproduced two figures from Nidelet et al. (2016). Figure 4.6 shows the between-strains coefficient of variation, that confirm that all strains seem to be optimized for glycolysis and ethanol production, while the most variable pathway was the pentose-phosphate.

Moreover, the analysis of pairwise correlations between fluxes revealed two antagonistic ways of functioning of central carbon metabolism (figure 4.7). Mitochondrial fluxes are positively correlated to each other, and positively correlated with pentose phosphate pathway, while glycolysis fluxes are positively correlated, and negatively correlated to mitochondrial fluxes. High glycolysis is associated with high biomass production.

Name	Origin	Succinate	Glycerol	Acetate	Pyruvate	AKG	Ethanol	Biomass
6464	Bread	2.29	27.04	2.31	0.87	0.20	276.56	1.24
CBS1171	Bread	1.58	26.51	1.47	0.99	0.13	256.72	1.66
CLIB215	Bread	2.34	26.00	0.98	1.01	0.28	270.48	2.02
CLIB215_3B	Bread	2.56	29.66	1.92	1.17	0.26	291.76	1.91
7_7	Flor	1.54	23.92	6.53	0.56	0.15	272.35	1.44
F25	Flor	0.87	22.68	6.38	0.49	0.11	251.43	1.22
FS2D	Flor	1.47	24.21	6.63	0.68	0.24	277.12	1.12
GUF54_A1	Flor	1.87	23.48	6.77	1.03	0.26	280.77	1.07
MJ73	Flor	1.63	23.37	6.02	0.82	0.32	272.92	1.63
P3_D5	Flor	1.49	20.48	4.77	0.85	0.19	272.30	1.26
TA12_2	Flor	1.66	22.88	4.76	0.96	0.27	277.92	1.50
TS12_A7	Flor	3.03	21.97	5.70	0.72	0.24	268.66	1.71
VPDN_Fino	Flor	3.24	21.63	3.23	0.84	0.26	269.14	1.13
OakR3	MedOak	1.69	27.16	4.42	1.05	0.17	293.53	1.60
ZP848	MedOak	1.88	26.75	4.75	1.04	0.16	277.65	2.09
ZP851	MedOak	1.75	26.54	3.56	1.02	0.16	276.84	1.68
OakA11	Oak	1.36	28.64	7.12	1.00	0.12	279.12	2.11
OakB21	Oak	1.45	27.38	6.74	1.00	0.12	270.88	2.13
ZP1050	Oak	1.93	27.34	5.54	0.92	0.16	297.84	1.71
ZP611	Oak	1.85	26.96	5.03	0.97	0.26	291.79	2.08
245	Rum	1.59	24.68	4.08	0.63	0.17	257.83	1.97
309	Rum	2.16	26.71	4.36	0.80	0.30	278.00	1.87
460	Rum	2.08	23.94	3.93	0.97	0.27	277.96	1.97
390_D2	Rum	2.05	24.37	3.95	0.94	0.21	276.10	1.38
CBS7957	Rum	1.79	22.50	3.70	0.91	0.15	271.30	1.70
CBS7959	Rum	1.68	22.63	4.75	0.98	0.19	282.28	1.35
EDV493	Rum	2.11	21.82	3.17	0.76	0.15	277.20	1.66
1014_F5	Wine	2.15	25.21	3.24	1.33	0.49	285.37	1.22
20B2	Wine	1.76	21.68	2.47	0.91	0.40	272.31	1.71
22A4	Wine	1.89	21.66	2.57	0.82	0.46	253.04	1.73
6320_A7	Wine	1.94	23.27	3.68	1.06	0.44	273.04	2.21
D47_6	Wine	2.03	23.18	1.88	0.76	0.20	324.46	1.87
EC1118	Wine	1.83	22.90	3.95	0.81	0.15	295.43	2.20
F12_3B	Wine	2.23	24.77	2.96	0.86	0.23	287.49	1.95
GE7_4A	Wine	1.97	22.07	2.31	0.84	0.22	309.87	1.44
K1_28_1A	Wine	2.37	21.58	2.05	0.83	0.39	282.86	1.34
L1414	Wine	1.66	22.15	3.39	0.86	0.36	292.66	1.94
Lava32_15	Wine	1.88	26.35	4.02	0.87	0.23	301.04	2.00
Lava32_6	Wine	1.77	22.99	3.23	0.86	0.30	284.45	1.61
M15-3B	Wine	1.89	21.88	3.61	0.79	0.20	295.00	1.46
MC10	Wine	1.85	22.22	3.49	0.96	0.31	289.20	1.62
MC3C	Wine	2.06	22.84	3.11	1.08	0.39	298.98	2.00
N15_4	Wine	1.91	21.43	1.20	0.76	0.35	280.96	1.72

Table 4.2: External metabolite and biomass fluxes measured for 43 yeast strains from different origins (Nidelet et al., 2016).

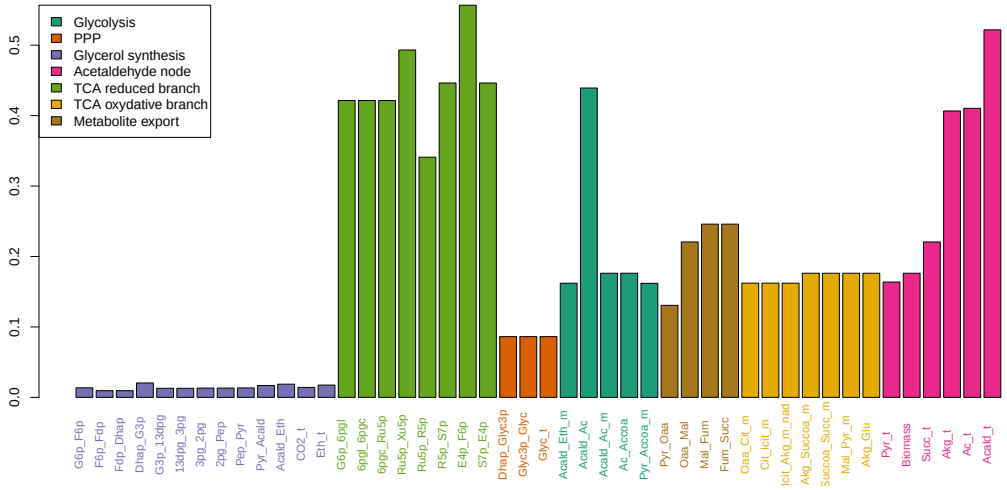


Figure 4.6: Barplot of between-strain coefficients of variation. The coefficient of variation (ratio of the standard deviation to the mean) of each flux is represented as a vertical bar. The vertical bars are ordered by metabolic pathways: glycolysis and ethanol synthesis (blue), PPP (green), glycerol synthesis (orange), acetaldehyde node (blue marine), reductive branch of the TCA (brown), oxidative branch of the TCA (yellow) and output fluxes (violet).

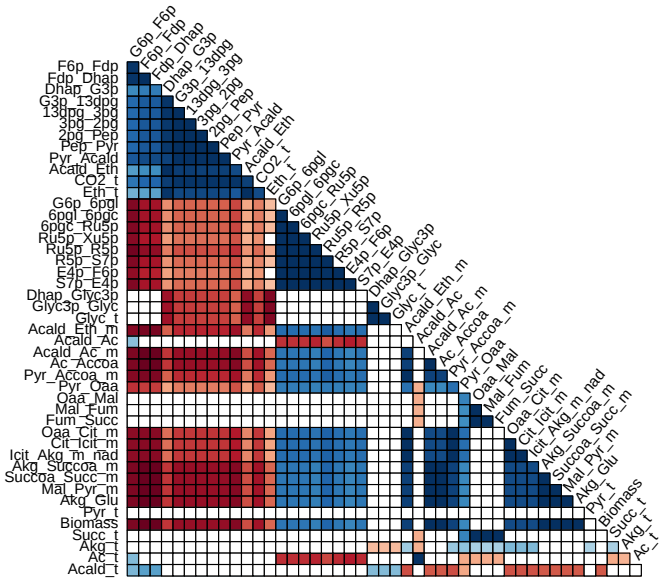


Figure 4.7: Correlation Matrix between internal metabolic fluxes. Pearson correlation values between each pair of fluxes are represented as gradient of colors from red, -1, to blue, $+1$. Fluxes belonging to the same pathway generally group together.

? Question

What if we constrain the flux solution space with the exchange fluxes only?

4.3.3 FBA solution versus the EP distribution

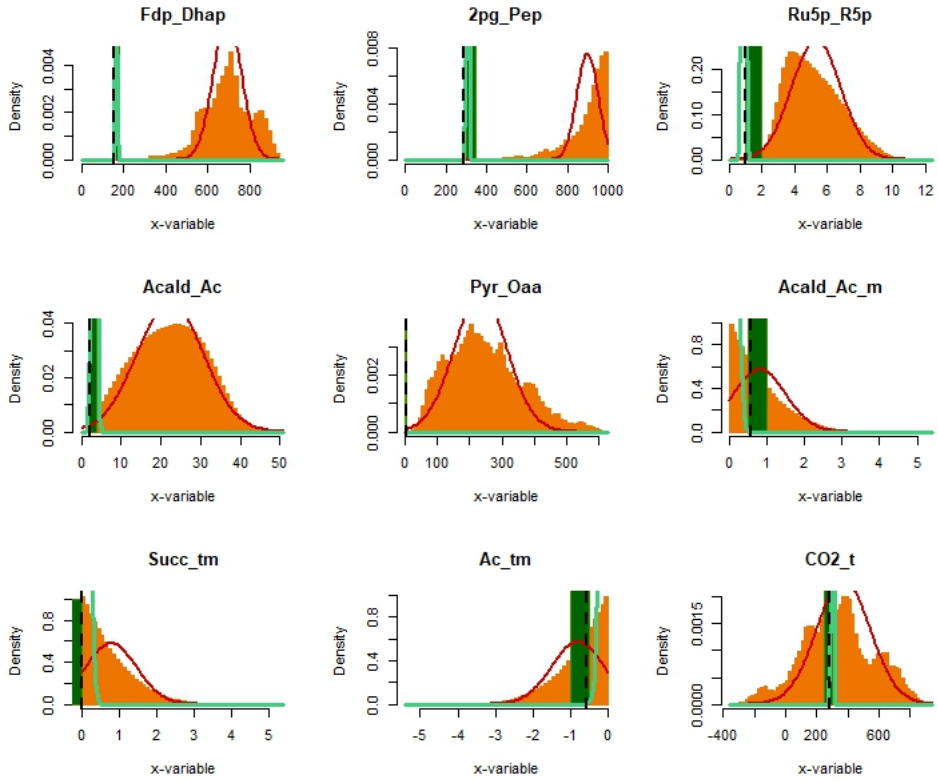


Figure 4.8: Probability distributions of the feasible solution space. In red (resp. orange) is indicated the null posterior distribution of fluxes through the EP algorithm (resp. HR sampling) when no experimental data is introduced; in light green (resp. dark green) the posterior distribution of fluxes through the EP algorithm (resp. HR sampling) when exchange fluxes are constrained by experimental observations. Dashed black line indicates the FBA solution obtained through minimization of glucose uptake, given the experimental observations.

HR and EP algorithms can be used to explore the probability distribution of the feasible solution space, with or without experimental observations, instead of minimizing an objective function. We used the DynamoYeast model to compute four probability distributions of fluxes (figure 4.8):

1. The feasible solution space computed with the HR algorithm, hereafter called "null distribution" (in orange in the figure).

2. Null posterior distribution obtained with the EP algorithm (red).

3. Posterior distribution obtained with the HR algorithm after constraining the range of variation of observed exchange fluxes (dark green).

4. Posterior distribution obtained with the EP algorithm after constraining the range of variation of observed exchange fluxes (light green).

All four distributions were compared to the FBA solution found in Nidelet et al. (2016) by minimizing glucose uptake and constraining the range of variation of exchange fluxes by the observations (black dotted line).

Figure 4.8 shows typical results obtained after a simulation run for one of the 43 yeast strains. Again, we show that the null distributions found by the EP algorithm are consistent with the

ones proposed by the HR algorithm. When constraining with the observed exchange fluxes, a much smaller range of the null feasible space is explored (compare the green distributions to the red ones). This tells us that the observation of exchange fluxes fully constrain the functioning of central carbon metabolism. Unsurprisingly, the CO_2 flux, which is directly observed, is correctly predicted.

The distributions of the HR solutions constrained by the observations appear as rectangles in figure 4.8. In order to better discriminate between the probabilities within the constrained range of variation, we would need to increase the total number of iterations. On the contrary, the EP algorithm provides a full probabilistic distribution of the constrained feasible space at a low computational cost.

Amazingly, the observed CO_2 flux is close to the *a posteriori* mode of the null EP distribution of the feasible space. In more than 50% of cases, the CO_2 flux can be higher than the observed fluxes. This shows that there could be other modes of cell functioning that would lead to higher rates of transformation of glucose into CO_2 and energy.

In all constrained cases, one can compare the mode of the *a posteriori* distribution (in green) to the FBA solution (dark dashed lines) (Figure 4.8). For most fluxes, the FBA solution does not correspond to the *a posteriori* mode of the constrained EP distribution. Remember that the constrained *a posteriori* distribution reflects all possible fluxes leading to exchange fluxes comparable to the observations. Among them, the FBA solution is the one corresponding to the lower consumption rate. Hence, all other solutions correspond to higher glucose consumption rate. The position of the FBA solutions within the posterior distribution is interesting. For the $Ru5p\_R5p$ reaction (pentose-phosphate), the FBA solution is at the right of the constrained distribution. Hence, lower fluxes in the pentose-phosphate pathways could lead to the same observations at the price of a higher glucose consumption rate. This suggests that pentose-phosphate pathway helps producing energy while saving resources. On the contrary, mitochondrial transport of succinate ($Succ\_tm$) and acetate (Ac_tm) shows a FBA solution at leftmost of the constrained solution space. Further comparisons, with alternative objective functions would be interesting in the future to better understand metabolic choices of living species.

Altogether, this study confirmed that it is possible to use the EP algorithm to find feasible ranges of non observed fluxes, once constraining the CBM with observations.

4.4 Conclusion

The EP algorithm is likely to give a good approximation for the posterior joint distribution of fluxes of the DynamosYeast model. In the following, we used this algorithm to predict unobserved metabolic fluxes for each strain per temperature combination from the HeterosYeast dataset through integration of proteomic data. In the HeterosYeast data set, we could not propose an objective function to minimize. Furthermore, the only observed exchange flux was the CO_2 flux. In Chapter 5, we propose a new method based on profile comparison (Lee et al., 2012) to integrate proteomic data information into the CBM.

Bibliography

Antoniewicz, M. (2015). Methods and advances in metabolic flux analysis: a mini-review., *J Ind Microbiol Biotechnol.* **42**(3): 317–25.

Becker, S. A. and Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments, *PLoS computational biology* **4**(5): e1000082.

Bordbar, A., Monk, J. M., King, Z. A. and Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions, *Nature Reviews Genetics* **15**(2): 107–120.
URL: *https://www.nature.com/articles/nrg3643*

Braunstein, A., Muntoni, A. P. and Pagnani, A. (2017). An analytic approximation of the feasible space of metabolic networks, *Nature Communications* **8**: 14915.
URL: *http://www.nature.com/doifinder/10.1038/ncomms14915*

Bélisle, C. J. P., Romeijn, H. E. and Smith, R. L. (1993). Hit-and-Run Algorithms for Generating Multivariate Distributions, *Mathematics of Operations Research* **18**(2): 255–266.
URL: *https://www.jstor.org/stable/3690278*

Celton, M., Goelzer, A., Camarasa, C., Fromion, V. and Dequin, S. (2012). A constraint-based model analysis of the metabolic consequences of increased NADPH oxidation in Saccharomyces cerevisiae, *Metabolic Engineering* **14**(4): 366 – 379.

Colijn, C., Brandes, A., Zucker, J., Lun, D., Weiner, B., Farhat, M., Cheng, T., Moody, D., Murray, M. and Galagan, J. (2009). Interpreting expression data with metabolic flux models: predicting mycobacterium tuberculosis mycolic acid production., *PLoS Comput Biol.* **5**(8): e1000489.

Fell, D. and Small, J. (1986). Fat synthesis in adipose tissue. an examination of stoichiometric constraints, *Biochem J.* **238**(3): 781–786.

Gudmundsson, S. and Thiele, I. (2010). Computationally efficient flux variability analysis, *BMC Bioinformatics* **11**: 489.

Lee, D., Smallbone, K., Dunn, W. B., Murabito, E., Winder, C. L., Kell, D. B., Mendes, P. and Swainston, N. (2012). Improving metabolic flux predictions using absolute gene expression data, *BMC Systems Biology* **6**(1): 73.
URL: *https://doi.org/10.1186/1752-0509-6-73*

Lobel, L., Sigal, N., Borovok, I., Ruppin, E. and Herskovits, A. A. (2012). Integrative Genomic Analysis Identifies Isoleucine and CodY as Regulators of Listeria monocytogenes Virulence, *PLoS Genetics* **8**(9): e1002887.
URL: *https://dx.plos.org/10.1371/journal.pgen.1002887*

Meersche, K. V. d., Soetaert, K. and Oevelen, D. V. (2009). xsample() : An *R* Function for Sampling Linear Inverse Problems, *Journal of Statistical Software* **30**(Code Snippet 1).
URL: *http://www.jstatsoft.org/v30/c01/*

Nicholson, J. and Lindon, J. (2008). Systems biology: Metabonomics, *Nature* **455**(7216): 1054–6.

Nidelet, T., Brial, P., Camarasa, C. and Dequin, S. (2016). Diversity of flux distribution in central carbon metabolism of S. cerevisiae strains from diverse environments, *Microbial Cell Factories* **15**(1).
URL: *http://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-016-0456-0*

Patil, K. R. and Nielsen, J. (2005). Uncovering transcriptional regulation of metabolism by using metabolic network topology, *Proceedings of the National Academy of Sciences of the United States of America* **102**(8): 2685–2689.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC549453/*

Riekeberg, E. and Powers, R. (2017). New frontiers in metabolomics: from measurement to insight., *F1000Research* **6**: 1148.

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. and Sauer, U. (2012). Multidimensional Optimality of Microbial Metabolism, *Science* **336**(6081): 601–604.
URL: *https://science.sciencemag.org/content/336/6081/601*

Töpfer, N., Caldana, C., Grimbs, S., Willmitzer, L., Fernie, A. and Nikoloski, Z. (2013). Integration of genome-scale modeling and transcript profiling reveals metabolic pathways underlying light and temperature acclimation in arabidopsis., *Plant Cell* **25**(4): 1197–211.

Varma, A. and Palsson, B. O. (1994). Metabolic flux balancing: basic concepts, scientific and practical use., *Nat. Biotechnol.* **12**: 994–998.

Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E. and Shlomi, T. (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model, *Bioinformatics (Oxford, England)* **26**(12): i255–260.

Chapter 5

Data integration uncovers the metabolic bases of phenotypic variation in yeast

Marianyela Petrizzelli*, Dominique de Vienne*, Thibault Nidelet and Christine Dillmann*,[1]**

*Génétique Quantitative et Evolution – Le Moulon, INRA, Université Paris-Saclay, Université Paris-Sud, CNRS, AgroParisTech, 91190 Gif-sur-Yvette, France, **SPO, INRA, Montpellier SupAgro, Univ Montpellier, Montpellier, France.

ABSTRACT The relationships between levels of integration, from gene expression to complex phenotypic traits, are a key feature for understanding the genotype-phenotype map. We proposed a novel method that incorporates protein abundance data into constraint-based modelling to elucidate the biological mechanisms underlying phenotypic variation. In particular, we studied yeast genetic diversity at three levels of phenotypic complexity: protein abundances, metabolic fluxes and life-history/fermentation traits. Protein abundances and life-history/fermentation traits were measured in a population obtained by pairwise crossing of strains representative of two yeast species, *Saccharomyces cerevisiae* and *S. uvarum*. Metabolic fluxes were estimated using Bayesian models by constraining a metabolic model of central carbon metabolism with measured abundances of involved enzymatic proteins. At the trait level, there were weak differences between species along with high diversity within species, and a strong negative correlation was observed between production traits like population carrying capacity (K) and traits associated to growth and fermentation rates ($J_{\max}$). The metabolic fluxes predicted from protein abundance variations revealed that this negative correlation is sustained by a differential usage of energy production pathways: TCA cycle and glycolysis. In addition, we identified protein sets that confirmed that high K was associated with high TCA fluxes, respiration and energy conversion levels, while high $J_{\max}$ was associated with high glycolytic fluxes, fermentation and protein recycling. By coupling phenomic data with mathematical modeling of metabolism, we explained the trade-off between two yeast life-history traits, K and $J_{\max}$, by a differential pathway usage for the production of energy reserves and cellular compounds.

KEYWORDS Data integration; Genotype-Phenotype map; Metabolism modelling; life-history trade-offs

Introduction

Phenotypic diversity within the living world results from millions of years of evolution. Most evolutionary pressures like mutation, random genetic drift, migration or recombination shape phenotypic diversity by directly changing the genetic composition of populations. The effects of selection are more difficult to predict because the fittest individuals are picked out from their phenotype, which results from a complex interaction between genotype and environment (Fisher 1930). An additional layer of complexity results from the fact that life-history traits (Stearns 1992) are the observable results of unobservable processes that occur at the cellular scale. During the last decades, there has been a growing interest for a better understanding of the so-called genotype-phenotype map in evolutionary biology (see *e.g.* Wagner and Zhang (2011)). In parallel, novel profiling technologies and accurate high throughput phenotyping strategies have led to the genome-scale characterization of genomic sequences as well as to the quantification of transcriptomic, proteomic and metabolomic data at the individual level. Linking cellular processes to observable phenotypic traits is becoming a new discipline in Biology, known as integrative biology.

Unicellular organisms are choice model species for integrative biology because most observable traits are direct products from cell metabolism, without the complications of the tissue and organ levels that need to be taken into account in multicellular species. Schematically, cells sense the environment and transfer the information *via* signal transduction chains that interact

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Tuesday 11th June, 2019
[1]To whom correspondence should be addressed: UMR Génétique Quantitative et Évolution - Le Moulon, Ferme du Moulon, 91190 Gif-sur-Yvette, France. E-mail: christine.dillmann@inra.fr

with the gene regulation network. The gene regulatory network modulates transcription, translation and post-translational modifications according to environmental signals, which results in variations of protein abundances. Differential abundances of enzymatic proteins affect the fluxes of matter and energy that are related to phenotypic traits, including life-history traits and fitness. Thus, in unicellulars, five integration levels are usually considered: genomic, transcriptomic, proteomic (including post-translational modifications), metabolic and cellular or observable trait level. The last level is the most integrated, and it encompasses a variety of traits more or less related to fitness.

While genomic, transcriptomic, proteomic and trait levels are now readily measurable on numbers of individuals thanks to technical progresses, metabolic fluxes are still difficult to measure. Metabolic Flux Analysis is powerful (Antoniewicz 2015). However, it is based on RMN and differential usage of radioactive isotopes. It remains low-throughput and cannot be applied on numerous individuals. Technical developments in mass spectrometry popularized metabolomics (Nicholson and Lindon 2008), which allowed to characterize the metabolome, that represents the complete set of metabolites in a cell, tissue, organ or organism. However, the technique still suffers from standardization procedures and does not allow for high-throughput quantitative comparisons (Riekeberg and Powers 2017).

Taking advantage from the recent progresses in genome-scale functional annotation, constraint-based metabolic models provide a mathematical framework that allows predicting internal cellular fluxes from *a priori* knowledge on thermodynamic constraints on individual enzymatic reactions, steady state hypotheses and the genome-scale stoichiometry matrix of all metabolic reactions. The idea is that a given set of environmental conditions will drive a cell to a steady state during which internal metabolites stay at a constant concentration while exchange fluxes are constant and correspond to a constant import/export rate. However, because the number of metabolites is much higher than the number of reactions, the system has an infinite number of solutions. Flux Balance Analysis (Fell and Small 1986; Watson 1984) consists in choosing, among all possible solutions, the one that maximizes the biomass pseudo-flux. From a population geneticist point of view, this method is questionable because evolution is not always based on optimization principles (Gould and Lewontin 1979). However, it was shown to be relevant in some cases, like chemostat culture of *Escherichia coli* (Edwards *et al.* 2001). Data-driven methods have also been proposed, that consist in choosing the most likely solution given observed transcriptomic, proteomic or metabolomic data (see the review by Töpfer *et al.* (2015)). Among all these methods, the one from Lee *et al.* (2012) sounds promising for studies at the population/species level. It is based on the realistic assumption that, at the genome scale, fluxes should covary with enzymatic protein abundances. Whatever the method, comparisons rely on the probability distribution of the solution space, which is analytically untractable because of the soichiometry constraints. Recently, Braunstein *et al.* (2017) proposed a bayesian probabilistic method to characterize the solution space, that proved to be much faster than the classical hit-and-run algorithm (Bélisle *et al.* 1993) and allow for analyses at both genome- and population-scales.

The so-called HeterosYeast project consisted in studying the molecular bases of heterosis in yeast species at two different levels of integration, the proteomic level and the observable trait level (Blein-Nicolas *et al.* 2013, 2015; da Silva *et al.* 2015). A diallel design including two yeast species involved in wine fermentation was realized and the hybrid and parental strains were monitored during fermentation on grape juice at two temperatures. Observable and proteomic traits were analyzed separately. Briefly, the most important findings were homeostasis of the interspecific hybrids observed at the trait level da Silva *et al.* (2015) and the predominance of inter-specific heterosis at the proteomic level Blein-Nicolas *et al.* (2015). A more careful analysis of genetic variance components confirmed that observable phenotypic traits tend to exhibit higher additive genetic variances and lower interaction variances than proteomic traits (Petrizzelli *et al.* 2019). Yet, the link between variation at the trait level and variation at the proteomic level is still missing.

Given the important yeast genomic ressources (Cherry *et al.* 2012a), a number of curated genome-scale metabolic models are now available (Caspi *et al.* 2014). Among those, the DynamoYeast model (Celton *et al.* 2012) describes yeast central carbon metabolism. It is small enough (70 reactions) to remain tractable, and has been tested against experimental data (Nidelet *et al.* 2016).

The availability of the HeterosYeast dataset, of a curated metabolic model of yeast central carbon metabolism and of a probabilistic approach to explore the solution space, encouraged us to integrate the experimental proteomic data in the metabolic model in order to predict unobserved metabolic fluxes. We used predicted fluxes to bridge the gap between proteomic data and observable traits, and better understand the metabolic bases of life-history traits variation.

Material and Methods

Materials

The HeterosYeast dataset. The genetic material of the experimental design consisted in 7 strains of *S. cerevisiae* and 4 strains of *S. uvarum* associated to various food-processes (enology, brewery, cider fermentation and distillery) or isolated from natural environment (oak exudates). The 11 parental lines were selfed and pairwise crossed, which resulted in a half-diallel design with a total of 66 strains: 11 inbred lines, 27 intra-specific hybrids (21 for *S. cerevisiae*, noted *S. c.*, and 6 for *S. uvarum*, noted *S. u.*) and 28 inter-specific (noted *S. u.* $\times$ *S. c*). The 66 strains were grown in triplicate in fermentors at two temperatures, 26°C and 18°, in a medium close to enological conditions (Sauvignon blanc grape juice, da Silva *et al.* (2015)). From a total of 396 alcoholic fermentations (66 strains $\times$ 2 temperatures $\times$ 3 replicas), 31 failed due to poor fermenting abilities of some strains. The design was implemented considering a block as two sets of 27 fermentations (26 plus a control without yeast to check for contamination), one carried out at 26°C and the other at 18°. The distribution of the strains in the block design was randomized to minimize the residual variance of the estimators of the strain and temperature effects, as described in Albertin *et al.* (2013b).

For each alcoholic fermentation, two types of phenotypic traits were measured or estimated from sophisticated data adjustment models: 35 fermentation traits and 615 protein abundances.

The fermentation traits were classified into four categories (da Silva *et al.* 2015):

- *Kinetics parameters*, computed from the CO_2 release curve modeled as a Weibull function fitted on CO_2 release quantification monitored by weight loss of bioreactors: the fermentation lag-phase, *t-lag* (h); the time to reach the inflec-

tion point out of the fermentation lag-phase, *t-V_{max}* (h); the fermentation time at which 45 gL^{-1} and 75 gL^{-1} of CO_2 was released, out of the fermentation lag-phase, *t-45* (h) and *t-75* (h) respectively; the time between *t-lag* and the time at which the CO_2 emission rate became less than, or equal to, $0.05gL^{-1}h^{-1}$, *AFtime* (h); the maximum CO_2 release rate, $V_{\max}$ ($gL^{-1}h^{-1}$); and the total amount of CO_2 released at the end of the fermentation, $CO_{2\max}$ (gL^{-1}).

- *Life history traits*, estimated and computed from the cell concentration curves over time, modeled from population growth, cell size and viability quantified by flow cytometry analysis: the growth lag-phase, *t-$N_0(h)$*; the carrying capacity, *K* (log[cells/mL]); the time at which the carrying capacity was reached, *t-N_{max}* (h); the intrinsic growth rate, *r* (log[cell division/mL/h]); the maximum value of the estimated CO_2 production rate divided by the estimated cell concentration, J_{max} ($gh^{-1}10^{-8}cell^{-1}$); the average cell size at *t-N_{max}*, *Size-t-N_{max}*(μm); the percentage of living cells at *t-N_{max}*, *Viability-t-N_{max}* (%); and the percentage of living cells at *t-75*, *Viability-t-75* (%).
- *Basic enological parameters*, quantified at the end of fermentation: *Residual Sugar* (gL^{-1}); *Ethanol* (%vol); the ratio between the amount of metabolized sugar and the amount of released ethanol, *Sugar.Ethanol.Yield* (gL^{-1}%vol^{-1}); *Acetic acid* (gL^{-1} of H_2SO_4); *Total SO_2* (mgL^{-1}) and *Free SO_2* (mgL^{-1}).
- *Aromatic traits*, mainly volatile compounds measured at the end of alcoholic fermentation by GC-MS: two higher alcohols (*Phenyl-2-ethanol* and *Hexanol*, mgL^{-1}); seven esters (*Phenyl-2-ethanol acetate*, *Isoamyl acetate*, *Ethyl-propanoate*, *Ethyl-butanoate*, *Ethyl-hexanoate*, *Ethyl-octanoate* and *Ethyl-decanoate*, mgL^{-1}); three medium chain fatty acids (*Hexanoic acid*, *Octanoic acid* and *Decanoic acid*, mgL^{-1}); one thiol *4-methyl-4-mercaptopentan-2-one*, *X4MMP*(mgL^{-1}) and the acetylation rate of higher alcohols, *Acetate ratio*.

For proteomic analyses the samples were harvested at 40 % of CO_2 release, corresponding to the maximum rate of CO_2 release. Protein abundances were measured by LC-MS/MS techniques from both shared and proteotypic peptides relying on original Bayesian developments (Blein-Nicolas *et al.* 2012). A total of 615 proteins were quantified in more than 122 strains × temperature combinations as explained in details in Blein-Nicolas *et al.* (2015).

Genetic value of protein abundances and fermentation/life-history traits. In this analysis we considered the genetic values of protein abundances and fermentation/life-history traits, rather than their measured/computed value. In a previous study, Petrizzelli *et al.* (2019) have decomposed the phenotypic values of a trait at a given temperature, P_T, into its genetic, G_T, and residual, ϵ, contributions:

$$P_T = G_T + \epsilon \tag{1}$$

The genetic value, G_T, has been decomposed in terms of additive and interaction effects, taking into account the structure of the half-diallel design. The presence of two different species and of the parental inbreds in the experimental design let them to further distinguish between intra- and inter-specific additive genetic effects ($\boldsymbol{A_w}$ and $\boldsymbol{A_b}$, respectively) and to decompose the interaction effects into inbreeding ($\boldsymbol{B}$) and intra- and inter-specific heterosis effects ($\boldsymbol{H_w}$, $\boldsymbol{H_b}$). Therefore, the genetic value of a trait at a given temperature T has been modeled by:

$$G_T^{p_i} = \mu_T + 2A_{w_i,\,T} + \beta_{s(i),\,T} + B_{i,\,T} \tag{2}$$

$$G_T^{H_{ij}^w} = \mu_T + A_{w_i,\,T} + A_{w_j,\,T} + H_{w_{ij},\,T}, \tag{3}$$

$$G_T^{H_{ik}^b} = \mu_T + A_{b_i,\,T} + A_{b_k,\,T} + H_{b_{ik},\,T}. \tag{4}$$

for a parental strain p_i (eq. 2), for a intra-specific hybrid H_{ij}^w between parents p_i and p_j (eq. 3), and for the inter-specific hybrid H_{ik}^b between parents p_i and p_k (eq. 4). μ is the overall mean and $\beta_{s(i)}$ is the deviation from the fixed overall effect for the species:

$$s(i) \in \{S.\ cerevisiae, S.\ uvarum\}$$

We retrieved the genetic values for all proteomic data. For the fermentation traits, the model did not converge for most of the ethyl esters (*Ethyl-propanoate*, *Ethylbutanoate*, *Ethyl-hexanoate*, *Ethyl-octanoate* and *Ethyl-decanoate*), as well as for *Acetate Ratio* and for *Acetic acid*. These traits were removed from the analysis.

Protein functional annotation

Cross-referencing MIPS micro-organism protein classification (Ruepp *et al.* 2004), KEGG pathway classification (Kanehisa and Goto 2000; Kanehisa *et al.* 2016, 2017) and Saccharomyces Genome database (Cherry *et al.* 2012b), we attributed each protein to a single functional category based on our expert knowledge.

The first two hierarchical levels of MIPS functional annotation have been taken into account to assign proteins into 34 different categories. For *01.metabolism*, *02.energy* and *10.cell cycle and DNA processing* categories all secondary levels were used, resulting in 20 different functional categories. The *11.transcription* category was subdivided in into the *transcription* sub-group (*11.06 and 11.02*) and into the *RNA processing* sub-group (*11.04*). Similarly, *12.protein synthesis* category was split into *ribosomal proteins* (*12.01*) and *translation* (*12.04, 12.07, 12.10*) sub-groups; *20.transport* category into *vacuolar transport* (*20.09*) and *transport* (*20.01, 20.03*) sub-groups.

Instead the first hierarchical category was used for *14.protein fate*, *30.signal transduction*, *32.detoxification*, *34.homeostasis*, *40.cell growth and death*, *42.cytoskeleton* Further, we fused the *16.binding function* and *18.02.regulation* category into *16.binding*, and *32.transposon movement* with *10.01.DNA processing*. Finally, *41.mating* and *43.budding* categories were included in *10.03.cell cycle* category.

DynamosYeast model

We exploited the DynamoYeast model, a previously developed constraint-based model of central carbon metabolism of *S. cerevisiae* (Celton *et al.* 2012). This model includes upper and lower glycolysis, the pentose phosphate pathway (PPP), the synthesis of glycerol, the synthesis of ethanol and the reductive and oxidative branches of the tricarboxylic acid (TCA) cycle as the main metabolic pathways. It consists of 60 metabolites and 70 reactions, including one input flux, the glucose uptake, and 10 output fluxes (Figure 1), taking place in the cytosol, in the mitochondria or in the extracellular medium.

The range of variation of the fluxes was fixed to allow alcoholic fermentation. Therefore, malate dehydrogenase, *Oaa_Mal*, fumarase, *Mal_Fum*, fumarate reductase, *Fum_Succ*, and mitochondrial malate dehydrogenase, *Oaa_Mal_m*, fumarase, *Mal_Fum_m*, fumarate reductase, *Fum_Succ_m* and citrate synthase, *Oaa_Cit_m*, reactions were imposed to be irreversible with $\boldsymbol{v}^{inf} = 0$. Furthermore, fructose flux was not included in the model, and mitochondrial glutamate dehydrogenase, *Glu_Akg_m* as well as butanediol formation, *Aceto_But*

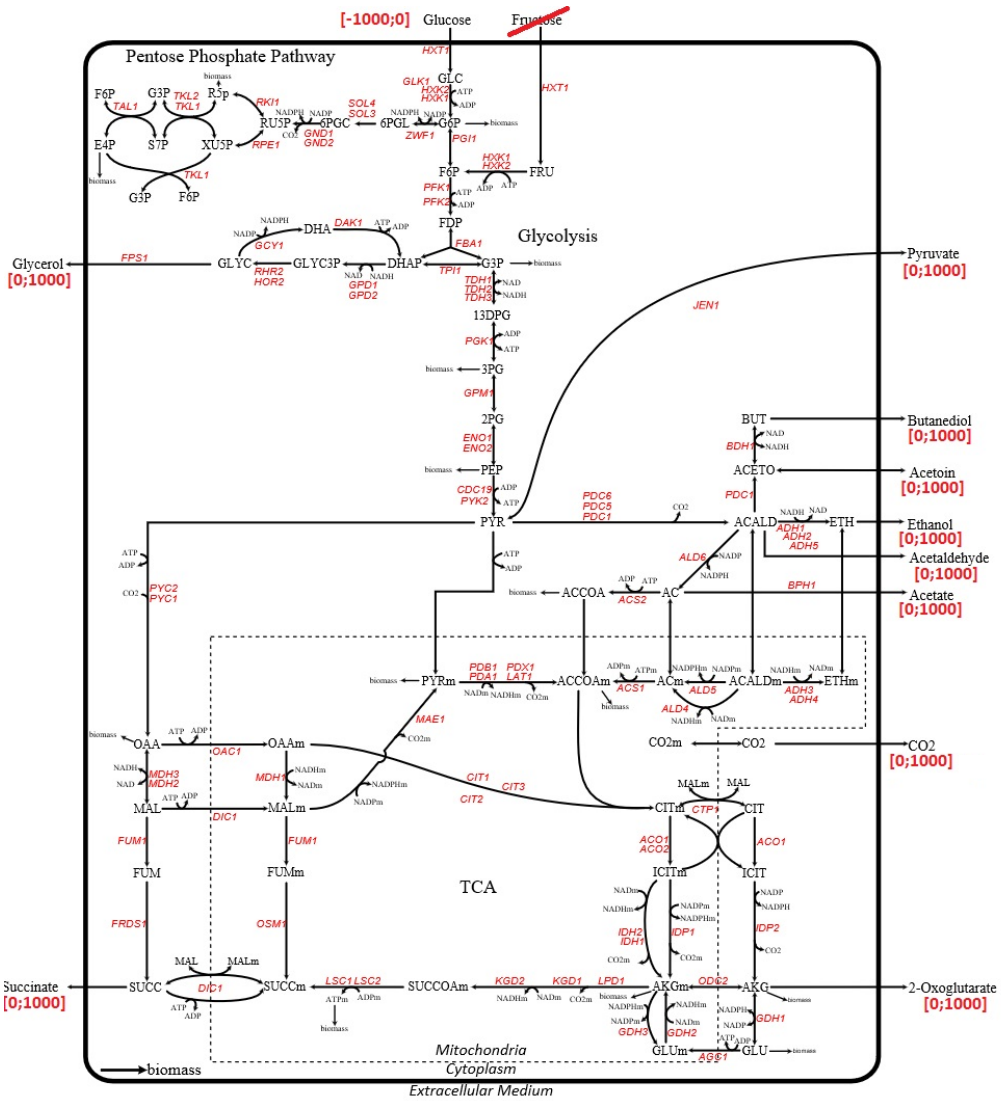


Figure 1 Representation of the DynamosYeast model of central carbon metabolism of *S. cerevisiae*. Metabolites are in black. Names of enzymatic proteins that catalyse the reactions are in red. Constraints on exchange fluxes are in red between square brackets and correspond to fermentation, with glucose as unique input flux.

reactions were set to zero. Overall, there were 16 reversible and 52 irreversible fluxes.

Following the conventions implemented by many genome-scale-metabolic models, many reactions of the DynamoYeast model for central carbon metabolism of *S. cerevisiae* are associated with genes and proteins via gene-protein-reaction (GPR) associations (Thiele and Palsson 2010).

In general, there can be a many-to-many mapping from genes to reactions, for example one reaction can be linked to protein ($P1$ and $P2$) or $P3$. The first Boolean AND relationship means that the reaction is catalyzed by a complex between two gene products. Since the maximum of the complex is given by the minimum of its components, the weighting of the complex is defined as: $P1$ AND $P2 = \min(P1, P2)$. The OR relationship allows for alternative catalysts to the reactions. As such total capacity is given by the sum of its components: ($P1$ AND $P2$) OR $P3 = \min(P1, P2) + P3$ (Lee *et al.* 2012). Following these rules, for each of the 11 yeast strains and the 55 hybrids at both temperatures, we estimated the protein abundances associated to the reactions in the DynamosYeast model, leading to a total of 33 reactions weightings out of 70.

Methods

Constraint-based modeling of metabolic networks

Metabolic networks can be described in terms of the relations between M metabolites, $\boldsymbol{m}$, and N reactions, $\boldsymbol{v}$, at a given time t:

$$(\boldsymbol{v}, \boldsymbol{m})_t$$

Their topology can be expressed through the $M \times N$ stoichiometric matrix $\boldsymbol{S}$, in which rows correspond to the stoichiometric coefficients of the corresponding metabolites in all reactions.

Under mass-balance assumption and thermodynamic bounds of reaction rates, the dynamics of the network is governed by

the linear system of constraints and inequalities:

$$S\boldsymbol{v} = \dot{\boldsymbol{m}} \tag{5}$$

$$\boldsymbol{v}^{inf} \leq \boldsymbol{v} \leq \boldsymbol{v}^{sup} \tag{6}$$

where $\dot{\boldsymbol{m}} \in \mathbb{R}^M$ is the vector of the M input/output rates of metabolites, $\boldsymbol{v} \in \mathbb{R}^N$ is the set of N reactions, and $\boldsymbol{v}^{inf}$, $\boldsymbol{v}^{sup}$ are the extremes of variation of the set of fluxes.

Under steady state assumption, $\dot{\boldsymbol{m}} = 0$ and the feasible space of solutions is expressed as:

$$L \equiv \{\boldsymbol{v} \in \mathbb{R}^N | S\boldsymbol{v} = 0,\ \boldsymbol{v}^{inf} \leq \boldsymbol{v} \leq \boldsymbol{v}^{sup}\} \tag{7}$$

In general, N is larger than M and the solution space L has infinite cardinality.

Prediction of the feasible space of solutions

We propose to characterize the feasible space of solutions L through the posterior probability of flux values obtained by the Expectation Propagation (EP) model described in Braunstein *et al.* (2017).

Instead of exploring L through sampling, as classical methods do, Braunstein *et al.* (2017) have proposed to combine statistical physics and Bayesian approaches to infer the joint distribution of metabolic fluxes. To do so, given the set of metabolite input/output rates, $\dot{\boldsymbol{m}}$, they encoded the stoichiometric constrains, within the likelihood posterior probability, defining a Boltzmann-like distribution with energetic quadratic function

$$\mathcal{E}(\boldsymbol{v}) = \frac{1}{2}(S\boldsymbol{v} - \dot{\boldsymbol{m}})^\top (S\boldsymbol{v} - \dot{\boldsymbol{m}}) \tag{8}$$

while the inequality constrains were encoded in the prior probability of fluxes. Via the Bayes theorem, this method provided a model for posterior density of flux distribution.

Therefore, each point $\boldsymbol{v}$ in L follows the truncated multivariate normal distribution

$$\forall \boldsymbol{v} \in L;\ \boldsymbol{v} \sim \mathcal{N}_T(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{v}^{inf}, \boldsymbol{v}^{sup}, \dot{\boldsymbol{m}}) \tag{9}$$

where $\boldsymbol{\mu}$ is the vector of the mean posterior values of fluxes and $\boldsymbol{\Sigma}$ the posterior variance-covariance matrix of fluxes estimated through the EP algorithm.

This formalism allows associating to each set of metabolic fluxes $\boldsymbol{v}$ its posterior probability of being observed

$$p_{\boldsymbol{v}} = P(\boldsymbol{v} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{v}^{inf}, \boldsymbol{v}^{sup}, \dot{\boldsymbol{m}}) \tag{10}$$

Different values for the extremes of variation can be supplied to model a particular process, for example for modeling reactions known to be irreversible in a specific context, *i. e.*

$$v_i^{inf} = 0 \text{ or } v_i^{sup} = 0$$

or for introducing experimental data constrains, *i. e.*

$$v_i = v_i^{obs} \pm \epsilon$$

for the i-th reaction.

Given that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ depend on the imposed range of internal and exchange fluxes, $\boldsymbol{v}^{inf}$, $\boldsymbol{v}^{sup}$, metabolic fluxes will take particular values with probabilities that depend on *a priori* knowledge and on the chosen metabolic processes.

The algorithm implemented in Braunstein *et al.* (2017) was translated into *R* code. Extraction of the stoichiometric matrix from the DynamosYeast model have been performed with the *sybil* package in R (Gelius-Dietrich *et al.* 2013).

Prediction of metabolic fluxes from proteomic data

In living systems, most metabolic reactions are catalyzed by enzymes, and quantitative proteomic data retain information about enzyme abundancies. Therefore, the metabolism of a cell, at a given time, is characterized by the set of fluxes, of metabolites and of protein abundances

$$(\boldsymbol{v}, \boldsymbol{m}, \boldsymbol{E})_t$$

where $\boldsymbol{E} = (E_1, E_2, \ldots, E_N)$, and E_i is the abundance of enzyme i associated with the reaction flux v_i. Indeed, even though reaction rates are not directly proportional to enzyme abundancies, a certain covariation between protein abundances and flux reaction rates is expected at the metabolic network scale. It can be used to infer intracellular metabolic fluxes with reasonable accuracy (Lee *et al.* 2012).

Among all possible solutions from the feasible space L, we proposed to choose the one that minimizes the objective function:

$$Z = \frac{1}{p_{\boldsymbol{v}}} \sum_{i=1}^{N} (E_i - |v_i|)^2 \tag{11}$$

i.e. the Euclidean distance between the quantified abundance of proteins $\boldsymbol{E}_{obs}$ and the associated fluxes, weighted by $p_{\boldsymbol{v}}$, the posterior probability of observing the set of metabolic fluxes $\boldsymbol{v}$.

The properties of the truncated multivariate normal distribution ensure that the solution of the objective function is unique and no sophisticated algorithm is needed to find this solution. For each set of observation $\boldsymbol{E}_{obs}$, we proposed to sample N_s points of the feasible space of solutions. Therefore, $\forall k \in \{1, 2, 3 \ldots N_s\}$, we got $\boldsymbol{v}^k \in L$ and $p_{\boldsymbol{v}^k}$. We calculated $Z^{(k)}$ and selected the set of flux values, $\boldsymbol{v}^{predicted}$, for which $Z^{(k)}$ was the minimum.

In practice, it is never possible to associate each reaction of the metabolic network with a protein abundance. First, quantitative proteomics is not exhaustive. Second, reactions of a metabolic model are not always associated with an enzyme. Assuming steady state condition and introducing information about protein abundances and measured external metabolic fluxes allows to describe the system as:

$$(\mathbb{1}_{obs}\boldsymbol{v} + \mathbb{1}_{\overline{obs}}\,\boldsymbol{v}, \boldsymbol{m}_{const}, \mathbb{1}_{obs}\boldsymbol{E} + \mathbb{1}_{\overline{obs}}\,\boldsymbol{E})_t$$

where $\mathbb{1}_{obs}$ ($\mathbb{1}_{\overline{obs}}$) is an indicator vector: its component-wise value would be equal to 1 if the associated flux/protein component have been observed (unobserved), 0 otherwise. Taking this into account, we reformulated the problem as following:

- Observed fluxes were introduced as additional constraints with
$$v_i \sim \mathcal{N}(v_i^{obs}, \sigma_{v_i}^2)$$
where σ_{v_i} was set to a small value.
- The objective function was calculated only on the subset of observed enzyme abundances:

$$Z = \frac{1}{p_{\boldsymbol{v}}} \sum_{i=1}^{N^{obs}} (E_i - |v_i|)^2$$

Prediction of metabolic fluxes have been performed by coupling the DynamosYeast model to our experimental data (protein abundances and the CO_2 reaction rate, the only measured flux in our study). We constrained the solution space L through the use of the maximum CO_2 release rate, measured at the same time

point as the one used for proteomics analyses (Blein-Nicolas *et al.* 2015). For each strain observed at each temperature, selection of a particular solution have been made through minimization of the objective function defined in eq. 11, given the observations.

Testing the prediction algorithm

The prediction algorithm is based on the assumption that fluxes and enzyme abundances covary. Indeed, any reaction rate can be expressed as a more or less complex function of enzyme abundances, kinetic constants and metabolite concentrations (Fell and Cornish-Bowden 1997):

$$v_i = k_{cat_i} E_i f(\boldsymbol{\kappa}, \boldsymbol{m}, E)$$

where k_{cat} is the catalytic constant, κ is a set of other kinetic constants, E is the set of abundances of enzymes other than enzyme i. The f function can be more or less complex depending on the mode of regulation.

To test the accuracy of the prediction of metabolic fluxes from protein abundance data, we used the feasible solution space of the Dynamoyeast model and different kinds of functions that relate reaction rates to enzyme abundances. Specifically, we inverted the relationship, expressing protein abundance as a function of the reaction rate from a simplified formalism derived from the Metabolic Control Theory (Kacser and Burns 1981):

$$v_{initial} = \frac{1}{\frac{1}{A_i E_i} + \sum_{j \neq i} \frac{1}{A_j E_j}}$$

where the $A'_j s$ are positive or negative constant terms. Given that enzyme concentrations cannot be negative, and taking $\forall j, \quad A_j = \pm 1$, we get the hyperbolic relation:

$$E_i = \left| \frac{v_{initial}}{1 - v_{initial}} \right| \tag{12}$$

We also tested the case where protein abundances and flux reaction rates were linearly related:

$$E_i = k \, |v_{initial}| \tag{13}$$

k being an uniform random number $k \sim \mathcal{U}(0.1, 3)$

Finally, we considered the case where protein abundances and flux reaction rates are linked by a sigmoidal function (Nijhout *et al.* 2003), which we approximated with a Hill function:

$$E_i = \left| \frac{v_{initial}^n}{1 - v_{initial}^n} \right| \tag{14}$$

where n is the Hill coefficient, sampled in the set $\Omega = \{2, 3, 4, 5\}$.

Formally, for each simulation, we sampled an initial set of fluxes $\boldsymbol{v}_{initial} \in L$. We estimated the complete set of enzymatic protein abundances, $\boldsymbol{E}_{initial}$ using (12, 13 or 14). Then, we minimized the Z objective function to predict the set of fluxes $\boldsymbol{v}^{predicted}$ that best fit enzyme abundances. Accuracy of the predictions was measured by the correlation coefficient between $\boldsymbol{v}^{predicted}$ and $\boldsymbol{v}_{initial}$. Computer simulations were performed to test the influence of two main parameters: (i) the number of sampled points N_s; (ii) the number of quantified proteins, N^{obs}, included in the minimization process.

Practically, we assumed to be under steady state condition ($\dot{\boldsymbol{m}} = 0$) and we sampled N_s points of the solution space from the multivariate posterior joint distribution of fluxes through the EP algorithm Braunstein *et al.* (2017). We drew an additional point in the solution space of L, $\boldsymbol{v}_{initial}$, and we calculated protein concentrations from the inverse problem. We retained the set of fluxes, $\boldsymbol{v}^{predicted}$ for which Z was minimum. The numbers N^{obs} and N_s were let to vary ($N_s \in \{10^2, 10^3, 10^4, 10^5, 10^6\}$ and $N^{obs} \in \{1, 2, 3 \dots\}$).

In terms of computational time, it would be expensive to consider all different combinations of observed enzymatic proteins associated to the metabolic model that can be included in eq. 11 (there are $N^{obs}(1 + (N^{obs} - 1) + (N^{obs} - 1)(N^{obs} - 2) + \cdots + (N^{obs} - 1)!)$ combinations). Therefore, for a given N_s, our strategy was to randomly choose one-by-one a protein to include in the computation of the Z function and therefore for the prediction of metabolic fluxes, $\boldsymbol{v}^{predicted}$.

We randomly choose one reaction, v_1, over the complete set of reactions in the model, and we minimized

$$Z_1 = \frac{1}{p_v}(E_1 - |v_1|)^2 \tag{15}$$

to select one over the N_s possible solutions of L, $\boldsymbol{v}_1^{predicted}$. At the next iteration, we randomly chose an additional flux v_2 and its associated protein abundance E_2, and we minimized

$$Z_2 = \frac{1}{p_v} \sum_{i=1}^{2} (E_i - |v_i|)^2 \tag{16}$$

to predict $\boldsymbol{v}_2^{predicted}$. This procedure is performed until the complete set of reactions is selected. Overall, simulations have been run a thousand of times for different values of N_s and N^{obs}.

Statistical Analysis

In order to study the main features characterizing fermentation and life-history traits in the HeterosYeast dataset, we analyzed the components of variation of a dataset consisting of three different levels of cellular organization: protein abundances **E**, metabolic fluxes **V** and fermentation/life-history traits **T**:

$$D = (\mathbf{E}, \mathbf{V}, \mathbf{T})$$

The total number of observations was 127 strain $\times$ temperature combinations (66 strains $\times$ 2 temperatures $-$ 5 missing data due to the poor fermenting abilities of some strains). The whole dataset consisted of 615 protein abundances, 70 metabolic fluxes and 28 fermentation and life-history traits.

Two types of analysis using several multivariate approaches were performed: an analysis at a single phenotypic level and an analysis integrating the different levels.

We run Principal Component Analyses (PCA) to identify the largest sources of variation in the datasets and the similarities/differences observed between the different phenotypic levels. We included prior knowledge regarding the yeast species in the analysis to perform a supervised analysis with sparse Partial Least Squares Discriminant Analysis (sPLS-DA) in order to extract and combine discriminating features that best separate the different groups. The number of selected features have been tuned using 3-fold cross-validation repeated 1000 times.

Furthermore, integration of the different levels of cellular organization have been performed in a unsupervised framework through a regularized Canonical Correlation Analysis (rCCA), using the *mixomics* package in R (Lê Cao *et al.* 2009; Rohart *et al.* 2017). We first searched for the key features that maximize the correlation between metabolic fluxes and fermentation traits.

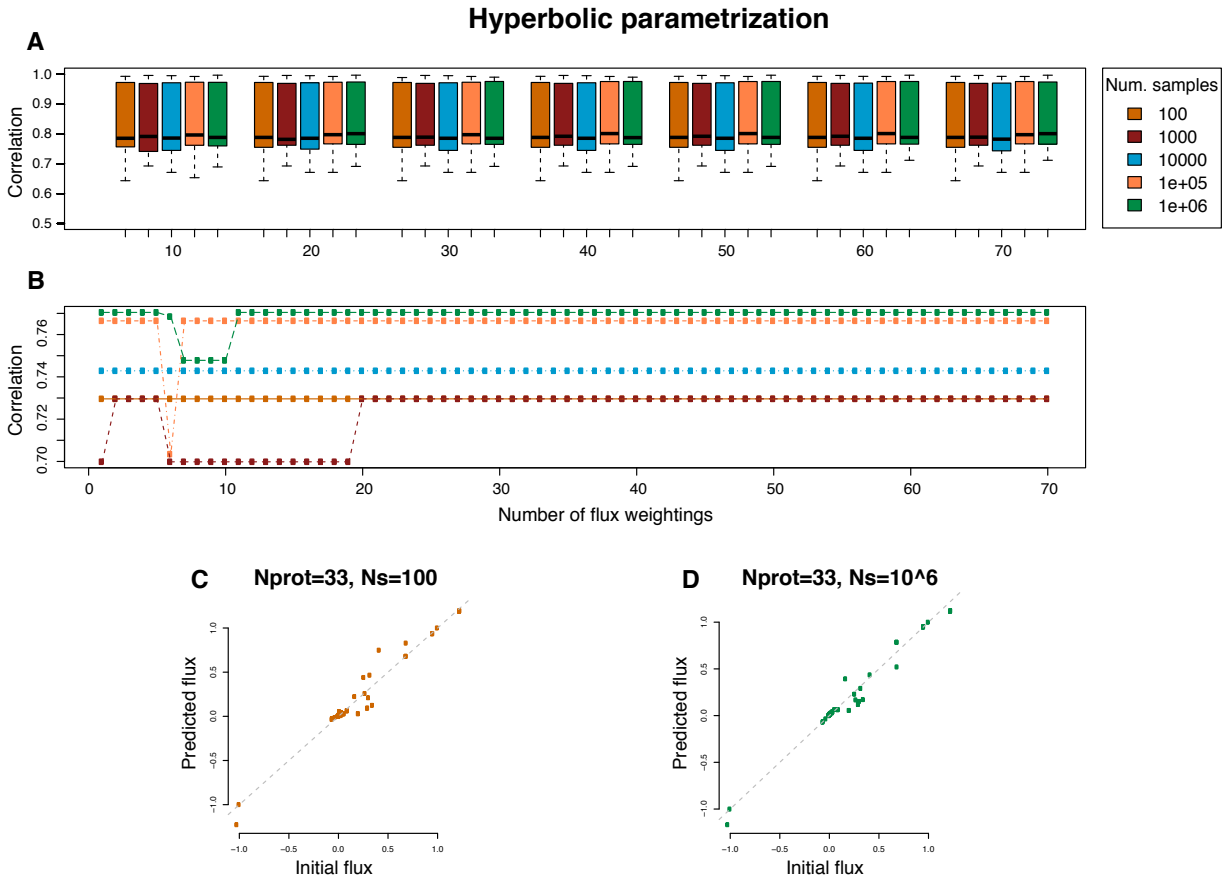


Figure 2 Correlations between initial and predicted fluxes in simulated datasets using the DynamoYeast model. Enzymatic protein abundances were expressed in terms of a hyperbolic function of the initial fluxes using eq. 12. Colors indicate the number of points N_s that were sampled in the solution space L. **A.** Boxplot representation as a function of the number N^{obs} of observed proteins. Each box represents thousand simulations. **B.** Changes observed for the correlation during a single simulation run when increasing one by one the number of observed proteins from 1 to 70. **C.** Relation between the initial and the predicted fluxes shown for one simulation with $N^{obs} = 33$ and $N_s = 10^4$. **D.** Relation between the initial and the predicted fluxes shown for one simulation with $N^{obs} = 33$ and $N_s = 10^6$.

Second, we looked for groups of proteins that maximized the correlation with the most integrated traits (tuning of the regularization parameters have been performed through leave-one-out cross-validation procedure on a 1000×1000 grid between 0.0001 to 1). Finally, Pearson's chi-square of enrichment was computed on protein functional category frequencies taking as prior probability the expected categorical frequency found in the MIPS database.

Since the correlation matrix between traits and fluxes was clearly structured, we computed the matrix of Euclidean distance between traits, based on the correlations with metabolic fluxes, and clustered traits using the *hclust* package in *R*. This procedure allowed us to define five trait groups that showed similar correlation patterns with fluxes of the central carbon metabolism. Finally, we stored the linear correlation coefficients between proteins ($P = 615$ proteins) and traits ($T = 28$) in a $(T \times P)$ matrix and ran a Linear Discriminant Analysis to seek for proteins that best discriminate between trait groups, considering traits as individuals. Functional analysis of proteins that best correlate with LDA axes was performed using the 34 protein functional categories defined above.

Results

The HetherosYeast dataset provided priceless observations on the genetic diversity of yeast strains involved in the winemaking process at different levels of cellular organization: phenotypic traits either related to life-history or fermentation (da Silva *et al.* 2015), and quantitative proteomic data (Blein-Nicolas *et al.* 2015). All traits were estimated or measured at 18°C and 26°C on a half-diallel design between 7 strains of *S. cerevisiae* and 4 strains of *S. uvarum*, with a total of 127 strain $\times$ temperature combinations. In order to access an intermediate level of integration between protein abundances and traits, we used a curated Constraint-Based Model (CBM) of yeast central carbon metabolism (Celton *et al.* (2012); Figure 1) to predict unobserved fluxes at the CBM scale that best match the observed patterns of variations of protein abundances.

To this end, the strategy we proposed was to: *(i)* characterize the feasible space of solution L through the posterior density distribution of fluxes, given by the EP algorithm (Braunstein *et al.* (2017)); *(ii)* select a unique solution through minimization of the objective function Z (eq. 11) that measures the Euclidean

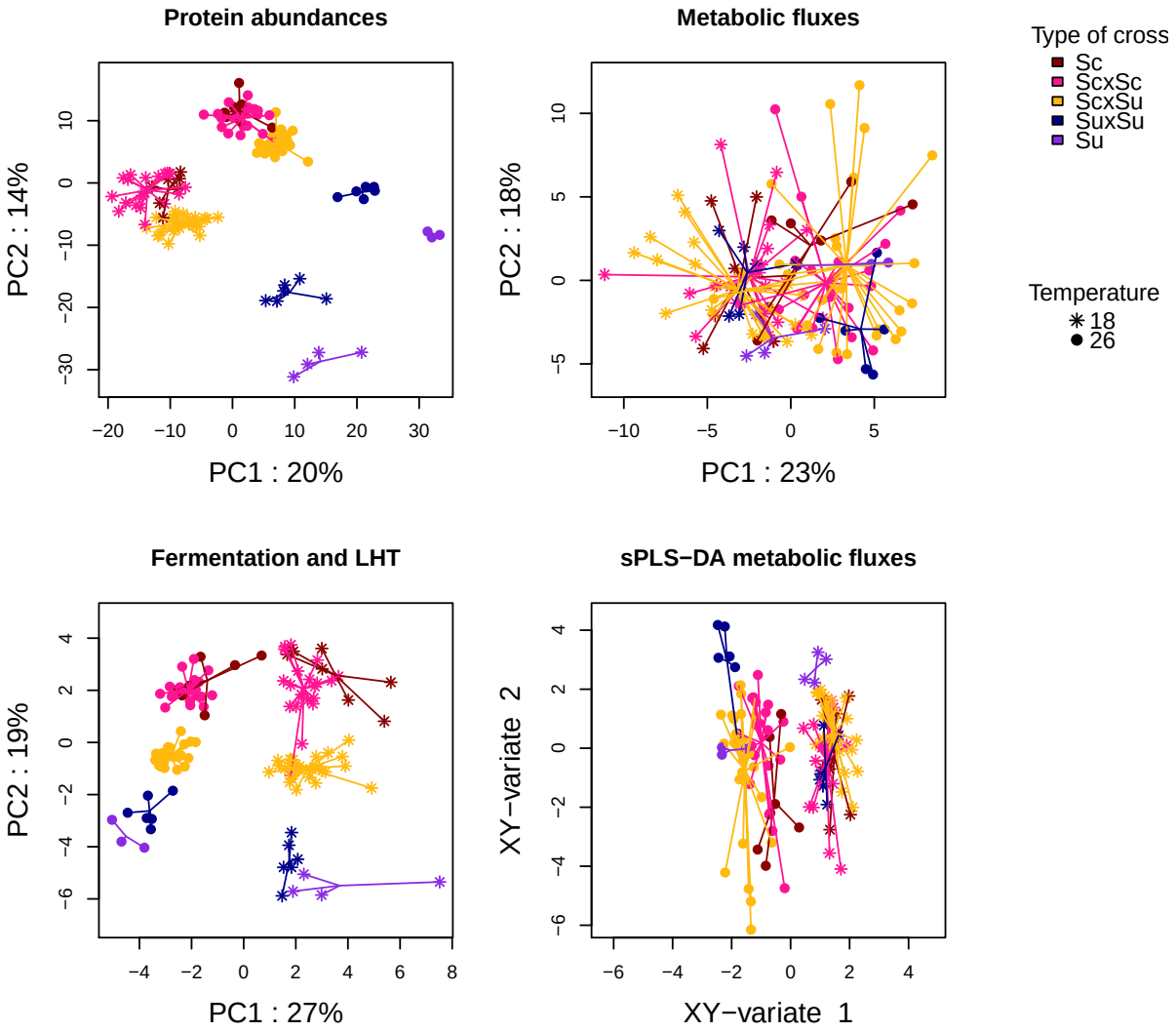


Figure 3 Principal Component Analysis and sparce Partial Least Square-Discriminant Analysis. PCA for protein abundances (top-left), metabolic fluxes (top-right) and fermentation/life-history traits (bottom-left). sPLS-DA for metabolic fluxes (bottom-right). Observations are represented on the first two PCA axes (sPLS-DA, respectively). Each dot correspond to a strain by temperature combination. Temperatures are differenciated by the type of dot, while type of crosses are identified by colours.

distance between observed enzyme abundances and reaction rates.

Below, we first describe the method and its validation using simulated datasets. Then, we analyze the relationships between the different integration levels, using predicted fluxes from central carbon metabolism and the HeterosYeast dataset.

Sampling the feasible solution space with the Expectation Propagation algorithm

Sampling points of the feasible space of solution L can be performed directly from the posterior truncated multivariate normal distribution of fluxes defined in eq. 9. We compared the Hit and Run (HR) algorithm (Meersche *et al.* 2009) to the EP posterior distribution of fluxes to test the goodness in prediction of the EP on the DynamosYeast posterior. The EP methodology gave a good approximation for the mean and variances of the posterior marginal distribution of fluxes (Supplementary method Sampling the solution space and Figure SF1-SF2), as well as for the variance-covariance matrix between fluxes (Figure SF3). These results are similar to the ones obtained in Braunstein *et al.* (2017). Therefore, we decided to rely on the EP algorithm to sample the feasible solution space of the CBM.

Protein abundances are good predictors of the initial set of metabolic fluxes

Computer simulations have been performed to access the goodness in prediction of the proposed method, as detailed in section Testing the prediction algorithm. The two main parameters to test were: *(i)* the number of sampled points N_s of L; *(ii)* the number N^{obs} of observed proteins to be included in Z (eq. 11). Simulations showed that minimization of eq. 11 leads to a high correlation between $v^{initial}$ and $v^{predicted}$ (Figure 2-A). Correlations ranged from 0.65 to 0.99 (p-value < 0.05). By increasing the number of sampled points in L, N_s, the mean correlation slightly

increased and its variance decreased. The number of observed protein abundances, N^{obs} had a more complex influence on the prediction accuracy. When increasing N^{obs}, the correlation between $\boldsymbol{v}^{initial}$ and $\boldsymbol{v}^{predicted}$ either increases, decreases or stays constant, as illustrated in Figure 2-B. However, the order of magnitude of the variations were small, and the correlation tends to be more stable for a high N_s value (Figure 2-B). When considering the actual number of enzyme abundances ($N_{obs} = 33$) that matched between the HeterosYeast proteomic data and the DynamoYeast CBM, we observed a high correlation between $\boldsymbol{v}^{initial}$ and $\boldsymbol{v}^{predicted}$ after setting $N_s = 10^6$ (Figure 2-C). Altogether, we considered that our algorithm was efficient to predict unobserved fluxes from enzyme abundances, given the structure of the metabolic network.

Predicting unobserved fluxes from the observed variation of protein abundances

The HeterosYeast proteomic data were used in the context of the DynamoYeast model of yeast central carbon metabolism. From the 615 protein abundances, we were able to quantify the proteins (or protein complexes) associated to 33 of the 70 reactions in the metabolic model. For each strain $\times$ temperature combination, observed CO_2 release rates were used as additional constraints in the form of *a priori* knowledge to get the feasible solution space L. We sampled $N_s = 10^6$ points in the space of solutions to select a unique solution of L that minimizes the Euclidean distance between fluxes and enzymes abundances. We therefore predicted the 69 unobserved fluxes in the CBM for each of the 127 strain $\times$ temperature combinations. Then statistical approaches have been used to investigate the components of variation and the structure of the new dataset consisting of 615 protein abundances ($\boldsymbol{E}$), 70 metabolic fluxes ($\boldsymbol{V}$) and 28 fermentation and life-history traits ($\boldsymbol{T}$):

$$D = (\mathbf{E}, \mathbf{V}, \mathbf{T})$$

Patterns of variation depend on the integration levels

The 127 observations of the new dataset D had a specific structure. There was 7 parental strains (*S.c.*) and 21 intraspecific hybrids (*S.c.*$\times$*S.c.*) from *S. cerevisiae*, 4 parental strains (*S.u.*) and 6 intraspecific hybrids (*S.u.*$\times$*S.u.*) from *S. uvarum*, and 28 interspecific hybrids (*S.c.*$\times$*S.u.*). All strains were observed during alcoholic fermentation on wine grape juice at two temperatures, 18°C and 26°C (da Silva *et al.* 2015).

To better understand the patterns of variation at each integration level, Principal Component Analysis (PCA) have been computed on each type of trait separately. Results are presented in Figure 3, where strains are identified by species, type of cross (intra-specific hybrid, inter-specific hybrid or parental strain) and temperature. The first PCA component accounted for 20%, 23% and 27% of the total variation, the second for 14%, 18% and 19% for protein abundances, metabolic fluxes and fermentation/life-history traits, respectively. Depending on the integration level, we observed different patterns of phenotypic diversity.

At the proteomic level ($\boldsymbol{E}$), the first two PCA axes contributed to both differences between temperatures and between species and type of cross. Heterosis is observed for all types of hybrids at both temperatures. First, *S.u.*$\times$*S.u.* hybrids are clearly differentiated from their *S.u.* parents. Second, *S.c.*$\times$*S.u.* interspecific hybrids are closer to their *S.c.* parents than to their *S.u.* parents. Finally, *S.c.*$\times$*S.c.* hybrids are close to their *S.c.* parents, but the range of variation between *S.c.*$\times$*S.c.* hybrids is larger than the one between parental strains. Altogether, the protein abundance of an hybrid strain cannot be predicted by the mean of its parental values.

At the trait level ($\boldsymbol{T}$), we observed a high temperature effect, with axis 1 (27% of the variation) separating clearly strains that grew at 26°C from those that grew at 18°C. At 26°C, strains were characterized by high growth rate (r), high CO_2 fluxes ($J_{\max}$ and $V_{\max}$), high *Hexanol* and *Decanoic acid* and low carrying capacity (K) and low fermentation times (*AFtime*, *t-lag*, *t-75*, *t-45*) (Figure SF4). At 18°C, strains were characterized by low growth rates and CO_2 fluxes and high K and fermentation times (Figure SF4). Those two groups of traits mostly vary with the temperature, although some differences between strains are observed within rather than between types of cross, especially at 18°C. At 26°C, *S.u.* strains perform slightly better than *S.c.* strains (higher growth rates, faster fermentation times). The types of cross are clearly separated along PCA axis 2. Again, heterosis is observed for intraspecific hybrids. However, interspecific hybrids seem to be in-between the two parental strains. Traits that explain the differences between observations along axis 2 were cell-size (*Size-t-*$N_{\max}$) and *Ethanol* at the end of fermentation (positively correlated to axis 2), aroma production at the end of fermentation, as well as *Sugar.Ethanol.Yield* (negatively correlated) (Figure SF4). Note that those traits are not influenced by the temperature. Hence, at the trait level, we observed differences between yeast species for traits related to aroma production that were not influenced by the temperature. Most fermentation and life-history traits showed a strong temperature effect and high differences between strains within type of cross, and a weak heterosis.

At the flux level ($\boldsymbol{V}$), temperature separated the observations on axis 1, but both axis 1 and axis 2 differentiated strains independently of their origin. Notice however that the range of variation of the hybrids is larger than the one of the parental strains, that indicates differences between inbred and hybrid strains. Altogether, central carbon metabolic fluxes were influenced by the temperature and showed strong differences between strains that were not related to the type of cross and the parental species. Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) have been computed on metabolic fluxes in order to select the main features characterizing species $\times$ temperature combinations (Figure 3). As previously, the first axis differentiated strains observed at different temperatures. Six fluxes contributed to the first axis of the sPLS-DA: CO_2, ethanol, pyruvate decarboxylase, alcohol dehydrogenase, 6-phosphogluconolactonase and phosphogluconate dehydrogenase fluxes (Figure SF5). All were negatively correlated with axis 1 and were involved in fermentation. This shows that fermentation was more efficient at 26°C. The second axis differentiated inbred strains from intraspecific hybrids with genotype $\times$ temperature interaction: both *S.u.*$\times$*S.u.* and *S.c.*$\times$*S.c.* hybrids have higher coordinates than their parents at 26°C, while *S.u.*$\times$ *S.u.* have lower coordinates than their parents at 18°C, and *S.c.*$\times$*S.c.* hybrids are confounded with their parental strains. Inter-specific hybrids are characterized by a wide range of variation at both temperatures. Fluxes that contributed to axis 2 were in majority mitochondrial fluxes. Mitochondrial acetyl-CoA formation, mitochondrial citrate synthase, mitochondrial aconitate hydratase, mitochondrial isocitrate dehydrogenase (NAD+) and mitocondrial transport fluxes of pyruvate, oxaloacetate and acetaldehyde were negatively correlated with the second axis, while mitochondrial transport of 2-oxodicarboylate, ethanol and CO_2 fluxes were positively correlated (Figure SF5).

In short, we found at each integration level a strong effect of the temperature, large differences between strains, and evidence for heterosis, *i.e.* differences between hybrids and mid-parent values. However, the patterns differed between the proteomic and the most integrated level. At the proteomic level, proteins involved in differences between strains were the same as the ones involved in differences between species and between temperature. At the flux level, there were few differences between species. Differences between temperatures were associated to enzymatic reactions related to fermentation, while differences between strains were associated to enzymatic reactions either involved in fermentation, or in the part of the TCA that occurs in the mitochondria. At the trait level, differences between temperatures were associated to differences in growth and fermentation traits, that were relatively conserved within species but showed between-strain variations. Differences between species mostly concerned volatile compounds at the end of fermentation, that are produced by the secondary metabolism.

Fermentation and life-history traits are associated with different metabolic pathways of the yeast carbon metabolism

Regularized Canonical Correlation Analysis (rCCA) have been performed to investigate correlations between metabolic fluxes and fermentation/life-history traits (Figure 4). Fermentation and life-history traits were divided mainly in two groups showing contrasting profiles. The first group consisted of traits that clustered with the carrying capacity, *K*. They were characterized by a negative correlation with fluxes involved in the glycolysis, ethanol synthesis and pentose phosphate pathway, and by a positive correlation with fluxes in the TCA reductive branch. In contrast, the second group consisted of traits that clustered with the intrinsic growth rate, *r*, and were characterized by a positive correlation with fluxes involved in the glycolysis, ethanol synthesis and pentose phosphate pathway and by a negative correlation with fluxes in the TCA reductive branch. Consistently, the *biomass* pseudo-flux was positively correlated with *r* and negatively with *K*.

When looking at the flux correlation structure revealed by Figure 4, we can see the opposition between the two well-known ways of producing energy in yeast. Fermentation is associated to an extensive usage of glycolysis and pentose-phosphate metabolic pathways, while respiration is associated to high TCA fluxes. Hence, high growth rate and CO_2 fluxes ($J_{\max}$, $V_{\max}$) and correspondingly fast fermentation (low fermentation times) seem to be associated to central carbon metabolism oriented towards fermentation, while high carrying capacity, low growth rate and slow fermentation seem to be associated to central carbon metabolism oriented towards respiration.

The *K* group could be divided into three subgroups, depending mainly on the correlations between the traits and the fluxes of glycerol synthesis and of acetaldehyde: **AFtime**, **K** and **CO2max** (subgroups designated by the name of the main trait in boldface). The **AFtime** subgroup showed a slightly negative correlation, the **K** subgroup a slightly positive correlation and the **CO2max** subgroup a positive correlation. **AFtime** grouped most traits correlated with the duration of fermentation, *AFTime*, *t-45*, *t-75*, *t-*N_{max}; **K** grouped traits measuring the lag time and beginning of fermentation (*t-lag*, *t-*V_{max}), the carrying capacity (*K*) and the level of *Octanoic acid* (fatty acid) at the end of fermentation, while the **CO2max** grouped traits correlated with fermentation products (*total* CO_2, *Ethanol* and *sugar-ethanol yield*), two volatile esters, *Isoamyl acetate* and *Phenyl-2-ethanol acetate*, as well as cell size and cell viability measured close to the end of fermentation, and *t-*N_0.

Similarly, within the *r* group we distinguished two clusters of traits: **Vmax** and **SO2**. **Vmax** grouped traits that correlated with V_{max} and *r*, as well as the amount of *hexanol* (alcohol) and *hexanoic* and *decanoic acids* (fatty-acid) that were quantified at the end of fermentation. **SO2** grouped basic oenological parameters measured at the end of fermentation (*total* and *free* SO_2, *residual sugar*), cell viability measured once carrying capacity is reached (*Viability-t-*N_{max}), and two volatile compounds *Phenyl-2-ethanol* (alcohol) *4-methyl-4-mercaptopentan-2-one* (thiol).

Briefly, we were able to associate fermentation and life-history traits to metabolic fluxes based on their correlation patterns. In particular, we found that the negative correlation between *r* and *K* is explained by a different pathway usage of the central carbon metabolism. High *r* and low *K* are associated with glycolysis and fermentation, while low *r* and high *K* are associated with TCA cycle and respiration.

Metabolic bases of yeast phenotypic trait variation

In order to confirm the association between integrated traits variation and differential usage of central carbon metabolism, we tried to identify the proteins outside the DynamoYeat model that were involved in the trait patterning, as observed from the correlation between traits and fluxes. We performed a Linear Discriminant Analysis on the correlation matrix between the *T* traits and the *E* proteins using as discriminant features the five groups of fermentation and life-history traits showing a similar correlation structure with metabolic fluxes, obtained in the previous analysis (see section Statistical Analysis).

Linear Discriminant Analysis clearly separated the five trait categories on the first axis, that explains 99% of the total variation (Figure 5). **AFtime** and **K** traits were close, and had positive coordinates on LDA1; **Vmax** had high negative coordinates, $\boldsymbol{SO_2}$ had a slightly negative mean and **CO2max** had a slightly positive mean on LDA1. Given the high discriminative power of LDA1, it is clear that proteins positively or negatively correlated to LDA1 participate to the differentiation between **AFtime** and **Vmax** trait groups.

Functional analysis of proteins that best correlate with the first axis of the LDA was performed on the group of proteins showing a correlation of 0.85 in the positive and in the negative direction. Pearson's chi square test of enrichment showed that the group of proteins negatively correlated to the first axis was enriched in proteins linked to protein fate, cytoskeleton, detoxification, growth and death but also to the fermentation, glycolysis and phosphate pathway. The group of proteins that positively correlated with LDA1 was enriched in proteins linked to energy conversion, nitrogen and sulfur pathway, metabolism, energy reserves, electron and respiration. This result was represented as a cloud of words on Figure 5.

In conclusion, the association between trait variation and central carbon metabolism observed at the flux level is confirmed by the proteomic analysis. Proteins that covary with traits of the **Vmax** group and with glycolytic and fermentation fluxes are enriched in proteins involved in glycolysis and fermentation, but also in protein synthesis and degradation (protein fate), and cytoskeleton, that can be associated to cell division. Proteins that covary with traits of the **AFtime** group and with TCA and respiration fluxes are enriched in proteins involved in TCA and respiration, but also in electron transport, energy conversion

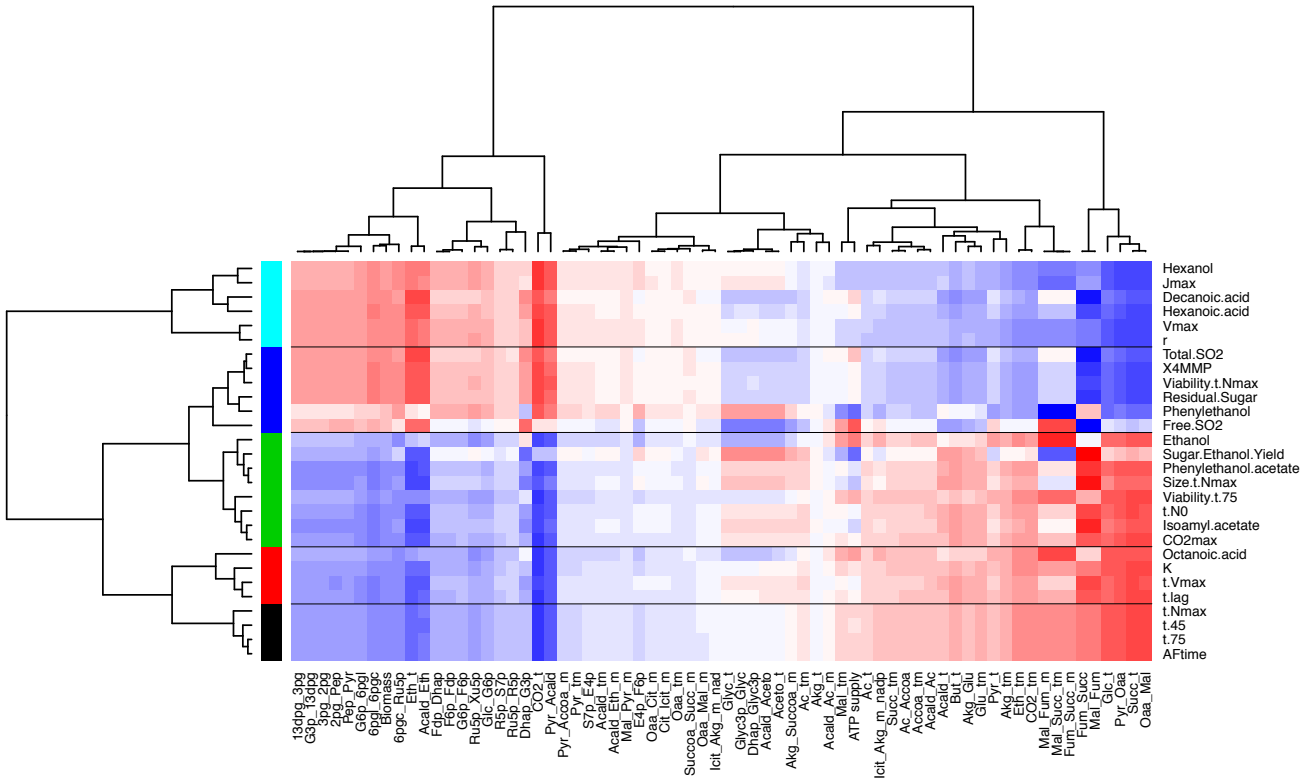


Figure 4 Regularized Canonical Correlation Analysis on metabolic fluxes and fermentation/life-history traits. Penalization parameters have been tuned through leave-one-out cross-validation method on a 1000 × 1000 grid between 0.0001 and 1 ($\lambda_1 = 0.8$, $\lambda_2 = 0.0001$). Canonical correlation values between metabolic fluxes and fermentation/life-history traits are represented as a gradient of colors from blue (-1) to red ($+1$). Metabolic fluxes and fermentation/life-history traits have been clustered using the *hclust* method. Colored row side bars indicate the five groups obtained on fermentation and life-history traits.

and nitrogen and sulfur metabolism.

Discussion

In this work, we applied cutting-edge methods for data integration to an original yeast dataset. The HeterosYeast dataset comprised quantitative proteomics, fermentation traits and life-history traits measured during wine fermentation on a wide range of strains from two yeast species. The objective was to integrate information at different levels of cellular organization (proteomic and metabolic fluxes) to better understand the metabolic bases of yeast phenotypic variation, in particular for life-history traits related to fitness. The key point of this study was to incorporate proteomic data in a constraint-based metabolic model to estimate unobserved metabolic flux values. Then, using a combination of multivariate analyses dedicated to a heterogeneous datasets of high dimension, we were able to show that the metabolic flux level retains information that was not directly interpretable at the proteomic or at the trait level. In particular, we showed that the negative correlation between traits associated with population growth rate and traits associated to maximal population size (carrying capacity) could be explained by a differential usage of central carbon metabolism: fermentation *versus* respiration.

Constraint-based modeling can predict unobserved fluxes from observations at the cellular level

Functional genome annotations, allied with current knowledge in biochemistry, now allows describing cell metabolism at genome-scale, using constraint-based metabolic models that take into account the stoichiometry of each reaction and incorporate thermodynamic constraints (Palsson 2015). Without any *a priori* knowledge, the number of steady-state solutions for reaction rates are infinite, but can be reduced by observations. Three types of experimental data can be used in this process: *(i)* exchange metabolic fluxes; *(ii)* metabolite input/output rates and *(iii)* protein abundances. External metabolic fluxes and metabolite input/output rates can be used directly in constraint-based models to reduce the feasible space of solutions, L (eq. 5 and ineq. 6) under the steady state assumption.

Protein abundances, linked to the metabolic fluxes in the model through GPR (gene-protein-reaction) association, carry information on the network functioning and on the state of the metabolic network at a given time and under a specific condition. Following Lee *et al.* (2012), we used protein abundance profiles to find the set of metabolic fluxes that minimized the Euclidean distance between metabolic fluxes and enzyme abundances. Indeed, even though the relationship between flux and enzyme abundances is commonly non-linear, the level of use of a given pathway is more or less associated with the abundance of its enzymes (Sabarly *et al.* 2016).

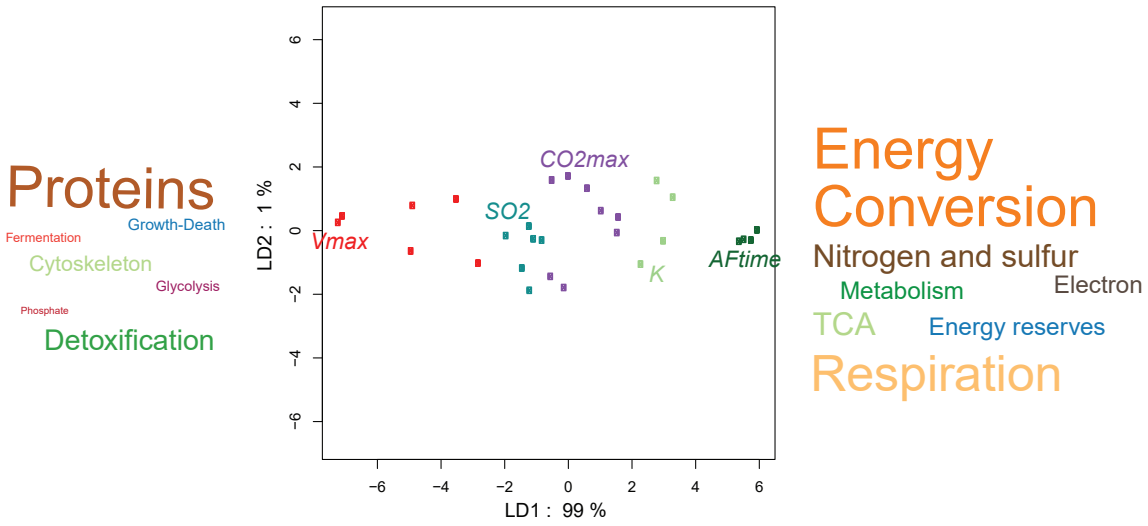


Figure 5 Projection of the 28 traits in the first two axes of a Linear Discriminant Analysis on protein abundances. Trait groups were constituted from their correlation with fluxes of central-carbon metabolism. Each dot is one fermentation or life-history trait. Colors correspond to trait groups, identified by one representative trait. The results confirm the structure of fermentation and life-history traits and reveal two trait groups with antagonistic proteomic pattern: the **AFtime** group and the **Vmax** group. Functional enrichment of proteins positively or negatively correlated to the first axis were represented by a cloud of words

The method that we propose relies on a probabilistic approach. Following Braunstein *et al.* (2017), we chose to characterize the feasible space of solutions L by means of its posterior density distribution through the Expectation Propagation (EP) algorithm. The computation time of EP algorithm is much shorter than the well known Hit and Run (Bélisle *et al.* 1993), and it provides both samples of metabolic fluxes in L and their associated posterior probability. In the selection process of a unique solution of L, we minimized Z, the Euclidean distance between the observed abundances of proteins and the associated metabolic fluxes weighted by the inverse of the probability of observing such set of fluxes, p_v (eq.11). This minimization process involved sampling in L, and selection was made after computation of the Z value over a high number of sampled points.

Computer simulations confirmed the good prediction efficiency of our method. In particular, we showed that the prediction efficiency was not affected by non linearities of the flux-enzyme relationship. The most important parameter was the number of reactions N^{obs} for which proteomic observations were available, as compared to the CBM size n. When N^{obs} was too low, adding a new information could lead to a decrease of the prediction efficiency. A decrease in the correlation between initial and predicted fluxes means that, once a new enzyme is added, the solution that minimizes the total Euclidean distance leads to flux predictions farther from their true value. This can occur whenever there is a weak correlation between the first $n-1$ fluxes, and the additional flux v_n. Therefore, it is important that observations on protein abundances do cover the main features in the architecture of the metabolic network. In our case, 33 reactions with observed protein abundances out of the 70 reactions of de DynamoYeast model were sufficient to reach a high prediction accuracy. Recent progresses in gel-free/label-free quantitative proteomics now allow to quantify thousands of proteins and should ensure a good coverage even for genome-scale metabolic models (Belouah *et al.* 2019).

Even though our flux predictions are not expected to be exact, we are confident that our method reveals the main orientations of cell metabolism. It takes advantage of additional information about the known architecture of the metabolic network to predict unobserved fluxes from observed protein abundances and globally add information on the system.

Unraveling the metabolic bases of life-history trait variation

The proposed approach has been used to predict metabolic fluxes from central carbon metabolism in a population obtained from a half-diallel cross between two yeast species, *S. cerevisiae* and *S. uvarum*, for which the genetic values of 615 protein abundances and 28 fermentation/life-history traits have been estimated under fermentation conditions at two different temperatures, 18°C and 26°C, leading to a total of 127 observations on 66 different yeast strains (Albertin *et al.* 2013a). As described above, we predicted metabolic fluxes for each strain $\times$ temperature combination by coupling the DynamosYeast model, a highly curated constrained based model of the central carbon metabolism (Celton *et al.* 2012), using the observed CO_2 release rate as *a priori* knowledge, and measurements of protein abundances associated to 33 out of the 70 reactions in the model.

The final dataset consisted in three matrices of 127 $\times$ 615 protein abundances, 127 $\times$ 70 central carbon fluxes, and 127 $\times$ 28 fermentation/life-history traits. The total number of phenotypes (713) greatly exceeded the number of observations and we used regularization techniques for the multivariate analyses (Rohart *et al.* 2017). In order to connect patterns of variation observed at different levels, we used a top-down strategy, from the most integrated to the less integrated level. First, we explored the correlations between traits and metabolic fluxes. Second, we tracked the proteins outside the metabolic model that best explained the correlation structure between traits and fluxes.

In our dataset, we observed a negative correlation between

traits associated to growth and CO_2 fluxes, and traits associated to population size and duration of the fermentation process. Those negative correlations resulted in different life-history strategies that have been observed elsewhere, on different yeast collections either from industrial (Albertin *et al.* 2013b) or natural origin (Spor *et al.* 2008, 2009). It roughly corresponds to the well-known *r-K* trade-off in ecology (Pianka 1970). More recently Collot *et al.* (2018) suggested that such trade-off could emerge from eco-evolutionary feedback loops because competing strains also modify their environment through the production of different sets of metabolites. The HerosYeast dataset shows that the choice of a strategy is plastic (da Silva *et al.* 2015) and can be modified by the environment (here the fermentation temperature).

Adding information about central carbon metabolic fluxes, we showed that such trade-off can be explained by metabolic switches between fermentation associated to glycolysis, and respiration, associated to TCA cycle. Such duality in the functioning of yeast central carbon metabolism has already been observed when matching the DynamoYeast model to experimentally measured exchange fluxes (Nidelet *et al.* 2016) in a collection of *S. cerevisiae* strains. The switch between the two modes of functioning (Figure 4) depends partly on the isoforms of the alcohol dehydrogenase (ADH). Interestingly, Albertin *et al.* (2013b) already found that the trade-off between cell-size and *K* was related to changes in the percentage of acetylation of the ADH 1p, with high levels being associated to large cells and low K.

Because this paper was devoted to a proof of concept, we deliberately chose to focus on central carbon metabolism and we used the DynamoYeast model because it describes a small number of reactions, as compared to available genome-scale models (Caspi *et al.* (2014)). Therefore, we were not able to explain between-strains variations for traits related to secondary metabolism like aroma production, that merely discriminated between the two yeast species of the HerosYeast dataset. Moreover, only a small subset of the proteomic data were coupled to the metabolic model. Seeking the proteins that most explain trait patterns that was revealed at the flux level, we were able to find proteins that were associated to the *r-K* trade-off at the trait level. The analysis of protein's functional annotations confirmed the already known link between glycolysis and pentose-phosphate pathways and fermentation, and the link between extensive usage of TCA and respiration.

Altogether, by coupling phenomic data with mathematical modeling of metabolism and cutting-edge statistical analyses taking into account high-dimensionality and heterogeneity of the measures, we were able to explain the commonly observed trade-off between two set of yeasts life-history traits by a differential pathway usage of energy production. Glycolysis and fermentation lead to fast growth and resource consumption. TCA and respiration lead to slow growth and high population sizes. The duality between the two alternative usages of central carbon metabolism is encoded into the architecture of the metabolic network.

Literature Cited

Albertin, W., T. da Silva, M. Rigoulet, B. Salin, I. Masneuf-Pomarede, *et al.*, 2013a The mitochondrial genome impacts respiration but not fermentation in interspecific saccharomyces hybrids. PLOS ONE **8**: 1–14.

Albertin, W., P. Marullo, M. Bely, M. Aigle, A. Bourgais, *et al.*, 2013b Linking Post-Translational Modifications and Variation of Phenotypic Traits. Molecular & Cellular Proteomics **12**: 720–735.

Antoniewicz, M., 2015 Methods and advances in metabolic flux analysis: a mini-review. J Ind Microbiol Biotechnol. **42**: 317–25.

Belouah, I., C. Nazaret, P. Pétriacq, S. Prigent, C. Bénard, *et al.*, 2019 Modeling protein destiny in developing fruit. Plant Physiology p. pp.00086.2019.

Blein-Nicolas, M., W. Albertin, T. da Silva, B. Valot, T. Balliau, *et al.*, 2015 A systems approach to elucidate heterosis of protein abundances in yeast. Mol Cell Proteomics **14**: 2056–71.

Blein-Nicolas, M., W. Albertin, B. Valot, P. Marullo, D. Sicard, *et al.*, 2013 Yeast proteome variations reveal different adaptive responses to grape must fermentation. Molecular Biology and Evolution **30**: 1368.

Blein-Nicolas, M., H. Xu, D. de Vienne, C. Giraud, S. Huet, *et al.*, 2012 Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics. PROTEOMICS **12**: 2797–2801.

Braunstein, A., A. P. Muntoni, and A. Pagnani, 2017 An analytic approximation of the feasible space of metabolic networks. Nature Communications **8**: 14915.

Bélisle, C. J. P., H. E. Romeijn, and R. L. Smith, 1993 Hit-and-Run Algorithms for Generating Multivariate Distributions. Mathematics of Operations Research **18**: 255–266.

Caspi, R., T. Altman, R. Billington, K. Dreher, H. Foerster, *et al.*, 2014 The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res. **42(Database issue)**: D459–71.

Celton, M., A. Goelzer, C. Camarasa, V. Fromion, and S. Dequin, 2012 A constraint-based model analysis of the metabolic consequences of increased NADPH oxidation in Saccharomyces cerevisiae. Metabolic Engineering **14**: 366 – 379.

Cherry, J., E. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, *et al.*, 2012a Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res. **40(Database issue)**: D700–5.

Cherry, J. M., E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, *et al.*, 2012b Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research **40**: D700–705.

Collot, D., T. Nidelet, J. Ramsayer, O. C. Martin, S. Méléard, *et al.*, 2018 Feedback between environment and traits under selection in a seasonal environment: consequences for experimental evolution. Proceedings of the Royal Society B: Biological Sciences **285**: 20180284.

da Silva, T., W. Albertin, C. Dillmann, M. Bely, S. la Guerche, *et al.*, 2015 Hybridization within saccharomyces genus results in homoeostasis and phenotypic novelty in winemaking conditions. PLOS ONE **10**: 1–24.

Edwards, J., R. Ibarra, and B. Palsson, 2001 In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. Nature Biotechnology **19**: 125–130.

Fell, D. and A. Cornish-Bowden, 1997 *Understanding the control of metabolism*, volume 2. Portland press London.

Fell, D. and J. Small, 1986 Fat synthesis in adipose tissue. an examination of stoichiometric constraints. Biochem J. **238**: 781–786.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Gelius-Dietrich, G., C. J. Fritzemeier, A. A. Desouki, and M. J.

Lercher, 2013 sybil – efficient constraint-based modelling in r. BMC Systems Biology **7**: 125.

Gould, S. and R. Lewontin, 1979 The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist programme. Proc. R. Soc. Lond. B **295**: 581–598.

Kacser, H. and J. A. Burns, 1981 The Molecular Basis of Dominance. Genetics **97**: 639–666.

Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, 2017 KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research **45**: D353–D361.

Kanehisa, M. and S. Goto, 2000 KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**: 27–30.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, 2016 KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research **44**: D457–462.

Lee, D., K. Smallbone, W. B. Dunn, E. Murabito, C. L. Winder, *et al.*, 2012 Improving metabolic flux predictions using absolute gene expression data. BMC Systems Biology **6**: 73.

Lê Cao, K.-A., I. González, and S. Déjean, 2009 integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics (Oxford, England) **25**: 2855–2856.

Meersche, K. V. d., K. Soetaert, and D. V. Oevelen, 2009 xsample() : An *R* Function for Sampling Linear Inverse Problems. Journal of Statistical Software **30**.

Nicholson, J. and J. Lindon, 2008 Systems biology: Metabonomics. Nature **455**: 1054–6.

Nidelet, T., P. Brial, C. Camarasa, and S. Dequin, 2016 Diversity of flux distribution in central carbon metabolism of S. cerevisiae strains from diverse environments. Microbial Cell Factories **15**.

Nijhout, H. F., A. M. Berg, and W. T. Gibson, 2003 A mechanistic study of evolvability using the mitogen-activated protein kinase cascade. Evolution & Development **5**: 281–294.

Palsson, B. O., 2015 *Systems Biology, Constraint-based Reconstruction and Analysis*. Cambridge University Press.

Petrizzelli, M., D. d. Vienne, and C. Dillmann, 2019 Decoupling the Variances of Heterosis and Inbreeding Effects Is Evidenced in Yeast's Life-History and Proteomic Traits. Genetics **211**: 741–756.

Pianka, E. R., 1970 On r- and K-Selection. The American Naturalist **104**: 592–597.

Riekeberg, E. and R. Powers, 2017 New frontiers in metabolomics: from measurement to insight. F1000Research **6**: 1148.

Rohart, F., B. Gautier, A. Singh, and K.-A. L. Cao, 2017 mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Computational Biology **13**: e1005752.

Ruepp, A., A. Zollner, D. Maier, K. Albermann, J. Hani, *et al.*, 2004 The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research **32**: 5539–5545.

Sabarly, V., C. Aubron, J. Glodt, T. Balliau, O. Langella, *et al.*, 2016 Interactions between genotype and environment drive the metabolic phenotype within Escherichia coli isolates. Environmental Microbiology **18**: 100–117.

Spor, A., T. Nidelet, J. Simon, A. Bourgais, D. de Vienne, *et al.*, 2009 Niche-driven evolution of metabolic and life-history strategies in natural and domesticated populations of Saccharomyces cerevisiae. BMC Evolutionary Biology **9**: 296.

Spor, A., S. Wang, C. Dillmann, D. d. Vienne, and D. Sicard, 2008 "Ant" and "Grasshopper" Life-History Strategies in Saccharomyces cerevisiae. PLOS ONE **3**: e1579.

Stearns, S., 1992 *The evolution of life histories.*. Oxford University Press.

Thiele, I. and B. O. Palsson, 2010 A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols **5**: 93–121.

Töpfer, N., S. Kleessen, and Z. Nikoloski, 2015 Integration of metabolomics data into metabolic networks. Frontiers in plant science **6**: 49.

Wagner, G. P. and J. Zhang, 2011 The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. Nat Rev Genet. **12**: 204–13.

Watson, M., 1984 Metabolic maps for the apple ii. Biochemical Society Transactions **12**: 1093–1094.

Chapter 6

Chapter 6

Conclusions and perspectives

Yeast species from the *Saccharomyces sensu stricto* phylogenetic group, including *S. cerevisiae* and *S. uvarum* studied here, are important in many areas such as agriculture, biotechnology and medicine. Beside its utility to meet human needs and customs, yeast represents a powerful model system to address core issues in biology. Its short generation time, and the fact that it is easy to grow and manipulate in the laboratory, have allowed to achieve major breakthroughs. In particular, the whole genome sequencing of *S. cerevisiae* (Goffeau et al., 1996) switched the focus from individual genes and functions to a global view of how the cellular networks interact, which has renewed interest on metabolism and its regulation (introduced in Chapter 4). A striking feature of metabolism is the similarity of the basic pathways, even between distant species such as yeast and human, which allows for instance the study in yeast of pathways involved in human diseases. However, pathway usage and regulation can drive huge phenotyic differences between close species, which raises the question of the genotype-phenotype map.

The main results of the thesis were obtained with two complementary modelling approaches applied to the same biological material, in order to: (i) analyze the phenotypic variation from a quantitative and population genetics perspective; (ii) investigate the genotype-phenotype map from an evolutionary systems biology point of view. These approaches were developed on a large yeast dataset collected on a diallel design (HeterosYeast project, chapter 2), where observations were organized in types of crosses (intra- and inter-specific hybrids or parental strains from two yeast species). Measurements were collected at different levels of phenotypic integration, from proteomic to life-history traits, during the wine fermentation process. This dataset allowed to question the complex relationship between genotypes, phenotypes and fitness in populations. Beside, developments related to a better-understanding of the structure of yeast phenotypic diversity and of the wine fermentation process, along with methodological developments, are proposed in my thesis. These methods have actually a broad applicability domain.

The evolution of life-history traits

The first modelling approach was introduced in Chapter 1, in which phenotypic variation is presented as the result of processes of evolution and adaptation. A key component of adaptation and evolvability is the partition of the phenotypic variance into additive and non-additive genetic components, and environmental components (G×E and residual). In this context, the diallel design of the HeterosYeast project was of particular interest. Among all statistical approaches proposed in the literature to analyze genetic and non genetic variance components from such designs, I decided to shape the model proposed by Lenarcic et al. (2012).

The results are reported in Chapter 3. Each measured trait was characterized by its variance components, and comparisons were performed among traits. This work revealed genotype × environment interactions at every level of cellular organization (variance components differed between the two temperatures). It allowed the classification of traits × temperatures combinations into a few number of clearly distinct groups of traits, that excluded the hypothesis that all traits have neutrally evolved under the same process. A possible interpretation is that traits sharing a similar

variance component profile have a common evolutionary history. Moreover, within some groups of traits we have shown that inbreeding and heterosis variance components were decoupled. This original result highlights that inbreeding and heterosis evolved independently. We also showed that epistasis is necessarily involved in this decoupling. These results call for theoretical developments in evolutionary genetics to identify the mechanisms and the driving forces at stake, and for experiments on others species to evaluate whether such findings are common in biological systems.

From a breeder's perspective, the above-mentioned analysis has allowed to infer the variance-covariance matrix between additive genetic effects for traits analyzed in the HeterosYeast population. By using the well-known equation of response to selection (Chapter 1, section 1.1.4), it is possible to predict the results of one generation of selection. Consider for example table 6.1 in which are listed desirable fermentation traits for white wine production (Philippe Marullo, pers. comm.). It is possible to construct a selection index that considers the observed value for the selected traits and the associated weighting coefficient. The naive approach is to consider the weighted sum between these two quantities. Thus, selection can be performed on crosses showing an index value above a certain threshold and the calculation of the selection gradient is straightforward. The response to selection equation would return the average expected phenotypic value of offspring's at the next generation. It would also return the expected response to selection for traits that are not selected directly. Hence, using the method I proposed for the estimation of variance components, we could predict the evolution of non selected traits, including protein abundances, after one generation of selection.

	Objectives		**Weighting**	
Trait	**Blanc 18 garde**	**Blanc 18 primeur**	**Blanc 18 garde**	**Blanc 18 primeur**
AFtime	min	min	1	1
t.lag	min	min	1	1
Hexanol	low	low	0,25	0,25
Octanoic acid	low		0,1	
Phenylethanol-acetate	low	high	0,5	0,5
Isoamyl-acetate	low	high	0,5	0,5
Residual sugar	<2	<2		
Phenylethanol	low	high<400	0,5	0,5
4MMP	high	high	1	1
Decanoic acid	low		0,1	
SO_2L/SO_2T	high	high	0,5	0,5

Table 6.1: Objectives in white grape must fermentation. Objectives for traits of enological interest for grape must fermentation at 18 degrees for *garde* and *primeur* wines. To each objective is given a weighting coefficient based on enological interests. Objectives may change with the desired type of wine.

Also, the additive components of the variance covariance matrix associated to the HeterosYeast population correspond to the famous G matrix of the adaptive fitness landscapes. Eigenvectors associated to the G matrix would reveal the possible directions for evolution and could help understanding the geometry of yeast multitrait fitness landscape.

In general, the statistical model proposed in this first modelling approach can be employed in any problem concerning pairwise interactions between physical or biological entities. In ecology, the same model could be used to investigate competition between individuals for resources in a given environment, to access the performances in mixtures and to quantify the mixing ability for panels of genotypes, populations or species.

Integrative biology

In the second modelling approach the high-level phenotypes are understood as resulting of the integration processes of multiple cellular scales. In this context, I predicted an intermediary level of cellular organization: the metabolic fluxes. The general mathematical framework for metabolism modelling (through constraint-based models) and the methods classically used in the inference of metabolic fluxes are presented in Chapter 4. The proposed modelling approach is illustrated throughout Chapters 4 and 5 and consisted in interfacing quantitative proteomic data with constraint-based metabolic models relying on

- Genome annotations that allow genome-wide association of enzymatic reactions to gene expression/protein abundances.
- The hypothesis that protein abundance drives pathway usage and that, at the genome-scale, there should be a correlation between protein abundances and fluxes.

Contrary to the existing approach based on the same principles (Lee et al., 2012), my approach is fully data-driven and does not rely on any hypothesis about optimization principles of cell metabolism, that are questionable from an evolutionary point of view. It relies on a probabilistic description of the feasible space for fluxes, given stoichiometric and thermodynamic constraints (Braunstein et al., 2017), and further reduction by observations of cellular fluxes introduced as additional constraints. Then, amongst all possible solutions, we chose the one that best matches the observed distribution of protein abundances.

Future prospects would be to apply the method to a yeast genome-scale model (Heavner et al., 2013), but also to other biological systems. For instance, there is in our laboratory a huge collection of proteomic and phenotypic data collected on maize leaf at different developmental stages, and there exists a genome-scale metabolic model for the maize leaf (Simons et al., 2014A,B). Combining the data, the genome-scale model and the proposed method, I am confident that it would help in underpinning the molecular bases of leaf development variation.

Using as a toy model a reduction of yeast central carbon metabolism through the proofed DynamoYeast model (Chapter 4, section 4.3), I was able to show that introducing an additional layer of phenotypic integration, namely metabolic fluxes, between proteomic and observable traits, allowed me to better-understand the well-known ecological $r-K$ trade-off as a trade-off between metabolic pathway usages. To do this, I used cutting-edge statistical methods designated for heterogeneous datasets of high dimensionality, that proved to be efficient. The $r-K$ trade-off could thus be associated with different modes of glucose consumption rates (high or low). The "ant" strategy recalled in chapter 2 was associated to quick reproduction, high carrying capacity and small cell size in fermentation and low reproduction rate in respiration (chapter 2 section 2.2.1), but also to a low glucose consumption rate, possibly associated to higher fluxes in the pentose-phosphate pathways.

Metabolic choices of living species are a kind of puzzle far from being fully understood. The preliminary analysis performed to investigate the FBA strategy of a lower consumption rate of glucose have been revisited in this work in Chapter 2 section 4.3.3. By comparing the FBA solution to the feasible space reduced by experimental observations, I showed that the usage of pentose-phosphate pathway is a way of economizing resources, *i.e.* producing energy at lower price, in terms of glucose consumption. Further comparisons with alternative objective functions would be interesting to better understand of the underlying metabolic bases of the variation of phenotypic traits.

Beyond methodological development that may be useful for the scientific community (hopefully!), my thesis shows that mathematical and statistical modelling allied with the evolutionary framework helps understanding the diversity of the living world.

Bibliography

Braunstein, A., Muntoni, A. P. and Pagnani, A. (2017). An analytic approximation of the feasible space of metabolic networks, *Nature Communications* **8**: 14915.
URL: *http://www.nature.com/doifinder/10.1038/ncomms14915*

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996). Life with 6000 Genes, *Science* **274**(5287): 546–567.
URL: *https://science.sciencemag.org/content/274/5287/546*

Heavner, B. D., Smallbone, K., Price, N. D. and Walker, L. P. (2013). Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance, *Database: The Journal of Biological Databases and Curation* **2013**.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3739857/*

Lee, D., Smallbone, K., Dunn, W. B., Murabito, E., Winder, C. L., Kell, D. B., Mendes, P. and Swainston, N. (2012). Improving metabolic flux predictions using absolute gene expression data, *BMC Systems Biology* **6**(1): 73.
URL: *https://doi.org/10.1186/1752-0509-6-73*

Lenarcic, A. B., Svenson, K. L., Churchill, G. A. and Valdar, W. (2012). A general bayesian approach to analyzing diallel crosses of inbred strains, *Genetics* **190**(2): 413–435.

Simons, M., Saha, R., Amiour, N., Kumar, A., Guillard, L., Clément, G., Miquel, M., Li, Z., Mouille, G., Lea, P. J., Hirel, B. and Maranas, C. D. (2014B). Assessing the Metabolic Impact of Nitrogen Availability Using a Compartmentalized Maize Leaf Genome-Scale Model, *Plant Physiology* **166**(3): 1659–1674.
URL: *http://www.plantphysiol.org/content/166/3/1659*

Simons, M., Saha, R., Guillard, L., Clément, G., Armengaud, P., Cañas, R., Maranas, C. D., Lea, P. J. and Hirel, B. (2014A). Nitrogen-use efficiency in maize (Zea mays L.): from 'omics' studies to metabolic modelling, *Journal of Experimental Botany* **65**(19): 5657–5671.

Appendix A

Appendix A

Supplementary materials for Petrizzelli et al. 2019

A.1 Subcompositional dominance and distances

We consider the central log-ratio transformation in order to pursue our analysis without considering both the block effect and residuals for the more integrated traits and residuals for protein abundances. We are aloud to do so since the clr-transformation satisfies the subcompositional dominance property, *i.e.*, for each couple of vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, and for each pair of subvectors $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, obtained by selecting the same set of components, the distance between the subvectors is always less than or equal to the distance between the original vectors, *i.e.*

$$d(\boldsymbol{x}, \boldsymbol{y}) \geq d(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \tag{S1}$$

Therefore, for each $\boldsymbol{z}$ such that $d(\boldsymbol{x}, \boldsymbol{y}) \geq d(\boldsymbol{x}, \boldsymbol{z})$, we have that, dividing eq.(S1) by $\frac{d(\boldsymbol{x},\boldsymbol{z})}{d(\hat{\boldsymbol{x}},\hat{\boldsymbol{z}})} \geq 1$

$$\alpha d(\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) \geq k d(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \tag{S2}$$

where $\alpha = \frac{d(\boldsymbol{x},\boldsymbol{y})}{d(\boldsymbol{x},\boldsymbol{z})} \geq 1$ and $k = \frac{d(\hat{\boldsymbol{x}},\hat{\boldsymbol{z}})}{d(\boldsymbol{x},\boldsymbol{z})} \leq 1$. So, since $k/\alpha \leq 1$

$$d(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \geq d(\hat{\boldsymbol{x}}, \hat{\boldsymbol{z}}) \tag{S3}$$

As a consequence, distance relationship between the original vectors is preserved by selected subvectors.

A.2 The fitting algorithm

The hglm package implements the estimation algorithm for hierarchical generalized linear models. It fits generalized linear models with random effects, where the random effect may come from a conjugate exponential-family distribution (Gaussian, Gamma, Beta or inverse-Gamma) and it is possible to explicitly specify the design matrices both for the fixed and random effects, which allows fitting correlated random effects as well as random regression models.
In order to perform the diallel analysis, we considered $\mathbf{y}$, the vector of observations for the trait of interest, and we re-wrote the model (eq.(1.56)) in matrix a form:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon} \tag{S4}$$

where X is the design matrix for the fixed effects, Z the design matrix for the random effects, $\boldsymbol{\beta} = (\mu, \beta_{S.uvarum}, \beta_{S.cerevisiae})$ and $\boldsymbol{u} = (\boldsymbol{A_w}, \boldsymbol{A_b}, \boldsymbol{B}, \boldsymbol{H_w}, \boldsymbol{H_b})$ are respectively the vectors of fixed effects parameters and random effects parameters, and $\boldsymbol{\epsilon}$ is the vector of random errors. With this notation, the construction of the model is straight forward since we just have to construct the design matrices for both fixed and random effects.

Let n be the number of observations, J the total number of parental strains, N_{intra} (resp. N_{inter}) the number of intra-specific (resp. inter-specific) crosses, and K the total number of random effects

parameters. X is a $n \times 3$ matrix, with, by construction, the first column equal to $(1, 1, ..., 1)$, while the elements of the second and third columns (for respectively *S. uvarum* and *S. cerevisiae*) are 1 or 0 depending on whether the strain is inbred and or not.
Z will be a $n \times K$ matrix and, more precisely, it can be thought as the following block matrix:

$$Z = \left[\begin{array}{c:c:c:c:c} Z_{\boldsymbol{A_w}} & Z_{\boldsymbol{A_b}} & Z_{\boldsymbol{B}} & Z_{\boldsymbol{H_w}} & Z_{\boldsymbol{H_w}} \end{array} \right] \tag{S5}$$

where $Z_{\boldsymbol{A_w}}$, $Z_{\boldsymbol{A_b}}$, $Z_{\boldsymbol{B}}$, $Z_{\boldsymbol{H_w}}$, $Z_{\boldsymbol{H_b}}$ denote the design matrices, respectively, of the random effect parameters $\boldsymbol{A_w}$, $\boldsymbol{A_b}$, $\boldsymbol{B}$, $\boldsymbol{H_w}$ and $\boldsymbol{H_b}$. In particular, $Z_{\boldsymbol{A_w}}$, $Z_{\boldsymbol{A_b}}$, $Z_{\boldsymbol{B}}$ are $n \times J$ matrices, $Z_{\boldsymbol{H_w}}$ is a $n \times N_{intra}$ matrix and $Z_{\boldsymbol{H_b}}$ is a $n \times N_{inter}$ matrix with entries:

$$z_{\boldsymbol{A_{w_{ij}}}} = \begin{cases} 2 & \text{If the } i\text{-observation belongs to a parental strain, the } j\text{-th;} \\ 1 & \text{If the } i\text{-observation belongs to an hybrid achieved through an intra- specific cross in which the parental strain } j \text{ is involved;} \\ 0 & \text{otherwise;} \end{cases} \tag{S6}$$

$$z_{\boldsymbol{A_{b_{ij}}}} = \begin{cases} 1 & \text{If the } i\text{-observation belongs to an hybrid achieved through an inter- specific cross in which the parental strain } j \text{ is involved;} \\ 0 & \text{otherwise;} \end{cases} \tag{S7}$$

$$z_{\boldsymbol{B_{ij}}} = \begin{cases} 1 & \text{If the } i\text{-observation belongs to a parental strain, the } j\text{-th;} \\ 0 & \text{otherwise;} \end{cases} \tag{S8}$$

and, enumerating the intra-specific/inter-specific hybrid strains with k_{intra}/k_{inter} from 1 to N_{intra}/N_{inter}, respectively,

$$z_{\boldsymbol{H_{w_{ik_{intra}}}}} = \begin{cases} 1 & \text{If the } i\text{-observation belongs to the } k_{intra}\text{- hybrid strain;} \\ 0 & \text{otherwise;} \end{cases} \tag{S9}$$

$$z_{\boldsymbol{H_{b_{ik_{inter}}}}} = \begin{cases} 1 & \text{If the } i\text{-observation belongs to the } k_{inter}\text{- hybrid strain;} \\ 0 & \text{otherwise;} \end{cases} \tag{S10}$$

A.3 Half-diallel simulation construction

In order to elucidate our findings about the decoupling of inbreeding and heterotic variances, we simulated a half-diallel between N parental strains. We supposed the phenotypic values of each trait to depend on a fixed number of loci, L, and we considered all the possible combinations of genetic effects, namely presence/absence of dominance, of additive × additive epistasis and of additive × additive epistasis.
We let the number of alleles at each locus to vary between 1 and N and we drew values for allele a at locus i (a_i) from a Gamma distribution ($\Gamma(k, \theta)$), for additive × additive epistatic effect between a_i and a_j (aa^{ij}) and for dominance × dominance epistatic effect (dd^{ij}) from a Gaussian distribution ($\mathcal{N}(0, \sigma^2)$). The dominance effect between alleles a and b at locus i (d^i_{ab}) are drawn from an uniform distribution $\mathcal{U}(0, m)$ with $m = 0.5$ for dominance of the strongest allele, and $m = 1$ for symmetrical dominance. Therefore, the phenotypic value of the parental lines P_k and

of the hybrid, H_{lk}, between parents P_k and P_l are given by:

1) Additive model

$$y_{P_k} = 2\sum_i k_i\,, \quad y_{H_{lk}} = \sum_i k_i + \sum_i l_i \tag{S11}$$

2) Additive model plus dominance

$$y_{P_k} = 2\sum_i k_i\,, \quad y_{H_{lk}} = \sum_i k_i + \sum_i l_i + \sum_i d^i_{kl} \tag{S12}$$

3) Additive model plus *additive* × *additive* effect

$$y_{P_k} = 2\sum_i k_i + \sum_{ij} aa^{ij}\,, \quad y_{H_{lk}} = \sum_i k_i + \sum_i l_i \tag{S13}$$

4) Additive model plus *dominance* × *dominance* effect

$$y_{P_k} = 2\sum_i k_i\,, \quad y_{H_{lk}} = \sum_i k_i + \sum_i l_i + \sum_{ij} dd^{ij} \tag{S14}$$

5) Additive model plus *additive* × *additive* and *dominance* × *dominance* effect

$$y_{P_k} = 2\sum_i k_i + \sum_{ij} aa^{ij}\,, \quad y_{H_{lk}} = \sum_i k_i + \sum_i l_i + \sum_{ij} dd^{ij} \tag{S15}$$

6) Additive model plus dominance and *additive* × *additive* effect

$$y_{P_k} = 2\sum_i k_i + \sum_{ij} aa^{ij}\,, \quad y_{H_{lk}} = \sum_i k_i + \sum_i l_i + \sum_i d^i_{kl} \tag{S16}$$

7) Additive model plus dominance and *dominance* × *dominance* effect

$$y_{P_k} = 2\sum_i k_i\,, \quad y_{H_{lk}} = \sum_i k_i + \sum_i l_i + \sum_i d^i_{kl} + \sum_{ij} dd^{ij} \tag{S17}$$

8) Additive model plus dominance, *additive* × *additive* and *dominance* × *dominance* effect

$$y_{P_k} = 2\sum_i k_i + \sum_{ij} aa^{ij}\,, \quad y_{H_{lk}} = \sum_i k_i + \sum_i l_i + \sum_i d^i_{kl} + \sum_{ij} dd^{ij} \tag{S18}$$

A.4 Inbreeding depression and heterosis variances are equal in three-parent diallel

Inbreeding and heterosis variances are equal in the particular case of a three-parent diallel when no maternal effect is present. It can be easily seen by the direct computation of their value.
In order to do that we decompose the phenotypic values of the $i-$parent, P_i, as

$$P^d_i = \mu + 2A_i \tag{S19}$$

and of the $i \times j$ hybrid, H_{ij}, as

$$H^d_{ij} = \mu + A_i + A_j \tag{S20}$$

where $\mu = \frac{1}{6}(P_1 + P_2 + P_3 + H_{12} + H_{13} + H_{23})$ is the mean phenotypic value of the population and

$$A_i = \frac{1}{3}(P_i + \sum_{j \neq i} H_{ij}) - \mu \tag{S21}$$

the GCA of strain i. Therefore, we can express the inbreeding depression variance as the deviation of the decomposed phenotypic value of the parents, P^d, and their true value P

$$Var(inbreeding) = \frac{1}{3}\sum_{i}(P_i^d - P_i - (\overline{P^d - P}))^2 \tag{S22}$$

and the heterosis variance analogously

$$Var(heterosis) = \frac{1}{3}\sum_{i<j}(H_{ij}^d - H_{ij} - (\overline{H^d - H}))^2 \tag{S23}$$

In which we have used the fact that $H_{ij} = H_{ji}$, since no maternal effects are present.

Substituting S19 and S21 in S22, we get

$$\begin{aligned}
Var(inbreeding) &= \frac{1}{3}\sum_{i=1}^{3}(\mu + 2A_i - P_i - \frac{1}{3}\sum_{k=1}^{3}(\mu + 2A_k - P_k))^2 = \\
&= \frac{1}{3}\sum_{i=1}^{3}(\mu + \frac{2}{3}(P_i + \sum_{j\neq i} H_{ij}) - 2\mu - P_i - \frac{1}{3}\sum_{k=1}^{3}(\mu + \frac{2}{3}(P_k + \sum_{j\neq k} H_{kj}) - 2\mu - P_k))^2 = \\
&= \frac{1}{243}\sum_{i=1}^{3}(6P_i + 6\sum_{j\neq i} H_{ij} - 9\mu - 9P_i - \sum_{k=1}^{3}(2P_k + 2\sum_{j\neq k} H_{kj} - 3\mu - 3P_k))^2 = \\
&= \frac{1}{243}\sum_{i=1}^{3}(6\sum_{j\neq i} H_{ij} - 9\mu - 3P_i + \sum_{k=1}^{3} P_k - 4\sum_{j<k} H_{kj} + 9\mu)^2 = \\
&= \frac{1}{243}\sum_{i=1}^{3}(-2P_i + P_j + P_k + 2H_{ik} + 2H_{ij} - 4H_{kj})^2
\end{aligned} \tag{S24}$$

where $i \neq j \neq k$. In the same way, substituting S20 and S21 in S23, we get

$$\begin{aligned}
Var(heterosis) &= \frac{1}{3}\sum_{i<j}(\mu + A_i + A_j - H_{ij} - \frac{1}{3}\sum_{k<m}(\mu + A_k + A_m - H_{km}))^2 = \\
&= \frac{1}{3}\sum_{i<j}(\frac{1}{3}(P_i + P_j + \sum_{k\neq i} H_{ik} + \sum_{k\neq j} H_{jk}) - \mu - H_{ij} - \\
&- \frac{1}{3}\sum_{k<m}(\frac{1}{3}(P_k + P_m + \sum_{l\neq k} H_{kl} + \sum_{l\neq m} H_{ml}) - \mu - H_{km}))^2 = \\
&= \frac{1}{243}\sum_{i<j}(3(P_i + P_j + \sum_{k\neq i} H_{ik} + \sum_{k\neq j} H_{jk}) - 9\mu - 9H_{ij} - (2\sum_{k} P_k + \sum_{k<m} H_{km} - 9\mu))^2 = \\
&= \frac{1}{243}\sum_{i<j}(3(P_i + P_j + \sum_{k\neq i} H_{ik} + \sum_{k\neq j} H_{jk}) - 9H_{ij} - 2\sum_{k} P_k - \sum_{k<m} H_{km})^2 = \\
&= \frac{1}{243}\sum_{i<j}(P_i + P_j - 2P_k - 4H_{ij} + 2H_{ik} + 2H_{jk})^2
\end{aligned} \tag{S25}$$

where again $i \neq k \neq j$. Therefore,

$$\begin{aligned} Var(inbreeding) = \\ = \frac{1}{243}\sum_{i=1}^{3}(-2P_i + P_j + P_k + 2H_{ik} + 2H_{ij} - 4H_{kj})^2 = \\ = \frac{1}{243}((-2P_1 + P_2 + P_3 + 2H_{12} + 2H_{13} - 4H_{23})^2 + (-2P_2 + P_1 + P_3 + 2H_{12} + 2H_{23} - 4H_{13})^2 + \\ +(-2P_3 + P_1 + P_2 + 2H_{13} + 2H_{23} - 4H_{12})^2 = \frac{1}{243}\sum_{i<j}(-2P_k + P_i + P_j + 2H_{ik} + 2H_{jk} - 4H_{ij})^2 = \\ = Var(heterosis) \end{aligned} \quad \text{(S26)}$$

A.5 Structuration of genetic variability at the fermentation trait level

A Gaussian mixture model is run to classify life-history and fermentation traits according to their genetic variance components.

The best model clearly identify three clusters (fig.S3 and fig.S6). Cluster 1 (99.9% of good assignments) is composed by 9 traits, characterized by having null inter-specific additive variance component, relatively low inter-specific heterosis variance and high intra-specific additive and inbreeding components. In this cluster we can find most volatile compounds such as *Octanoic acid* and *Hexanol* at both temperatures, *Phenyl-2-ethanol, Phenyl-2-ethanol acetate* and *Decanoic acid* at 18°C, the kinetic parameter CO_{2max} and the life-history trait *Size-t-*N_{max} at 26°C. Cluster 2 (98.9% of good assignments) consists of 28 traits that are characterized by high inter-specific additive and inbreeding components ($\sigma^2_{\boldsymbol{A_b}}$ and $\sigma^2_{\boldsymbol{B}}$), relatively low heterosis ($\sigma^2_{\boldsymbol{H_w}}$ and $\sigma^2_{\boldsymbol{H_b}}$) and intra-specific additive variances ($\sigma^2_{\boldsymbol{A_w}}$). Most kinetic parameters and life-history traits belongs to this cluster: *t-lag*, V_{max}, *t-45*, *r*, *t-*N_{max}, J_{max} and *Viability-t-*N_{max} at both temperatures; *t-*V_{max} and *t-75* at 26°C; *AFtime*, *t-*N_0, *Size-t-*N_{max} at 18°C. We can also find some basic enological parameters and aromatic traits - *Isoamyl acetate* and *Hexanoic acid* at both temperatures; *Phenyl-2-ethanol* and *Phenyl-2-ethanol acetate* at 26°C; *X4MMP*, *Free* SO_2 and *Total* SO_2 at 18°C. Traits attributed to cluster 3 (19 traits, 97.3% of good assignments) have high additive and heterotic variances and null inbreeding variance. The rest of the basic enological parameters and aromatic traits along with some kinetics parameters and life-history traits belongs to it.

As for protein abundances, we choose to consider life-history and fermentation traits at two temperatures (18°C and 26°C) as different traits. Indeed, after computation of genetic variance components for each trait, correlations between temperatures are not found to be significant except for 6 traits (*t-*V_{max}, *t-45*, *r*, *t-*N_{max}, *Viability-t-*N_{max} and *Hexanol*) that are highly and positively correlated. All of them fall in the same cluster at the two temperatures, except *t-*V_{max}. Overall, we find that 79% of traits do not belong to the same cluster at the two temperatures. Further, Pearson's correlation tests are performed to investigate the correlation between genetic effects at the two temperatures. They were not significant except for the additive inter-specific component ($cor = 0.74$, p-value<0.05). Therefore, at the fermentation trait level, genotype by environment interactions predominate.

Globally, correlations between variance components, when present, are found to be negative (fig.S4). However, the pattern changes when considering intra-group correlations. Indeed, in cluster 2, even if inbreeding is negatively correlated to the heterotic variances, it is positively correlated to the additive inter-specific variance, and in cluster 3, additive genetic variances are positively correlated to each other. In cluster 1, there is no statistical significant correlation between genetic effects (fig.S7).

Therefore, we can state that three well defined groups of traits can be differentiated according to their genetic variance profiles and we show that the part of phenotypic variation explained by the model's parameters depends on trait's category and temperature: in cluster 1, we can find

mostly aromatic traits; in cluster 2 kinetics parameters and life-history traits and in cluster 3 most enological parameters. Further, closely related phenotypes show similar profiles in terms of variance components, such as CO_{2max}, *Ethanol* and *Residual Sugar* that clusters together at 18°C; *Total* SO_2 and *Free* SO_2 are found in cluster 2 at 18°C and in cluster 3 at 26°C; *t-*N_0 and *t-lag* in cluster 2 at 18°C. We finally see that inbreeding variance can be either negatively, or not correlated to heterotic effects.

A.6 Strain characterization

We characterized the strains based on their genetic contribution to the total phenotypic value of a trait at a certain temperature (fig. S11). Strain D1 is found to be the strain with the lowest additive contribution for *Phenyl-2-ethanol* at both temperatures and for *Sugar.Ethanol.Yield* (except in inter-specific crosses at 18°C), with the highest additive intra-specific contribution for *Decanoic* and *Octanoic acid*, while displaying the highest heterosis contribution for *Octanoic acid* when crossed with E2 at 18°C, with E5 and U1 at 26°C, and for *Decanoic acid* when crossed with E4 at 26°C and U2 at 18°C. D2 and E2 strains have the highest or lowest additive contributions across almost all traits, mostly fermentation kinetics parameters and life history traits. In particular, D2 strain shows the highest intra- and inter-specific additive effects, and inbreeding values for *t.45*, *t.75* and *AFtime* at both temperatures, where the highest heterosis effect is achieved when crossed with E2, U1 for *t*.45 at 18°C, with E5 and U1 for *t*.75 with the first at both temperatures and the latter at 18°C. Similarly, the additive intra-specific effect of U4 is the highest or the lowest for almost all aromatic traits at 18°C (higher for *Phenyl-2-ethanol, Hexanol* and *Hexanoic acid*; lowest for *Decanoic acid* and *Octanoic acid*). Strain U1 shows the highest additive inter-specific effect in aromatic traits at 26°C (*Phenyl-2-ethanol, Phenyl-2-ethanol acetate, Hexanol, Hexanoic acid* and *Octanoic acid*). In particular, the heterosis effect in the inter-specific cross with strain D2 is the highest for *Hexanol* and with strain E2 for *Phenyl-2-ethanol*. For all traits, E5 produces intermediate heterosis values when crossed with E2, E3, E4, W1, U1 and U4 at 18°C, but its cross with E4 results in the highest heterosis value for $t.N_{max}$, and the lowest for *Decanoic acid* with E3 and for *Total* SO_2 with W1 at 26°C. In the same way, crosses between E3 and U1, U2 or U3, between E4 and U1 or W1 never show extreme heterosis values for any trait.

A.7 Supplementary tables

Table S1: Diallel table representing the mitochondrial inheritance for each phenotyped cross: the data clearly shows too many *unknown*s to enter a mitochondrial effect in the model. Backslashes indicate the not phenotyped reciprocals.

P1\ P2	D1	D2	E2	E3	E4	E5	W1	U1	U2	U3	U4
D1	D1	D2	unknown	\	unknown	\	unknown	\	U2	U3	U4
D2	\	D2	E2	\	E4	\	W1	\	\	\	\
E2	\	\	E2	unknown	unknown	E5	unknown	\	U2	U3	U4
E3	D1	D2	\	E3	unknown	\	W1	\	U2	U3	\
E4	\	\	\	\	E4	E5	W1	U1	U2	U3	U4
E5	D1	unknown	\	unknown	\	E5	unknown	U1	U2	U3	\
W1	\	\	\	\	\	\	W1	\	U2	U3	\
U1	D1	D2	E2	E3	\	\	CW1	U1	\	\	\
U2	\	D2	\	\	\	\	\	unknown	U2	\	\
U3	\	D2	\	\	\	\	\	unknown	unknown	U3	\
U4	\	unknown	\	E3	\	E5	W1	unknown	unknown	unknown	U4

Table S2: Pearson's chi-square. For each cluster and at each temperature (18° and 26°) we tested for enrichment in proteins belonging to a certain functional category using as prior probability the frequency of proteins functional category based on MIPS database. In yellow (resp. pink) are highlighted the functional category enhanced (resp. depleted) for each cluster and at each temperature when the statistical test was significant.

Cluster	T	P-val	Metabolism (other)	Amino acid metabolism	Nucleotide metabolism	Carbon metabolism	Replication	Transcription	Protein synthesis	Protein fate	Transport	Stress	Mating	Signal	Unknown
1	18	0.873	-0.092	0.592	-0.580	0.296	-0.637	-0.845	0.859	1.091	1.214	-0.591	-0.880	-0.456	-1.021
1	26	0.000	-1.853	-0.053	-0.230	-1.990	-1.102	2.009	-2.702	-2.022	-1.458	10.183	5.900	0.446	-2.336
2	18	0.000	-0.468	3.799	0.519	3.424	-1.970	-1.784	-1.269	3.335	-2.024	2.721	-2.322	-1.409	-2.094
2	26	0.000	-0.431	3.850	0.543	3.477	-1.425	-1.764	-1.235	2.697	-2.000	2.186	-2.303	0.061	-2.425
3	18	0.042	-0.062	2.012	2.550	1.332	-1.239	-0.327	-1.193	-0.387	-1.379	1.564	0.829	-0.886	-1.421
3	26	0.001	0.428	-0.053	1.100	-0.199	-1.102	-1.319	-0.466	3.951	-1.458	2.989	-1.447	0.446	-1.929
4	18	0.026	-0.532	3.519	0.755	-1.109	-1.220	-0.283	0.424	1.345	-1.341	1.622	-0.397	-0.873	-1.383
4	26	0.001	-0.206	2.549	0.301	-1.490	-0.691	-0.175	-0.319	1.483	-1.747	4.180	-0.864	0.989	-1.793
5	18	0.001	2.239	2.060	3.638	0.219	-1.393	-1.262	2.081	-1.803	-1.685	-0.488	-1.359	0.031	-1.731
5	26	0.001	2.063	2.300	2.324	-0.165	-1.645	-2.067	2.744	-0.319	-0.712	-0.931	-1.107	-0.856	-2.440
6	18	0.008	-2.145	0.044	2.562	1.287	-1.048	-1.243	3.483	0.101	-0.943	-0.882	-0.900	0.515	-0.596
6	26	0.118	-1.933	0.855	2.056	1.939	-0.617	-0.642	1.237	-0.284	-0.114	0.350	-1.330	1.094	-1.193
7	18	0.000	0.891	6.649	-1.135	4.290	-1.673	-3.194	1.885	-0.304	-2.198	1.672	-1.634	-2.007	-3.807
7	26	0.000	1.379	7.162	-0.865	5.001	-1.397	-3.095	1.398	-1.032	-2.895	1.776	-1.220	-2.237	-4.176
8	18	0.000	-3.121	-2.200	2.324	-2.679	-1.645	0.603	-2.935	7.348	-2.367	7.463	3.682	-0.186	-3.093
8	26	0.000	-2.899	-2.046	2.651	-2.506	-1.510	0.529	-2.693	5.751	-2.143	7.516	4.186	-0.743	-2.902
9	18	0.000	-1.884	-0.085	-0.252	-1.568	-1.120	1.959	-2.728	-2.050	-1.488	10.081	5.819	0.424	-2.360
9	26	0.013	-1.032	1.433	2.002	-0.882	-0.643	2.830	-1.584	-0.824	2.372	-0.488	-0.230	0.031	-1.731

A.8 Supplementary figures

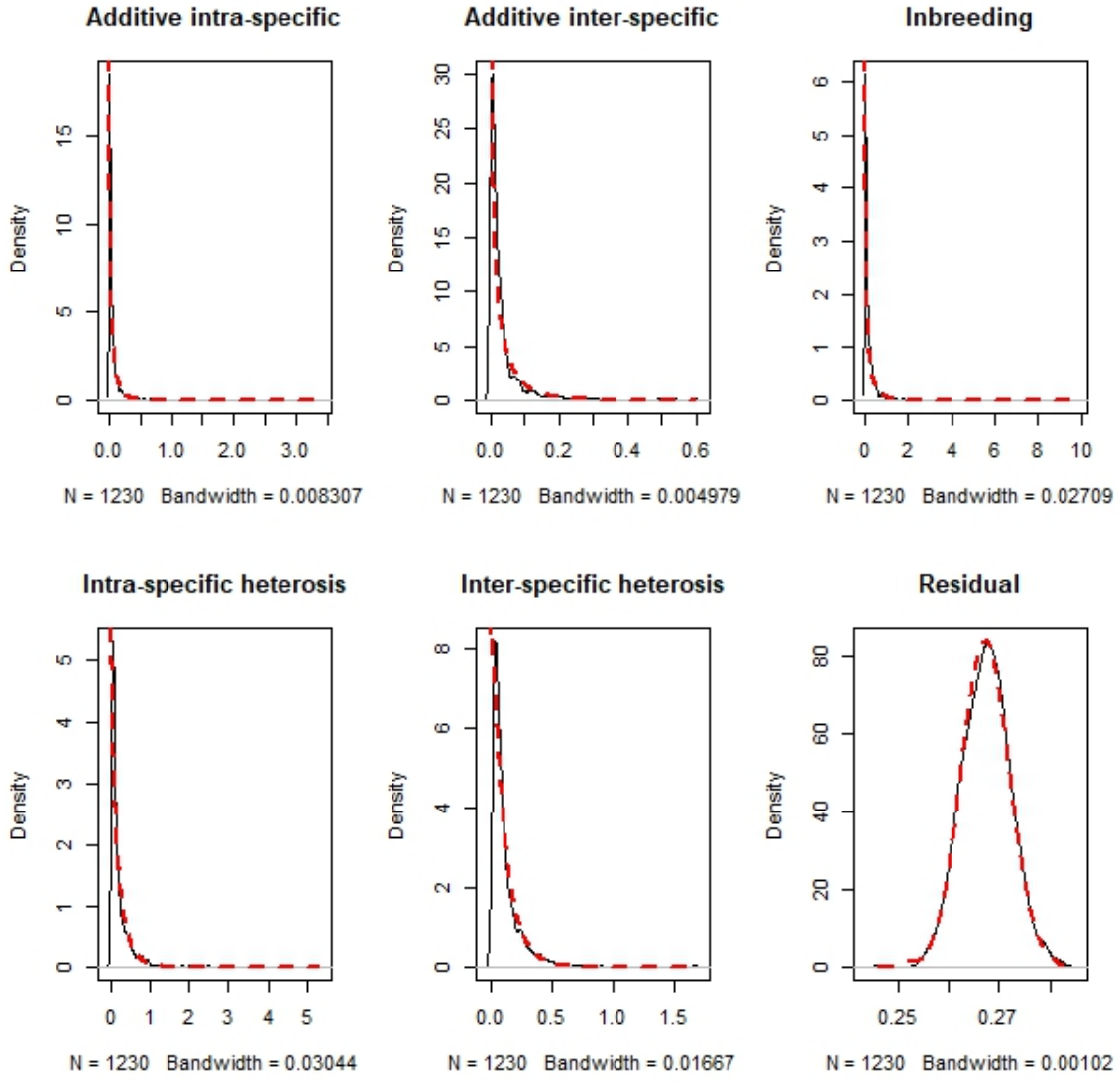


Figure S1: Density of the variance components estimated by the $hglm$ algorithm for the 1230 proteins. Red dashed lines represent the fitted distributions used to simulate and test parameter inference of the proposed model.

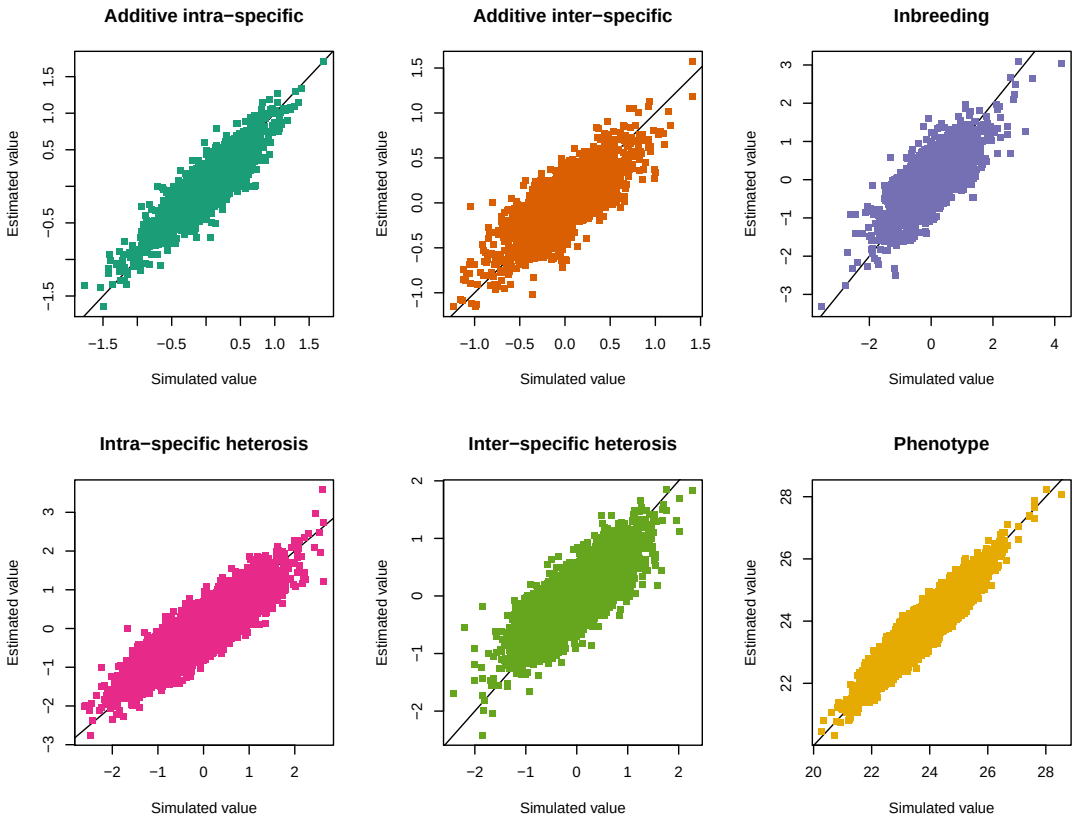


Figure S2: Fitted Best Linear Unbiased Predictors of the random effects parameters and predicted phenotypic value plotted against the simulated genetic parameters and the simulated phenotypic value. Fixed the number of parental strains and the number of individuals of each species, we performed the simulation 1000 times. Here, we show the case of eleven parents, with 7 belonging to one specie and 4 to the other.

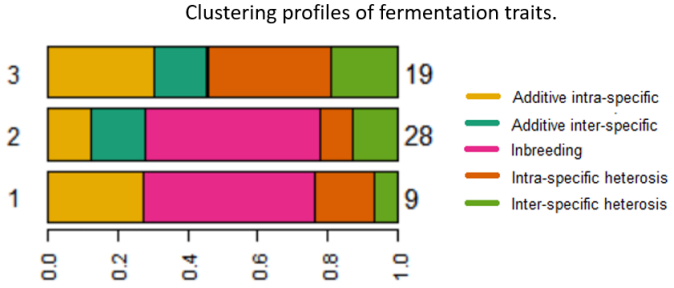


Figure S3: Clustering profiles of fermentation and life-history traits. Clusters number are reported on the left, on the right the number of traits found in each cluster.

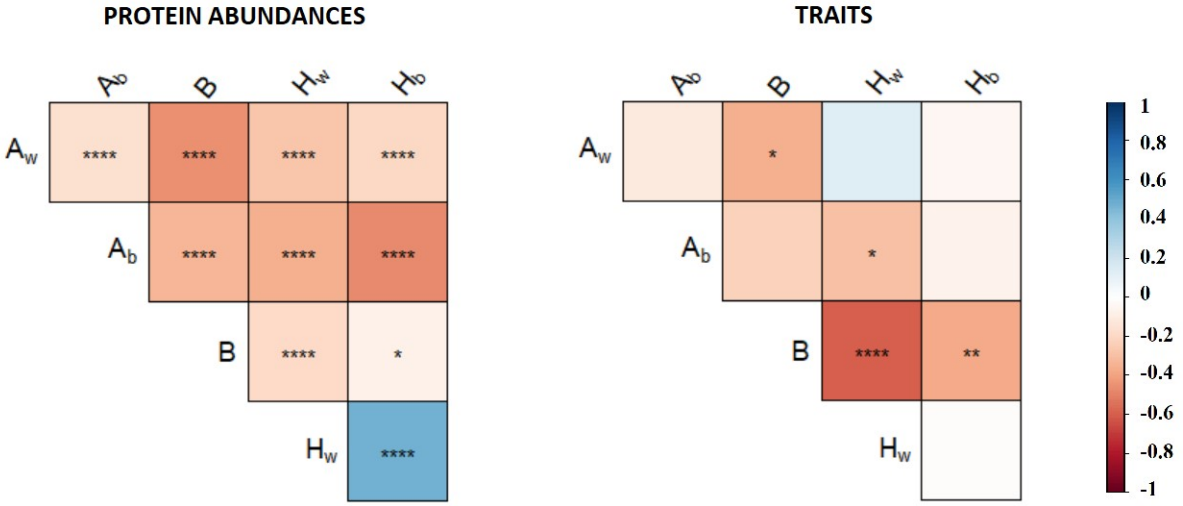


Figure S4: Global correlations between genetic variance components: on the left correlations at the proteomic level, on the right at the more integrated level. * significant at $p < 0.05$; ** significant at $p < 5 \cdot 10^{-3}$; *** significant at $p < 5 \cdot 10^{-4}$; **** significant at $p < 5 \cdot 10^{-5}$. No symbol: not significant.

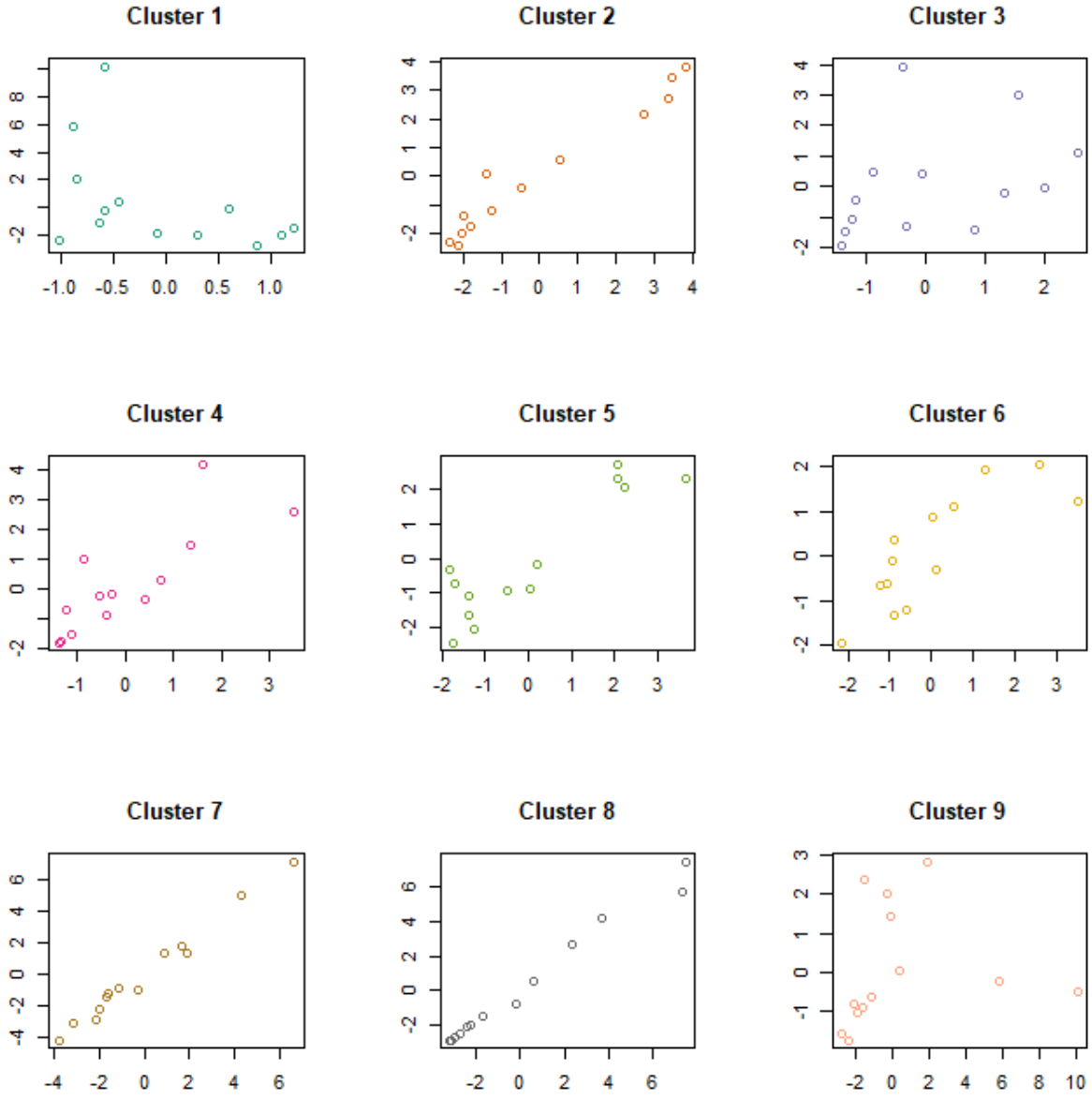


Figure S5: Pearson's chi-square test of enrichment: For each cluster are represented the chi-square standardized residuals at 18° (abscissa) and at 26° (ordinate).

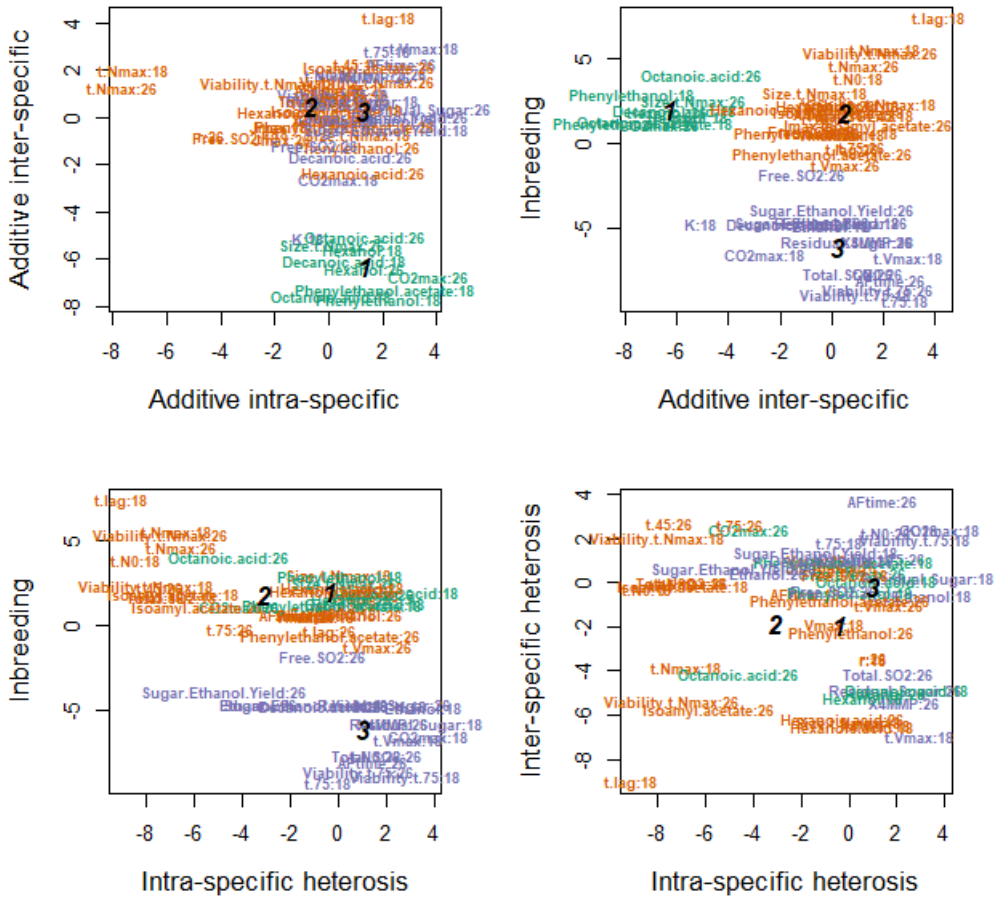


Figure S6: Life-history and fermentation traits profiles. Traits are identified by their label, color combinations identify the clusters obtained by their classification based on a Gaussian Mixture model.

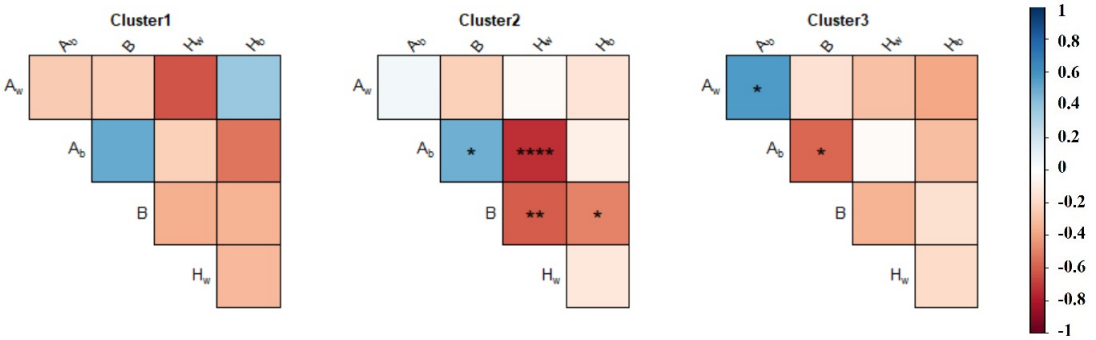


Figure S7: Pearson's correlation test performed to investigate the intra-cluster correlations at the trait level: for each cluster, the figure shows the correlation between variances of the genetic effects. * significant at $p < 0.05$; ** significant at $p < 5 \cdot 10^{-3}$; *** significant at $p < 5 \cdot 10^{-4}$; **** significant at $p < 5 \cdot 10^{-5}$. No symbol: not significant.

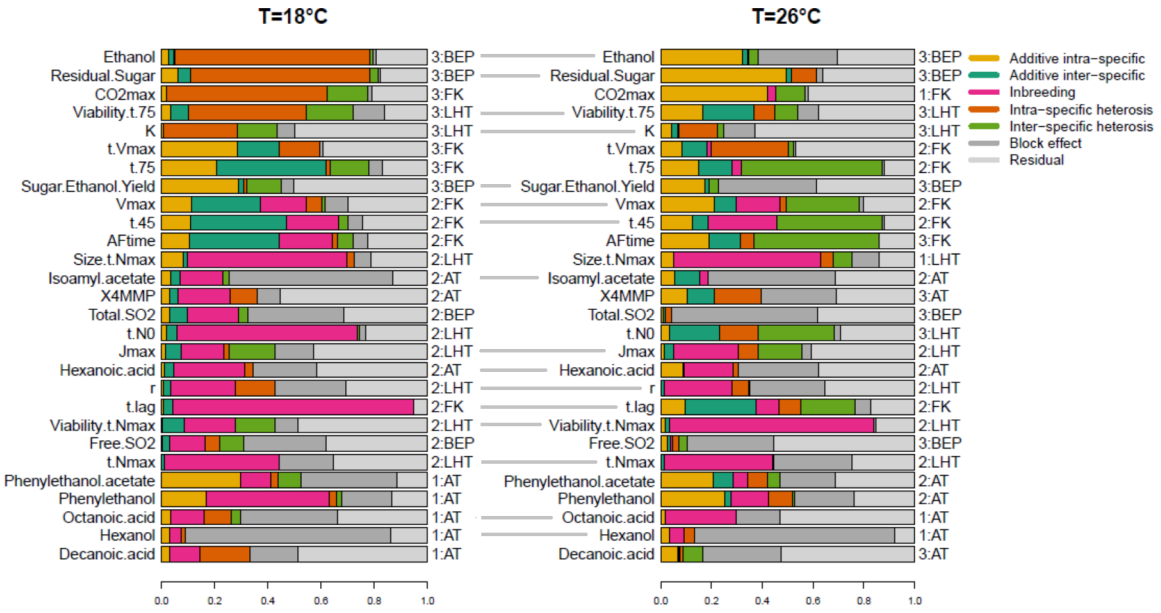


Figure S8: Variance components of fermentation traits. Left: Traits measured at 18°C. Right: Traits measured at 26°C. Each variance component is attributed a different color. Traits are ranked according to their cluster number at 18°C. Trait category and cluster number is indicated on the right-hand-side of the plot.

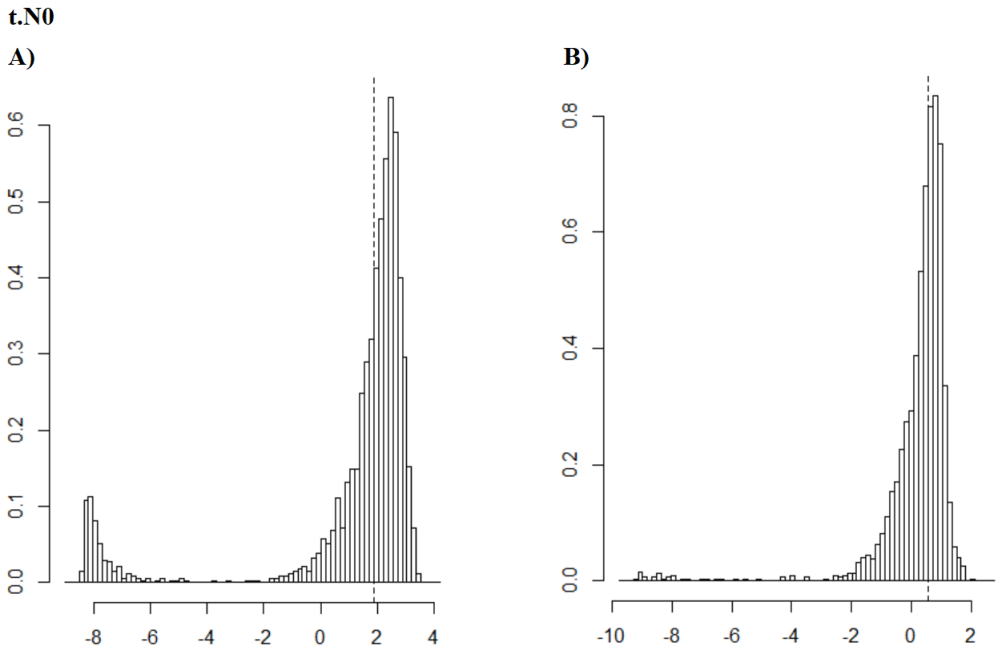


Figure S9: Bootstrap summary example: Distribution of intra-specific variance estimates for the growth lag-phase, *t.N0*, at A) 18° and B) 26°C.

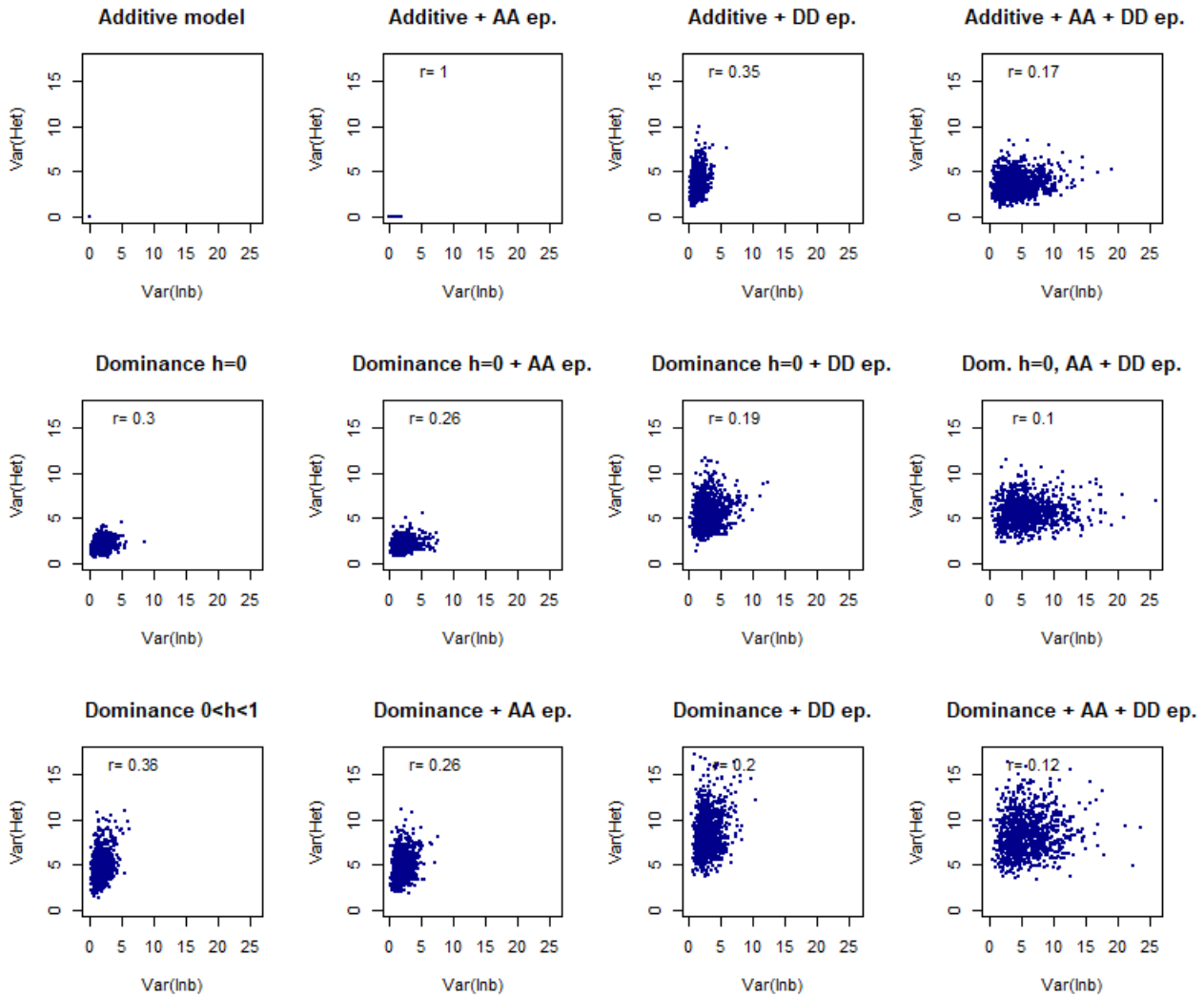


Figure S10: Simulations genetic models. Correlation between inbreeding and heterosis variance computed on the basis of a combination of multilocus genetic models: Additive model with/without additive by additive, dominance by dominance, or both epistatic effects, dominance of the strongest allele with/without additive by additive, dominance by dominance, or both epistatic effects; symmetrical dominance with/without additive by additive, dominance by dominance, or both epistatic effects. The simulated half-diallel consisted of 11 parental lines. Phenotypic values were supposed to depend on 10 loci, and the number of alleles per loci was imposed to 11. Alleles values were drawn from a gamma distribution (k=10, θ=20) and epistatic effects from a normal distribution ($\mathcal{N}(0,3)$).

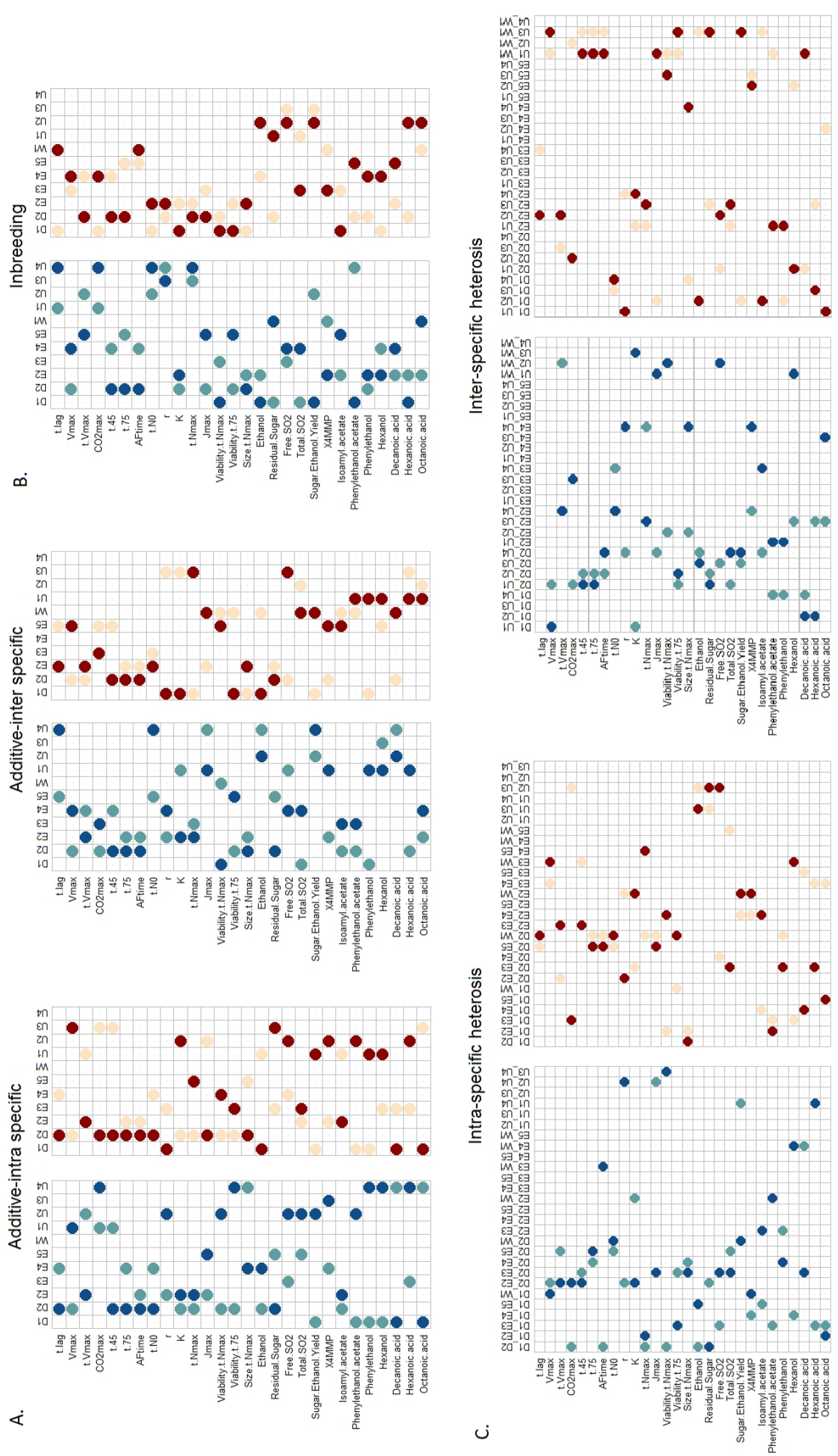
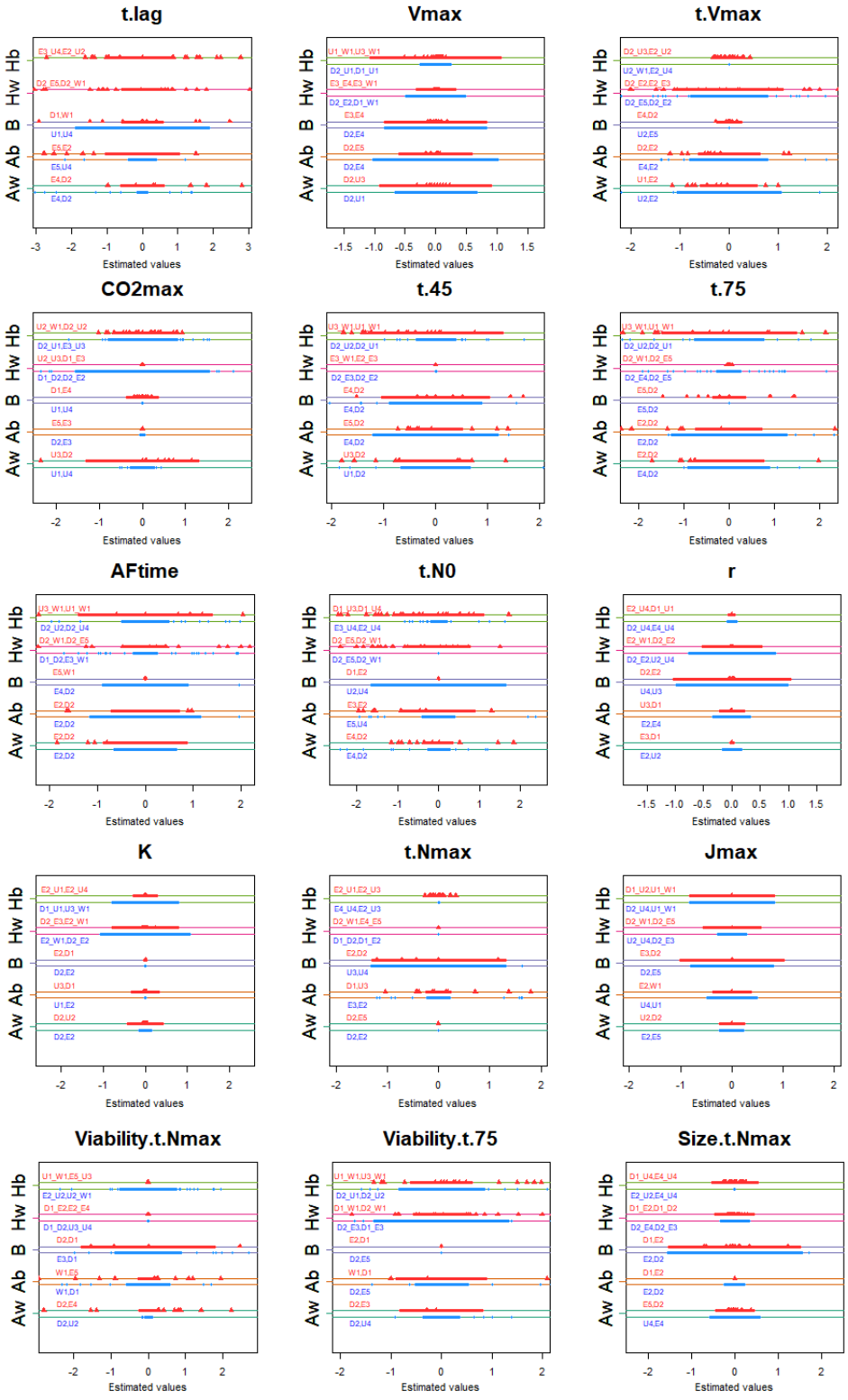


Figure S11: Dots show strains with highest and lowest genetic contribution per trait and temperature, in blue at 18°C and in red at 26°C. Dark and light colors report the strains with the highest and lowest additive (A), inbreeding (B) and heterosis (C) contributions, respectively. Strains are sorted by species and traits by category.


t.lag
Vmax
t.Vmax
CO2max
t.45
t.75
AFtime
t.N0
r
K
t.Nmax
Jmax
Viability.t.Nmax
Viability.t.75
Size.t.Nmax
Aw Ab B Hw Hb
Estimated values

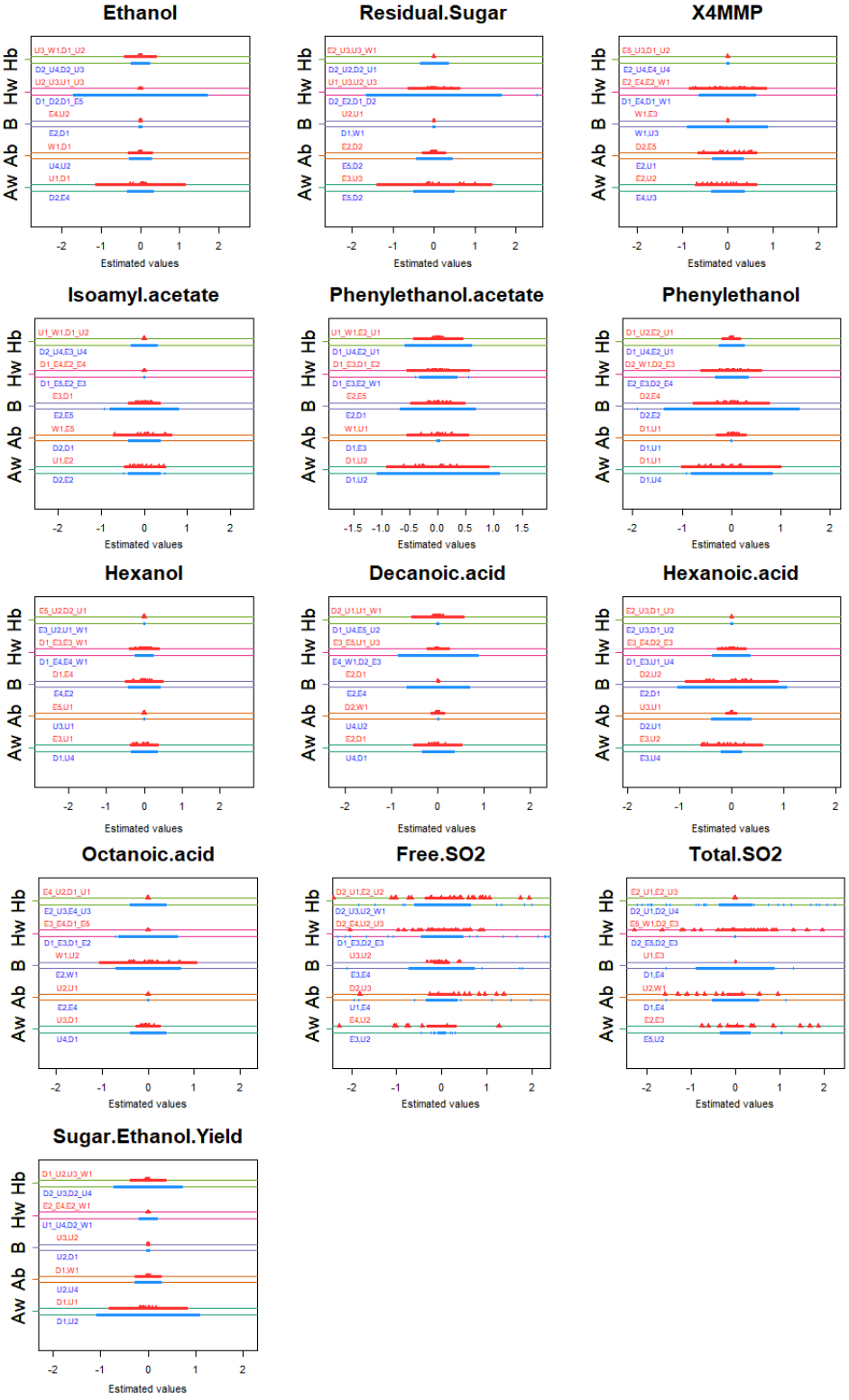


Figure S12: Interval plots. For each fermentation and life-history trait we plot the Best Linear Unbiased Predictors of the random genetic effects estimated through the decomposition of our diallel design. The random genetic effect estimates, namely $\hat{A}_w$, $\hat{A}_b$, $\hat{B}$, $\hat{H}_w$, $\hat{H}_b$ are plotted in blue (18°C), or in red (26°C). Horizontal bars are added to show, for each parameter, the region of highest density that covers nearly 95% ($\sim \pm 2\hat{\sigma}_{\boldsymbol{q}}$) of the parameter density. On the left hand-side of each plot we list, for each genetic effect, the strains which have the lowest and the greatest value of the respective genetic effect. The plot shows that: (i) genetic effects differ in a large extent between the two temperatures; (ii) additive and heterosis effects depend on the type of cross in which a line is involved (intra- or inter-specific); (iii) for some traits, genetic variances are strongly influenced by a particular hybrid combination.

Appendix B

Appendix B

Supplementary materials for "Data integration uncovers the metabolic bases of phenotypic variation in yeast"

B.1 Sampling the solution space

Let L denote the solution space of eq. 5 with constraints (eq. 6). Our aim is to sample random elements in the convex set L in order to characterize it by means of posterior joint distribution between fluxes. In order to do so, classical methods, like the well-known **Hit and Run** algorithm are available (Meersche *et al.* 2009). Braunstein *et al.* 2017 turn to map the original problem of sampling the feasible space of solutions L into the inference problem of the joint distribution of metabolic fluxes, letting the linear and inequality constrains to be encoded within the likelihood and the prior probabilities, which via the Bayes theorem provides a model for the flux posterior distribution density.

We compared the posterior density distribution obtained by the Hit and Run (HR) sampling to the Expectation Propagation algorithm (EP). We run the HR with a burning length equal to 10^6 and with a jump of 0.5, for a number of iteration from 10^6 to 10^7, and the EP algorithm with a high β parameter (Boltzmann inverse temperature parameter). Fig. SF1 shows the sampled space of solutions through the HR (histograms) and the EP estimate (red curve). Fig. SF2 shows the Pearson correlation coefficients between variances and means estimated through EP and HR for different number of iterations. As can be seen, the Pearson correlation increases as the number of the HR samples increases. Assuming that HR samples the true distribution of fluxes, means are well predicted by the EP algorithm, although variances are underestimated.

We further investigated if the EP algorithm well predicted the variance-covariance matrix of the DynamoYeast fluxes. In fig. SF3 are shown the relation between 8 pairwise fluxes chosen at random, and the correlation ellipses (red curve) computed by the EP algorithm. As can be seen, the EP algorithm well predicts the variance-covariance matrix between fluxes satisfying eq. 5, on the basis of the HR predictions.

B.2 Supplementary figures

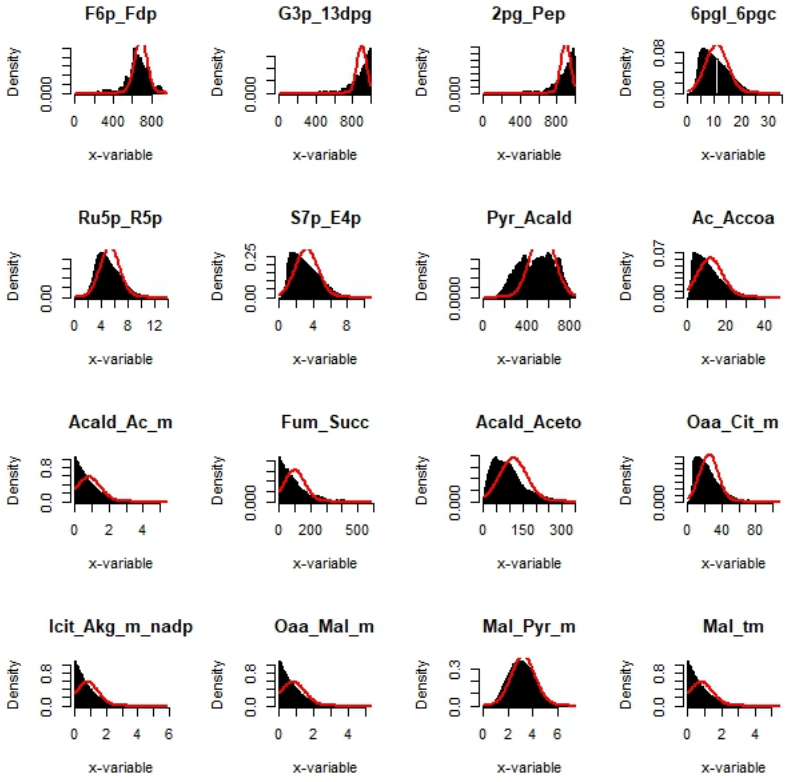


Figure SF1: Marginal probability densities of sixteen fluxes of the yeast carbon metabolism, randomly chosen. The histograms represent the result of the HR for $T \sim 10^7$ sampling points. The red line is the result of the EP estimate.

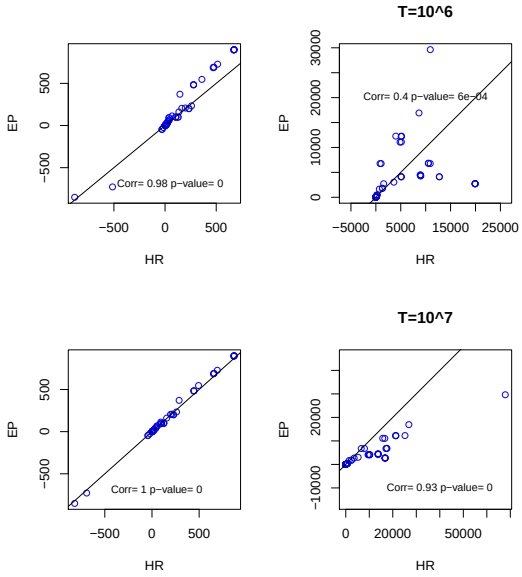


Figure SF2: Comparison of the results of HR versus EP. The plots on the left are scatter plots of the means and on the right variances of the approximated marginals computed via EP against the ones estimated via HR for an increasing number of explored configurations T, top $T \sim 10^6$, bottom $T \sim 10^7$.

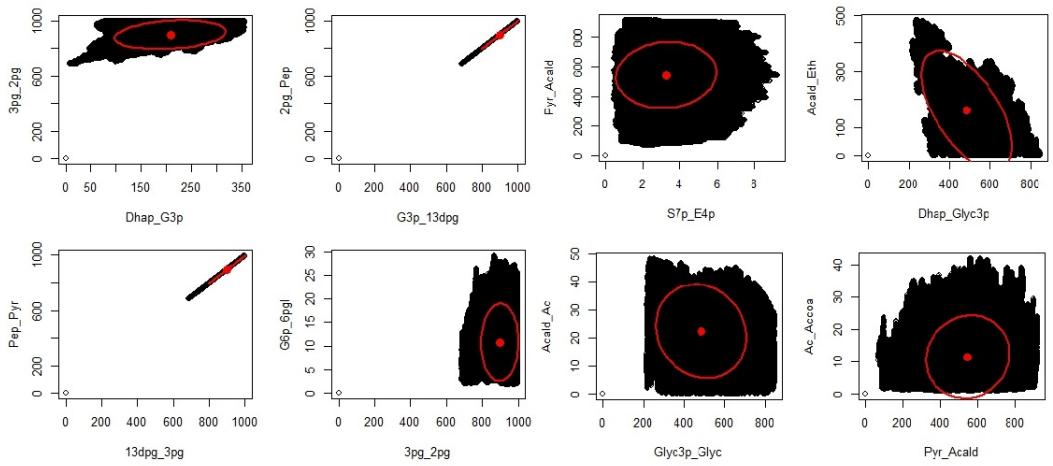


Figure SF3: Comparison of the results of HR versus EP. The plot shows the relation between 8 pairwise fluxes. Correlation ellipses, computed by the EP algorithm are drawn in red. Dot points represent the mean value of fluxes computed through EP. For HR samples, $T \sim 5 \cdot 10^6$.

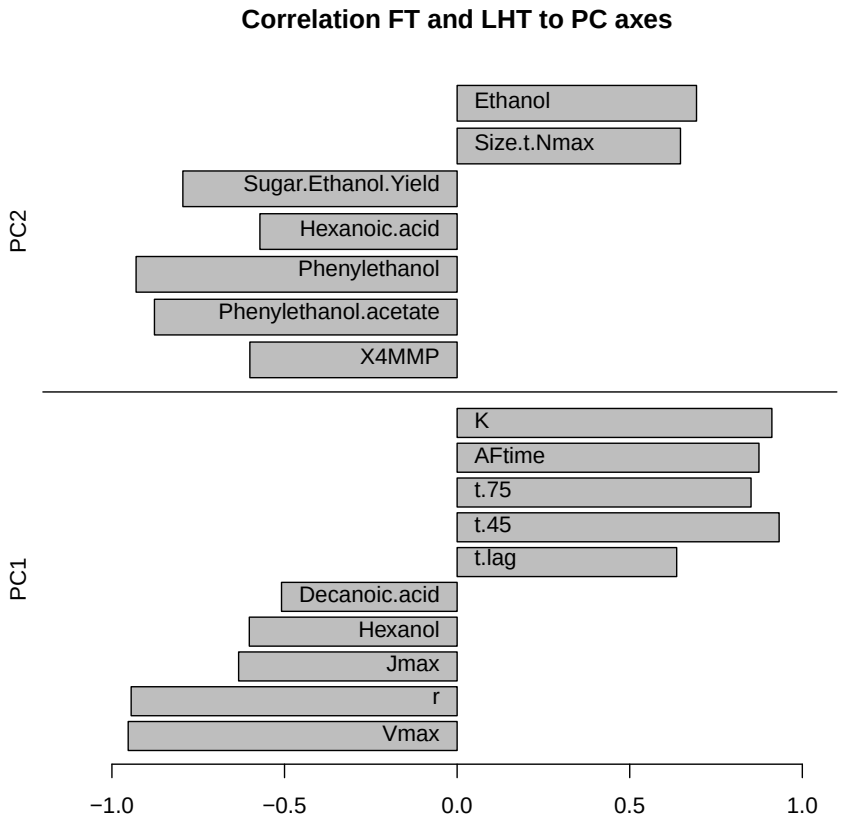


Figure SF4: Correlation between fermentation and life-history traits and the first two axis of the Principal Component Analysis. The figure shows traits for which the correlation was more more that 0.5 or less to -0.5 (p-value$<$0.05). The first axes is negatively correlated to the growth rate (r), CO_2 fluxes ($J_{\max}$ and $V_{\max}$), *Hexanol* and *Decanoic acid* and positively with the carrying capacity (K) and fermentation times (*AFtime*, *t-lag*, *t-75*, *t-45*). The second axes is positively correlated to cell-size (*Size-t-*$N_{\max}$) and *Ethanol* at the end of fermentation, while negatively with aroma production at the end of fermentation, as well as *Sugar.Ethanol.Yield*.

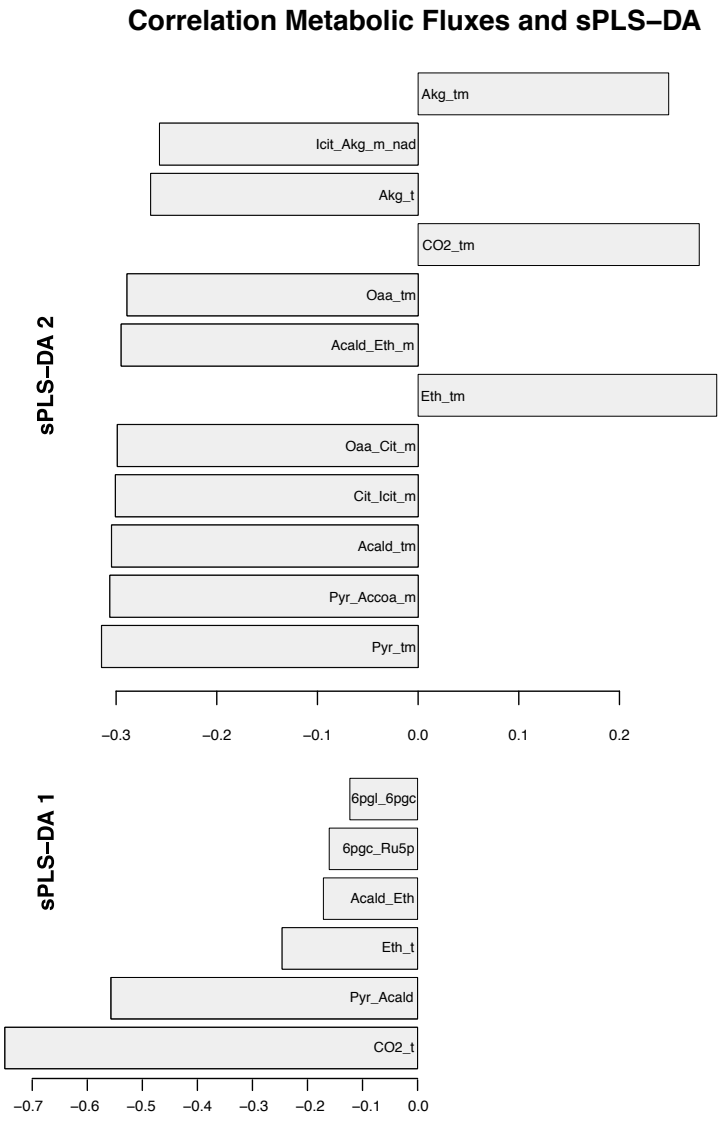


Figure SF5: Correlation between metabolic fluxes and the first two axis of the sparse Partial Least Square Discriminant Analysis. The CO_2, pyruvate decarboxylase, ethanol, alcohol dehydrogenase, 6-phosphogluconolactonase and phosphogluconate dehydrogenase fluxes contributed to the first axis of the sPLS-DA, and all were negatively correlated to it. The second axis was negatively correlated to the mitochondrial acetyl-CoA formation, mitochondrial citrate synthase, mitochondrial aconitate hydratase, mitochondrial isocitrate dehydrogenase (NAD+) and mitocondrial transport fluxes of pyruvate, oxaloacetate and acetaldehyde fluxes, while positively to mitochondrial transport of 2-oxodicarboylate, ethanol and CO_2 fluxes.

Appendix C

frontiers in Genetics

ORIGINAL RESEARCH
published: xx Month 2019
doi: 10.3389/fgene.2019.00833

Probabilities of Multilocus Genotypes in SIB Recombinant Inbred Lines

Kamel Jebreen[1,2†], *Marianyela Petrizzelli*[1†] *and Olivier C. Martin*[1*]

[1] *INRA, Univ. Paris-Sud, CNRS, AgroParishTech, Université Paris-Saclay, Gif-sur-Yvette, France,* [2] *Department of Mathematics, An-Najah National University, Nablus, Palestine*

Recombinant Inbred Lines (RILs) are obtained through successive generations of inbreeding. In 1931 Haldane and Waddington published a landmark paper where they provided the probabilities of achieving any combination of alleles in 2-way RILs for 2 and 3 loci. In the case of sibling RILs where sisters and brothers are crossed at each generation, there has been no progress in treating 4 or more loci, a limitation we overcome here without much increase in complexity. In the general situation of L loci, the task is to determine 2^L probabilities, but we find that it is necessary to first calculate the 4^L "identical by descent" (IBD) probabilities that a RIL inherits at each locus its DNA from one of the four originating chromosomes. We show that these 4^L probabilities satisfy a system of linear equations that follow from self-consistency. In the absence of genetic interference—crossovers arising independently—the associated matrix can be written explicitly in terms of the recombination rates between the different loci. We provide the matrices for L up to 4 and also include a computer program to automatically generate the matrices for higher values of L. Furthermore, our framework can be generalized to recombination rates that are different in female and male meiosis which allows us to show that the Haldane and Waddington 2-locus formula is valid in that more subtle case if the meiotic recombination rate is taken as the average rate across female and male. Once the 4^L IBD probabilities are determined, the 2^L probabilities of RIL genotypes are obtained *via* summations of these quantities. *In fine*, our computer program allows to determine the probabilities of all the multilocus genotypes produced in such sibling-based RILs for $L \geq 10$, a huge leap beyond the $L = 3$ restriction of Haldane and Waddington.

Keywords: Recombinant Inbred Line population, sibling mating, crossover, self-consistency, structure population

OPEN ACCESS

Edited by:
Chaeyoung Lee,
Soongsil University,
South Korea

Reviewed by:
Christophe Lambing,
University of Cambridge,
United Kingdom
Xuehui Huang,
Shanghai Normal University,
China

***Correspondence:**
Olivier C. Martin
Olivier.c.martin@inra.fr

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *21 June 2019*
Accepted: *13 August 2019*
Published: *xx Month 2019*

Citation:
Jebreen K, Petrizzelli M and Martin OC (2019) Probabilities of Multilocus Genotypes in SIB Recombinant Inbred Lines. Front. Genet. 10:833. doi: 10.3389/fgene.2019.00833

INTRODUCTION

There are numerous inference problems in population and quantitative genetics that require comparing experimental frequencies of genotypes to those expected "theoretically." Examples include genetic mapping of genomic markers, localizing causal factors of diseases and quantitative traits, performing marker assisted selection etc. (Lander and Schork, 1994; Weir, 1996; Walsh and Lynch, 2018). The *expected* frequencies of genotypes, hereafter referred to as probabilities, of interest in such studies often involve multiple loci (Buckler et al., 2009) and are strongly dependent on population structure. In population genetics studies, the structure of *natural* populations is rarely perfectly known. That partly explains why, in both animal and plant genetics, controlled crosses are widely produced to ensure a specific population structure. Arranging the crosses to lead to homozygous lines is greatly advantageous as such lines can be reproduced "identically and indefinitely." The simplest

situation satisfying these criteria is that of *recombinant inbred lines* (RILs) (Crow, 2007) founded from two parents as displayed in **Figure 1**. Given two (generally homozygous) parents that are the founders of the RIL construction (F_0), one first produces the associated hybrids (F_1). Second, starting with these F_1 individuals, one produces a sequence of generations F_2, F_3, etc by iterative inbreeding, crossing male and female siblings until formally at F_∞ one reaches full homozygozity (fixation of the alleles at all loci). As seen in **Figure 1**, the genomes of the homozygous lines produced by this process are mosaics of the parental genomes.

Consider the allelic content at some set of L genomic markers or loci. There are then 2^L possible RIL genotypes, each having a probability that depends on how meioses generate recombinations between these different loci. In the case of plants that allow for selfing, the same individual is both the mother and the father of its offspring; the RILs are then produced *via* single seed descent (SSD) as opposed to *via* sibling (hereafter denoted SIB) mating, this second case being the focus of the present work.

There are numerous generalizations of the RIL construction just given. Instead of using two parents to initiate the inbreeding, the use of 2^k parents leads to 2^k-way RILs (Broman, 2005). 2^k-way RILs start with 2^k parents to form 2^{k-1} offspring that are themselves crossed iteratively following a funnel (specifically a binary tree) pattern. Once the root of this tree is reached, the usual RIL inbreeding process is applied. For instance, the so called "Collaborative Cross" which has been a key community tool for mouse genetics, corresponds to $k = 3$; the choice there of using 8 founding parents at the top of the funnel allows for significantly greater allelic diversity than when using just 2-way RILs. Another generalization is the so called Advanced Intercross RIL (AI-RIL, sometimes referred to as Intermated RIL or IRIL) in which several generations of panmixia are inserted before applying the inbreeding to produce the RILs (Darvasi and Soller, 1995; Winkler et al., 2003; Rockman and Kruglyak, 2008). Other generalizations include Multi-parent Advanced Generation Inter-Cross (MAGIC) (El-din El-Assal et al., 2001), *nested association mapping* (NAM) populations (Buckler et al., 2009) etc. All of these population constructions involve some initial generations of allelic shuffling followed by the RIL (inbreeding) construction per se. Those *early* generations produce in effect initial conditions on the genotypes that are at the origin of the RILs and these initial conditions can be computed by direct recurrence from one generation to the next. In contrast, the RIL phase requires crossings that continue until all loci are homozygous and thus—at least mathematically—this phase involves an infinite number of generations. As a result, the computation of the probabilities of multilocus genotypes in RILs

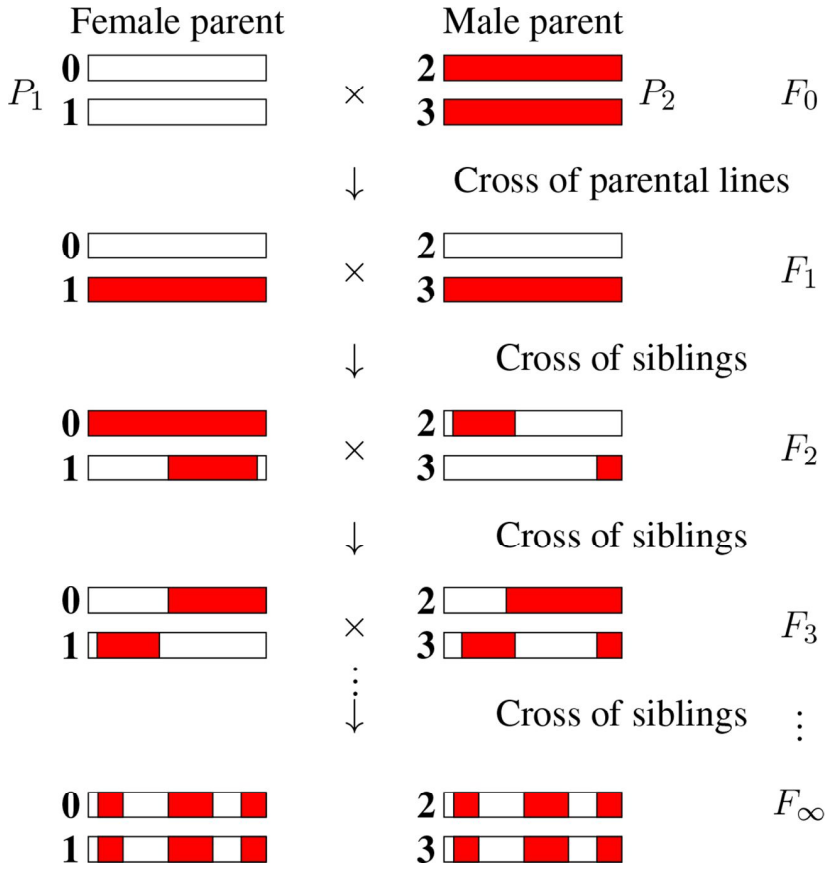


FIGURE 1 | The production of recombinant inbred lines by sibling mating (SIB RILs). The homologous chromosomes (numbered 0, 1, 2, and 3) inherit genomic segments from their parents, the boundaries of which identify crossover positions. The allelic content becomes fixed for "enough" generations, and so we introduce the limit of generations with the notation F_∞.

does not follow from a simple recursion over a fixed number of generations: either an extrapolation has to be made to deal with the infinite number of generations or some mathematical trick has to be devised to bypass the infinite nature of the process. This fact is at the heart of the difficulty of obtaining exact probabilities of multilocus genotypes in RILs.

The mathematical derivation of such RIL probabilities for two and three loci was provided by Haldane and Waddington (Haldene and Waddington, 1931) for bi-parental RILs in 1931. For two loci, by considering successive generations, they produced recursion equations for the probabilities of the corresponding (fixed or not) SIB genotypes which they then extrapolated to an infinite number of generations. This was quite a feat as they had to solve 22 simultaneous equations, leading *in fine* to their celebrated relation:

$$R = \frac{4r}{1+6r}. \quad (1)$$

In this formula, R is the probability for a RIL two-locus genotype to be recombinant (have the allele of one F_0 parent at one locus and the allele of the other at the other locus) while r is the recombination rate per meiosis between the two loci, assumed identical across male and female meiosis. We will rederive this formula using our framework in the section *Case of Two Loci: Recovering the Haldane–Waddington Result and Allowing for Sex-Dependent Recombination Rates* because to our knowledge, the generalization of the Haldane–Waddington formula to situations where male and female recombination rates differ has not been published and our framework allows to deal with this extension.

Given R, it is easy to derive the probabilities of the four different RIL genotypes (each of the two loci can be fixed for either of the two parental alleles). Indeed, the two recombinant genotypes have the same probability and the sum of these two probabilities is precisely R. The probability of each of the two recombinant (respectively non-recombinant) RIL genotypes is then R/2 (respectively (1-R)/2).

Haldane and Waddington further showed that this two-locus result also determined the three-locus probabilities. A way to see this is to notice that for three loci ($L = 3$) there are $2^L = 8$ different RIL genotypes (at each locus the homozygous allelic state comes from one of the two parents). These 8 genotypes can be grouped into 4 pairs such that within each pair one genotype is obtained from the other by exchanging the alleles of the parents; for instance if the alleles of the parents are denoted by (A, B, C) and (a, b, c) at the three successive loci, the 4 pairs are {(A, B, C), (a, b, c)}, and {(A, B, c), (a, b, C)}, {(A, b, C),(a, B, c)}, and {(a, B, C), (A, b, c)}. In each pair, the two complementary genotypes have the same probability so in effect it is enough to find the probabilities of each of the 4 pairs. These probabilities add up to one, providing a first equation. Then, labeling the loci as 1, 2, and 3, if the three meiotic recombination rates $r_{1,2}$, $r_{2,3}$ and $r_{1,3}$ are known, the three RIL recombination rates $R_{1,2}$, $R_{2,3}$ and $R_{1,3}$ are also. These quantities provide three further equations relating the four pair probabilities. These four equations uniquely determine the four pair probabilities and thus the probabilities of the 8 RIL genotypes.

Since that 1931 Haldane–Waddington landmark paper, some works have provided generalizations of Eq. 1, for instance in the case of 2^k-way RILs (Broman, 2005; Teuscher and Broman, 2007) and in the case of IRILs (Winkler et al., 2003; Teuscher and Broman, 2007). However, the problem of dealing with more than three loci seems substantially more difficult. Following the Haldane–Waddington algebraic approach, if there are L loci, there are 16^L possible allelic combinations at each generation and so it is necessary to diagonalize a $16^L \times 16^L$ matrix; that task takes on the order of 16^{3L} operations and thus cannot be done on a standard computer even for $L = 4$. To our knowledge, the only work providing closed-form expressions for 4 or more loci is that of (Samal and Martin, 2015), but their framework for determining exact probabilities of RIL multilocus genotypes applies only to single seed descent RILs, not to SIB RILs. The contribution of the present work is to show that the case of SIB RILs is also to a large extent tractable. In particular, (i) we give the analytic expressions for treating four loci in the absence of crossover interference, and (ii) we show that our framework allows to tackle more loci, though at a computational cost (CPU time and also computer memory) that increases roughly as 16^L. Specifically, our computer scripts, written in R (Ihaka and Gentleman, 1996), can treat $L = 8$ loci in approximately 5 min when run on a desktop computer while a high-end server allows us to go up to $L = 10$ loci. Lastly, to illustrate an application of our theoretical framework to a practical situation, we construct a maximum likelihood algorithm to impute missing data in RIL populations. In contrast to the standard approach which infers probabilities using machine learning, our method exploits the exact multilocus RIL genotype probabilities. By comparing the two approaches we show that the use of the exact probabilities significantly increases the reliability of the missing data imputation.

OVERVIEW OF THE METHOD

In the less complex case of single seed descent RILs, it was possible to determine the probabilities of the 2^L RIL multilocus genotypes by writing self-consistent equations directly associated with these unknowns (Samal and Martin, 2015). However, in the case of SIB RILs, the situation is more subtle because the allele carried by a RIL genotype may come from *either of the two siblings* at the F_1 generation and thus "identical by descent" (IBD) does not reduce to identity by state (having the same allelic content) as can be seen in **Figure 1.** As a result, it is necessary to first work with the 4^L probabilities that a RIL inherits IBD at the L loci from any of the four F_1 homologous chromosomes. After introducing in the Section *Probabilities of Multilocus IBD Inheritances in RILs and the Set of Non-Equivalent Q's* the 4^L RIL multilocus IBD probabilities, we show in the section *Self-Consistent Equations for the IBD Probabilities* that each of these unknowns satisfies a self-consistent equation relating it to the others. These equations allow to overcome the technical obstacle of there being an unlimited number of generations in the process of generating RILs all the way to complete fixation. Although these 4^L self-consistent equations constrain the 4^L unknowns, we show in the section *Adding One Linear Inhomogeneous Equation to Uniquely Specify All IBD Probabilities* that one additional equation is necessary to specify the solution. For that last constraint we use the fact that the sum of all probabilities is 1. In the section *Reducing the System of Equations to Treat Only the Non-Equivalent 's*, we show how the complexity of the problem can be reduced by

working with a subset only of the unknowns. Finally, upon solving the system of equations to determine the IBD quantities, each of the 2^L RIL multilocus genotype probabilities follows by summing the probabilities of all compatible IBDs as will be shown in the section *Extracting the Probabilities of RIL Genotypes*.

Probabilities of Multilocus IBD Inheritances in Rils and the Set of Non-Equivalent Q's

For a given RIL L-locus genotype (specified formally at generation F_∞), the genomic content at any locus ℓ ($\ell \in \{1,\ldots, L\}$) will be IBD with exactly one of the four F_1 homologous chromosomes. (One may note that the allelic fixation can happen before the IBD fixation, but no matter what, after an infinite number of generations both the IBD and the allelic states are fixed, that is they are identical across the four chromosomes of the SIB pair.) We number those four chromosomes 0, 1, 2 and 3 as indicated in **Figure 2** and use the same labeling for the later generations too. The IBD case illustrated is such that the RIL inherits from the F_1 chromosome 2 at the first locus and from the F_1 chromosome 1 at the second locus. (By convention we order the loci from left to right.) More generally, let us introduce the probability $Q(i_1, i_2,\ldots,i_L)$ that a RIL inherits IBD from F_1 chromosome i_ℓ for locus ℓ, $\ell = 1,\ldots, L$ where $i\ell = 0, 1, 2, 3$. Naturally the sum of these 4^L probabilities (there are four possible values of i_ℓ at each locus ℓ) is equal to 1.

For $L = 1$, there are four IBD probabilities: $Q(0)$, $Q(1)$, $Q(2)$ and $Q(3)$. We shall assume Mendelian segregation with no bias in favor of any particular allele and so in particular the two homologues within each sex are equivalent. Then $Q(i) = 1/4$ for all $\in \{0, 1, 2, 3\}$.

Moving on to $L = 2$ for which there are 16 Q's, the equivalence of homologues leads to the equalities $Q(0,0) = Q(1,1)$, $Q(0,1) = Q(1,0)$, $Q(2,2) = Q(3,3)$, and $Q(2,3) = Q(3,2)$ but also to equalities between mixed terms, $Q(0,2) = Q(0,3)$, $Q(1,2) = Q(1,3)$ etc. Furthermore, if female and male meiosis behave in the same way (so that in particular they have the same recombination rates), we can also conclude that $Q(0,0) = Q(2,2)$ etc so that finally there are just *three* probabilities to determine, $Q(0,0)$, $Q(0,1)$ and $Q(0,2)$ instead of the initial 16. More generally, if there are L loci, how many *non-equivalent* Q's are there? We shall assume there is no segregation bias and that female and male meioses have statistically identical behavior. Then it is possible to show (see **Supplementary Material** for details) that the number of non-equivalent Q's is exactly

$$N_Q(L) = 2^{L-2}\left(2^{L-1}+1\right). \quad (2)$$

For example $L = 1$ leads to $N_Q(L) = 1$ while $L = 2$ leads to $N_Q(L) = 3$. The number of these non-equivalent Q's grows roughly as $(1/8) \times 4^L$ to be compared with the total number ignoring equivalence of 4^L. The factor $(1/8)$ clearly makes it worthwhile to use such a reduction in the number of unknowns to simplify the task of writing and solving the equations. The proof of Eq. 2 in the **Supplementary Material** provides a way to enumerate the Q's to be kept and schematically goes as follows. First, because all four chromosomes play equivalent roles, we can force i_1 to be 0. Second, i_2 can be constrained not to take the value 3 since that value can be replaced by 2, this time by equivalence of chromosomes 2 and 3. If i_2 takes the value 0 or 1, we can again constrain i_3 to be different from 3 by the same reasoning. If instead $i_2 = 2$, then i_3 must be allowed to take all values 0, 1, 2 and

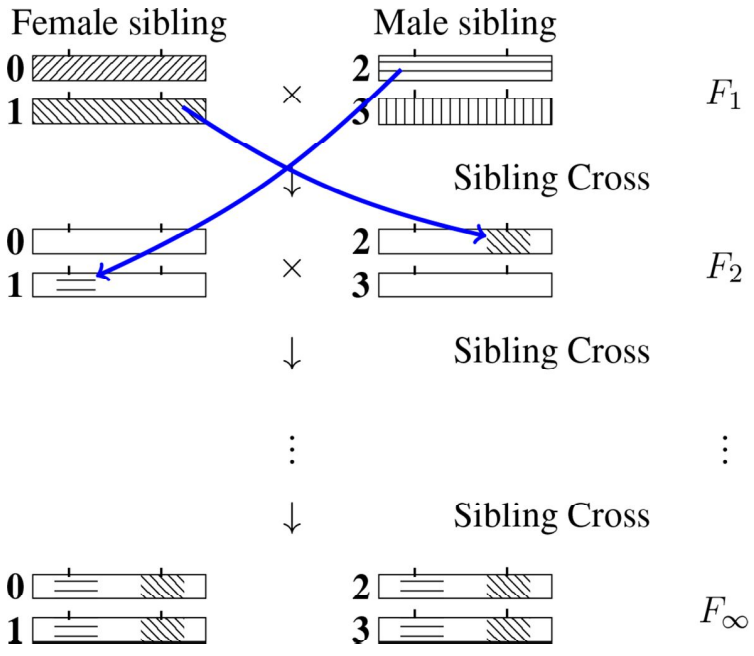


FIGURE 2 | Inheritance during SIB mating and illustration of the construction of a self-consistent equation for any IBD probability. At each generation the homologous chromosomes are labeled 0, 1 (for the female) and 2, 3 (for the male). Note that the chromosomes labeled 0 and 2 are the outcomes of female meiosis while the chromosomes labeled 1 and 3 are the outcomes of male meiosis. The drawing illustrates the transition probability $T[(2, 1) \rightarrow (1, 2)Q(1, 2)]$ entering the self-consistent equation (*cf.* Eq. 3 when the left-hand side is $Q(2, 1)$).

3. We can proceed in this way to define the rules to be applied to the successive i_p. As long as the current list consist of 0s and 1s, the next i can be constrained to not take the value 3 by equivalence between chromosomes 2 and 3, but for all entries after the *first* occurrence of a 2, all values must be allowed (see the **Supplementary Material** for the final steps required to prove Eq. 2). As an illustration, the reader can check that for $L = 3$ loci, this construction leads to 10 non-equivalent Q's, namely $Q(0,0,0)$, $Q(0,0,1)$, $Q(0,0,2)$, $Q(0,1,0)$, $Q(0,1,1)$, $Q(0,1,2)$, $Q(0,2,0)$, $Q(0,2,1)$, $Q(0,2,2)$, and $Q(0,2,3)$.

Self-Consistent Equations for the 4^L IBD Probabilities

The IBD inheritance needs an infinite number of generations to become fixed with certainty, at least in principle. Our strategy consist in mapping such an infinite process into a finite one by relying on self-consistency. The probability for F_∞ siblings to inherit IBD the sequence of "indices" $(i_1,i_2,\ldots,i_L)$ from the F_1 chromosomes can be decomposed into trajectories where the inheritance indices a t the F_2 level are also made explicit. If we denote these by $(i'_1,i'_2,\ldots,i'_L)$, we can reinterpret $Q(i_1,i_2,\ldots,i_L)$ as a sum of contributions:

$$Q\left(i_1,i_2,\ldots,i_L\right)=\sum_{\left(i'_1,i'_2,\ldots,i'_L\right)} T\left[\left(i_1,i_2,\ldots,i_L\right)\rightarrow\left(i'_1,i'_2,\ldots,i'_L\right)\right]Q\left(i'_1,i'_2\ldots,i'_L\right) \tag{3}$$

where $T[(\cdot)\rightarrow(\cdot)]$ is the transition probability of having the IBD propagate from the first list of indices to the second list of indices when going from the F_1 to the F_2 generation. $T[(\cdot)\rightarrow(\cdot)]$ is illustrated graphically in **Figure 2** by considering the case of two loci and having $i_1 = 2$, $i_2 = 1$, $i'_1 = 1$ and $i'_2 = 2$.

Clearly $T[(\cdot)\rightarrow(\cdot)]$ depends on the meiotic process, and thus in particular on the recombination rates between loci. To simplify the notation, let us set $u = (i_1,i_2,\ldots i_L)$ and $v = (i'_1,i'_2,\ldots,i'_L)$. These transition probabilities $T[(\cdot)\rightarrow(\cdot)]$ satisfy three properties. First, if $i_k = 0$ or 1, $T[u\rightarrow v] = 0$ unless $i'_k = 0$ or 2. Similarly, if $i_k = 2$ or 3, $T[u\rightarrow v] = 0$ unless $i'_k = 1$ or 3. We summarize this *via* the rules

$$i'\in\begin{cases}\{0,2\} & \text{if } i\in\{0,1\}\\ \{1,3\} & \text{if } i\in\{2,3\}\end{cases} \tag{4}$$

where i and $i'\in\{0,1,2,3\}$. Second, it turns out that the matrix T is "doubly stochastic" meaning that the sum of its entries in any row or in any column is exactly 1. The result that the sum over elements in a row is 1 follows from the fact that this sum gives the probability of having any of the possible outcomes of inheritances for a given starting point. Analogously, the result that the sum over all elements in a column is 1 corresponds to the fact that a given v is reached by some u and that summing over all possibilities for u again leads to 1. Third, each element of T decomposes into four factors,

$$T\left[u\rightarrow v\right]=P_0\left[u\rightarrow v\right]P_1\left[u\rightarrow v\right]P_2\left[u\rightarrow v\right]P_3\left[u\rightarrow v\right] \tag{5}$$

where the subscript of each P labels the chromosome of interest (and therefore the meiosis) at the F_2 generation, thus P_j is a probability associated with the meiosis that produces chromosome j when going from F_1 to F_2. Consider for specificity the term P_3. For the computation of this probability, only the entries in v equal to 3 matter. The corresponding *indices* specify which loci are thereafter IBD from chromosome 3 when considering the F_∞ inheritance from the F_2 generation. If those loci numbers are say 2, 5, and $L - 1$, then $P_3[u\rightarrow v]$ is the probability for the loci 2, 5 and $L - 1$ to inherit IBD from i_2, i_5 and i_{L-1} during the meiosis producing chromosome 3 when going from the F_1 generation to the F_2 generation. Note that all the other loci and chromosomes are irrelevant for this factor. The probability of that event is 0.5 (for the probability that the locus 2 will inherit IBD from chromosome i_2) times the probability that the successive intervals 2-5 and 5-$(L-1)$ will be as required – recombinant or not – by the values of i_5 and i_{L-1}. Let us suppose that meioses arise without genetic interference, that is, according to the so-called Haldane model (Haldane et al., 1919). (Note that the values of these Ps are the only part of our framework where crossover interference affects our computations; if these single-meiosis probabilities are known, then our framework provides the probabilities of all RIL multilocus genotypes just as in the case of no interference.) For specificity, if there is no interference and both intervals 2–5 and 5(– $(L - 1)$ are recombinant, the associated (meiotic) probability P is simply $0.5\times r_{2,5}\times r_{5,L-1}$. Such a reasoning is easily extended to any situation, leading to the formula

$$P_j\left[u\rightarrow v\right]=0.5\prod_{\langle l,l'\rangle} r_{l,l'}^{e_{l,l'}}\left(1-r_{l,l'}\right)^{1-e_{l,l'}} \tag{6}$$

where the locus indices l and l' are such that $v_l = v_{l'} = j$, j being the index appearing in the probability P_j. In addition, the $e_{l,l'}$ are defined as

$$e_{l,l'}=\begin{cases}1 & \text{if the interval is "recombinant"}\\ 0 & \text{if the interval is not "recombinant".}\end{cases} \tag{7}$$

For Eq. 6, an interval $\langle l,l'\rangle$ is called "recombinant" if and only if i_l and $i_{l'}$ differ. Lastly, we need to specify the actual pairs of loci l and l' that are to be used in that equation. To do so, we first construct the list of ordered indices that satisfy the constraint $v_l = v_{l'} = j$. The product in Eq. 6 is then over the *successive* pairs of this list. If the list is empty, $P_j = 1$ while if there is only one element in the list, $P_j = 0.5$. The interpretation of Eq. 6 is then as follows: there is a factor $r_{l,l'}$ if the u list imposes that the interval $\langle l,l'\rangle$ be recombinant and a factor $1 - r_{l,l'}$ otherwise. Putting together Eqs.3 and 5 specifies the 4^L linear homogeneous equations for the Q's. In our computer software, we determine the matrix elements of T as formal mathematical functions of the $r_{l,l'}$. In these general expressions it is possible to substitute the numerical values of the $r_{l,l'}$ when necessary.

Adding One Linear Inhomogeneous Equation to Uniquely Specify All 4^L IBD Probabilities

Eq. 3 can be rewritten as

$$Q(u)-\sum_v T\left[u\rightarrow v\right]Q(v)=0 \tag{8}$$

for all choices of u, corresponding to a set of 4^L linear homogeneous equations. Given one has as many equations as unknowns, one might hope that this system would determine the Q's but that is not the case because these 4^L equations are not independent. Indeed, consider the sum of all the equations in the system:

$$\sum_u Q(u) - \sum_u \sum_v T[u \to v] Q(v) = 0. \tag{9}$$

By interchanging the order of the sums this becomes

$$\sum_u Q(u) - \sum_v \left(\sum_u T[u \to v] \right) Q(v) = 0 \tag{10}$$

which is automatically satisfied because T is doubly stochastic so that $\sum_u T[u \to v] = 1$. To overcome the problem coming from this dependence amongst the homogeneous self-consistent equations, we need to include further information. We choose to do that by adding the constraint that the sum of all 4^L IBD probabilities equals 1:

$$\sum_u Q(u) = 1. \tag{11}$$

The inclusion of this (inhomogeneous) linear equation then uniquely specifies the values of all Q's.

Reducing the System of Equations to Treat Only the $N_Q(L)$ Non-Equivalent Q'S

As mentioned previously, it is advantageous to work with a subset of non-equivalent Q's because this substantially reduces the complexity of the operations to be performed. Specifically, we modify the above approach by considering self-consistent equations only for the reduced list of unknowns—the $N_Q(L)$ non-equivalent Q's chosen in the section *Probabilities of Multilocus IBD Inheritances in RILs and the Set of Non-Equivalent Q's*—so instead of having 4^L homogeneous equations of the type Eq. 8 we have only $N_Q(L)$ of them. In these $N_Q(L)$ equations, we replace each $Q(v)$ by an equivalent $Q(v')$ where $Q(v')$ belongs to our list of $N_Q(L)$ unknowns. This recipe leads to $N_Q(L)$ linear homogeneous equations for our unknowns. Furthermore, we also apply these substitutions to the inhomogeneous equation Eq. 11, with the previously mentioned rule. As a result, by counting the number of Q's arising in each equivalence class defined in the section *Probabilities of Multilocus IBD Inheritances in RILs and the Set of Non-Equivalent Q's*, $Q(u)$ occurs with weight 4 if the entries of u are all different from 2 and with weight 8 otherwise.

In practice, to solve this set of equations, it is convenient to have as many equations as unknowns so we remove exactly one of the homogeneous equations. In our computer algorithm we remove the last of these homogeneous equations but any other choice is just as valid. Having obtained as many independent equations as there are unknowns, the direct solution of this linear system (a linear algebra problem) provides the (unique) values of our N_Q(L) non-equivalent Q's

Extracting the 2^L Probabilities of RIL Genotypes

Once the Q's are determined, the probabilities of RIL multilocus genotypes can be computed by summing all IBD probabilities that are *compatible* with the RIL allelic content. Let us refer to the allelic content of parent 1 as a series of A alleles and that of parent 2 as a series of a alleles. Consider then a RIL multilocus genotype, written as a list $G = (\alpha_1, \alpha_2, \ldots, \alpha_L)$ of L alleles, α_k being A or a. The probability of a genotype G is obtained by summing over all $Q(u)$ for which the u is compatible with the allelic content of G. The compatibility rule can be summarized as follows: if $\alpha_k = A$, then u_k must be 0 or 2, while if $\alpha_k = a$, then u_k must be 1 or 3. This is formalized mathematically by the following equation

$$P(G = (\alpha_1, \alpha_2, \ldots, \alpha_L)) = \sum_u Q(u) \tag{12}$$

where the sum is restricted to the u's satisfying the compatibility rule. Note that the Q's on the right-hand side of Eq. 12 in general will not belong to our list of non-equivalent Q's. As before, just omit all the terms associated with Q's that are not in this list and multiply the other terms by either 8 or 4 depending on whether the associated u has one of its indices u_k equal to 2 or not, again because of the size of the equivalence classes.

RESULTS

We illustrate the power of our framework by considering increasing number of loci. The case of two loci is presented both for pedagogical reasons and to give the novel (as far as we know) values of the IBD probabilities when allowing for sex-dependent recombination rates. For three loci we detail the derivation of the coefficients of the self-consistent equations by giving associated graphical representations in the **Supplementary Material**. For four loci the analytical expression of the 40 × 40 matrix is also given explicitly. For more loci, the mathematical steps become too cumbersome to be dealt with by hand, but our computer code (in the form of R functions) can be used to first generate the analytic expressions for the linear system of equations, then to solve that system for the Q's, and finally to produce the probabilities of all the RIL multilocus genotypes. The complexity of the computations provided by our framework can be summarized *via* the dimensionality of the linear system of equations used to compute the Q's. This dimension increases roughly by a factor 4 for each additional locus for the simple reason that the number of unknowns increases in that way (*cf.* Eq. 2). Lastly, in the section *Application to Imputing Missing Data* we will apply our method to the problem of imputing missing values in RIL genotyping data, demonstrating the benefit of using exact multilocus genotypes.

Case of Two Loci: Recovering the Haldane–Waddington Result and Allowing for Sex-Dependent Recombination Rates

Haldane and Waddington (1931) derived the formula for the probabilities of 2-locus RIL genotypes and (Teuscher and Broman, 2007) gave an alternative more compact approach. We will derive that Haldane–Waddington result here using our self-consistency approach. Then we show how to extend our framework to the case where female and male recombination rates differ.

Let $r_{l,l'} = r_{1,2}$ denote the recombination fraction between the two loci (this recombination rate is for the moment taken to be the same in female and male as assumed by Haldane and Waddington). Furthermore, let a^l denote the allele at locus l, $l \in \{1,\ldots,L\}$, on any of the homologous chromosomes in the RIL. By Eq. 2, for $L = 2$ there are 3 unknown Q's. The indices u for each of these Q's are such that they are not related by the symmetry between chromosomes. Our choice is to use $Q(0,0)$, $Q(0,1)$ and $Q(0,2)$. To build the 3×3 system of equations, begin with the inhomogeneous linear equation

$$4Q(0,0)+4Q(0,1)+8Q(0,2)=1. \tag{13}$$

where the respective factors 8 and 4 follow from whether or not the $\boldsymbol{u}$ list of indices contains a 2. The next step is to write the self-consistent equation for each of the $N_Q(L) - 1$ non-equivalent Q's. For instance for $u = (0,0)$, by Eq. 3 applied to this case and using the rules for the vanishing of the elements of the matrix T, one has

$$\begin{aligned} Q(0,0) \quad &= T\left[(0,0)\to(0,0)\right]Q(0,0) \\ &+T\left[(0,0)\to(0,2)\right]Q(0,2) \\ &+T\left[(0,0)\to(2,0)\right]Q(2,0) \\ &+T\left[(0,0)\to(2,2)\right]Q(2,2). \end{aligned} \tag{14}$$

The matrix elements $T[u \to v]$ are determined by Eqs. 5 and 6. Direct calculation gives $(1 - r_{1,2})/2$, 1/4, 1/4, and $(1 - r_{1,2})/2$ respectively. To obtain a self-consistent equation involving only our three non-equivalent Q's, we rewrite Eq. 14 by replacing $Q(2,0)$ by $Q(0,2)$ and $Q(2,2)$ by $Q(0,0)$, leading to

$$r_{1,2}Q(0,0)-Q(0,2)/2=0. \tag{15}$$

The self-consistent equation for $Q(0,1)$ is obtained by the same method. Eq. 13 together with Eq. 15 and its analogue for $Q(0,1)$ then lead to the system

$$\begin{bmatrix} 4 & 4 & 8 \\ r_{1,2} & 0 & -\frac{1}{2} \\ r_{1,2} & -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} Q(0,0) \\ Q(0,1) \\ Q(0,2) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \tag{16}$$

(Compared to Eq. 8, we have changed the signs of each homogeneous equation to obtain a more readable matrix.) This system can be solved by hand, leading to

$$Q(0,0)=\frac{1}{4+24r_{1,2}},\quad Q(0,1)=Q(0,2)=\frac{r_{1,2}}{2+12r_{1,2}}. \tag{17}$$

Given these three values, we can compute the RIL recombination rate R by summing all the probabilities of IBD events that produce recombinant RILs:

$$\begin{aligned} R = {} & Q(0,1)+Q(0,3)+Q(2,1)+Q(2,3) \\ & +Q(1,0)+Q(1,2)+Q(3,0)+Q(3,2). \end{aligned} \tag{18}$$

Using the equivalences ($Q(3,0) = Q(0,2)$ etc), this gives $R = 4Q(0,1) + 4Q(0,2)$; substituting the values from Eq. 17 leads directly to the Haldane-Waddington formula, Eq. 1.

How do these results extend to the case where female and male have different recombination rates, r^f and r^m? The main complication comes from the fact that the symmetries of the system are reduced: one can no longer exchange the roles of female and male SIBs. As a result, there are 6 non-equivalent IBD probabilities. Without loss of generality, we take these to be $Q(0,0)$, $Q(0,1)$, $Q(0,2)$, $Q(2,0)$, $Q(2,2)$, and $Q(2,3)$. The determination of these six unknowns follows the same logic as when $r^f = r^m$. First, use the inhomogeneous equation specifying that the Q's are probabilities that add up to 1:

$$2Q(0,0)+2Q(0,1)+4Q(0,2)+4Q(2,0)+2Q(2,2)+2Q(2,3)=1. \tag{19}$$

Second, determine the homogeneous equations associated with the self-consistency for the first $N_Q(L) - 1$ non-equivalent Q's. This then leads to the following system of equations:

$$\begin{bmatrix} 2 & 2 & 4 & 4 & 2 & 2 \\ \left(\frac{1}{2}\bar{r}^f-1\right) & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2}\bar{r}^f & 0 \\ \frac{1}{2}r^f & -1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2}r^f & 0 \\ 0 & \frac{1}{4} & \frac{3}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{2}\bar{r}^m & 0 & \frac{1}{4} & \frac{1}{4} & \left(\frac{1}{2}\bar{r}^m-1\right) & 0 \\ \frac{1}{2}r^m & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2}r^m & 1 \end{bmatrix} \begin{bmatrix} Q_{(0,0)} \\ Q_{(0,1)} \\ Q_{(0,2)} \\ Q_{(2,0)} \\ Q_{2,2} \\ Q_{(2,3)} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{20}$$

For the matrix elements in this system of equations, we have used the notation $\bar{r} = 1 - r$ to designate the complementary value of the recombination rate, such a notation allowing for more

compact expressions. The linear system Eq. 20 can be solved by hand, leading to

$$Q(0,0)=\frac{-\frac{1}{2}r^f+\frac{1}{2}r^m+1}{4(3r^f+3r^m+1)} \qquad Q(0,1)=\frac{3r^f+r^m}{8(3r^f+3r^m+1)}$$

$$Q(0,2)=Q(2,0)=\frac{r^f+r^m}{4(3r^f+3r^m+1)} \qquad Q(2,2)=\frac{\frac{r^f}{2}-\frac{r^m}{2}+1}{4(3r^f+3r^m+1)}$$

$$Q(2,3)=\frac{r^f+3r^m}{8(3r^f+3r^m+1)}.$$

(21)

Note that except for $Q(0, 2)$ and $Q(2, 0)$, all the Q's are *asymmetric* functions of r^f and r^m. Furthermore, the equality $Q(0, 2) = Q(2, 0)$ follows from the special symmetry of replacing the left-right convention that orients chromosomes by one using the right-left orientation.

Given the non-trivial result of Eq. 21, we can ask what is the consequence for R, the RIL recombination rate. The calculation is straightforward:

$$\begin{aligned}R &= Q(0,1)+Q(0,3)+Q(1,0)+Q(1,2)\\ &\quad +Q(2,1)+Q(2,3)+Q(3,0)+Q(3,2)\\ &= 2(Q(0,1)+Q(0,2)+Q(2,0)+Q(2,3))\\ &= \frac{2(r^f+r^m)}{3(r^f+r^m)+1}.\end{aligned}$$

(22)

Interestingly, this result depends only on the mean of the female and male recombination rates, in spite of the fact that such a property does not hold at the level of the individual Q's. Furthermore, it shows that the Haldane-Waddington relation (Eq. 1) can be used when recombination rates are sex-dependent if in that formula the (sex-independent) recombination rate is replaced by the sex-averaged recombination rate.

Although this example was very simple (it involved only two loci), it should be clear that our framework is generally applicable, for any number of loci, whether the female and male recombination rates are identical or not.

Case of Three Loci

Haldane and Waddington showed that the probabilities of two-locus RIL genotypes may be used to derive the probabilities of the three-locus RIL genotypes. Teuscher and Broman also provided this result when they introduced their approach (Broman, 2005; Teuscher and Broman, 2007). In the introduction we explained why such a relation holds and so one might expect a similar conclusion to hold for the Q's, but this is not so. Indeed, for this $L = 3$ case, as mentioned in the section *Probabilities of Multilocus IBD Inheritances in RILs and the Set of Non-Equivalent Q's*, there are $N_Q(L) = 10$ unknown Q's to determine, corresponding to 9 degrees of freedom, but the information from the $L = 2$ level only provides 6 constraints, two for each pair of loci ($6 = 2 \times 3$).

To determine the values of all the IBD probabilities, we simply apply our framework when using $L = 3$. We begin by specifying the set of non-equivalent Q's that are our unknowns, following the logic of the general case as exposed in the section *Probabilities of Multilocus IBD Inheritances in RILs and the Set of Non-Equivalent Q's*. We thus choose $Q(0, 0, 0)$, $Q(0, 0, 1)$, $Q(0, 0, 2)$, $Q(0, 1, 0)$, $Q(0, 1, 1)$, $Q(0, 1, 2)$, $Q(0, 2, 0)$, $Q(0, 2, 1)$, $Q(0, 2, 2)$, and $Q(0, 2, 3)$. Second, we write the single inhomogeneous equation that sums all Q's (before applying equivalences). Third, we construct the self-consistent equations for the first 9 of our non-equivalent Q's, assuming no genetic interference. The **Supplementary Material** provides a graphical representation of the T $[u \to v]$ entries to be explicit, our R code constructs this matrix automatically. These successive steps lead to the following linear system for our 10 unknowns:

$$\begin{bmatrix}
4 & 4 & 8 & 4 & 4 & 8 & 8 & 8 & 8 & 8\\
\bar{r}_{12}\bar{r}_{23}-1 & 0 & \frac{\bar{r}_{12}}{2} & 0 & 0 & 0 & \frac{\bar{r}_{13}}{2} & 0 & \frac{\bar{r}_{23}}{2} & 0\\
r_{23}\bar{r}_{12} & -1 & \frac{\bar{r}_{12}}{2} & 0 & 0 & 0 & \frac{r_{13}}{2} & 0 & \frac{r_{23}}{2} & 0\\
0 & \frac{\bar{r}_{12}}{2} & \frac{\bar{r}_{12}-2}{2} & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4}\\
r_{12}r_{23} & 0 & \frac{r_{12}}{2} & -1 & 0 & 0 & \frac{\bar{r}_{13}}{2} & 0 & \frac{r_{23}}{2} & 0\\
r_{12}\bar{r}_{23} & 0 & \frac{r_{12}}{2} & 0 & -1 & 0 & \frac{r_{13}}{2} & 0 & \frac{\bar{r}_{23}}{2} & 0\\
0 & \frac{r_{12}}{2} & \frac{r_{12}}{2} & 0 & 0 & -1 & 0 & \frac{1}{4} & 0 & \frac{1}{4}\\
0 & 0 & 0 & \frac{\bar{r}_{13}}{2} & 0 & \frac{1}{4} & \frac{\bar{r}_{13}-2}{2} & 0 & 0 & \frac{1}{4}\\
0 & 0 & 0 & \frac{r_{13}}{2} & 0 & \frac{1}{4} & \frac{r_{13}}{2} & -1 & 0 & \frac{1}{4}\\
0 & 0 & 0 & 0 & \frac{\bar{r}_{23}}{2} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{\bar{r}_{23}-2}{2} & 0
\end{bmatrix}
\begin{bmatrix}
Q(0,0,0)\\ Q(0,0,1)\\ Q(0,0,2)\\ Q(0,1,0)\\ Q(0,1,1)\\ Q(0,1,2)\\ Q(0,2,0)\\ Q(0,2,1)\\ Q(0,2,2)\\ Q(0,2,3)
\end{bmatrix}
=
\begin{bmatrix}
1\\0\\0\\0\\0\\0\\0\\0\\0\\0
\end{bmatrix}$$

(23)

where as before $\bar{r} = 1 - r$ denotes the complementary value of the recombination rate and $r_{ij} = r_{i,j}$. The solution of Eq. 23 can be obtained either numerically or analytically—that is as an explicit function of the three recombination rates—using e.g., Maple or Mathematica since a treatment by hand would be very tedious.

Four and More Loci

The previous methodology can be extended to more loci but quickly becomes too cumbersome to manage manually. For illustration, in the case $L = 4$, there are 40 Q's to determine (*cf.* Eq. 2). The system of 40 linear inhomogeneous equations determining these unknowns is given in Eq. 24 and barely fits on one page as a figure.

In that display including a 40×40 matrix, we have used the same compact notation as for $L = 3$. Our software produces this system of equations and then can solve for the Q's for any particular values of the r_{ij}. Computing the corresponding probabilities of RIL genotypes is then straightforward and in practice the computer does this very quickly.

It is of course possible to go to larger values of L but then it becomes unweildly to show the corresponding matrix. As expected, the computation time required by our R code grows

Q17

$$
\left[\begin{smallmatrix}
4 & 4 & 8 & 4 & 4 & 8 & 8 & 8 & 8 & 8 & 4 & 4 & 8 & 4 & 4 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 & 8 \\
\bar{r}_{12}\bar{r}_{23}\bar{r}_{34}-1 & 0 & \frac{\bar{r}_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & \frac{\bar{r}_{12}\bar{r}_{24}}{2} & 0 & \frac{\bar{r}_{12}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}\bar{r}_{34}}{2} & 0 & \frac{\bar{r}_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}\bar{r}_{14}}{2} & 0 & \frac{\bar{r}_{23}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
r_{34}\bar{r}_{12}\bar{r}_{23} & -1 & \frac{\bar{r}_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & \frac{\bar{r}_{12}r_{24}}{2} & 0 & \frac{\bar{r}_{12}r_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{13}}{2} & 0 & \frac{r_{24}\bar{r}_{13}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}\bar{r}_{23}}{2} & 0 & \frac{r_{34}\bar{r}_{23}}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{\bar{r}_{12}\bar{r}_{23}}{2} & \frac{\bar{r}_{12}\bar{r}_{23}}{2}-1 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{12}}{4} & 0 & \frac{\bar{r}_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}}{4} & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}}{4} & 0 & \frac{\bar{r}_{23}}{4} & 0 & 0 & 0 & 0 \\
r_{23}r_{34}\bar{r}_{12} & 0 & \frac{r_{23}\bar{r}_{12}}{2} & -1 & 0 & 0 & \frac{\bar{r}_{12}\bar{r}_{24}}{2} & 0 & \frac{r_{34}\bar{r}_{12}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}r_{34}}{2} & 0 & \frac{r_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}\bar{r}_{14}}{2} & 0 & \frac{r_{23}r_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
r_{23}\bar{r}_{12}\bar{r}_{34} & 0 & \frac{r_{23}\bar{r}_{12}}{2} & 0 & -1 & 0 & \frac{r_{24}\bar{r}_{12}}{2} & 0 & \frac{\bar{r}_{12}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}\bar{r}_{34}}{2} & 0 & \frac{r_{13}r_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}r_{14}}{2} & 0 & \frac{r_{23}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{r_{23}\bar{r}_{12}}{2} & \frac{r_{23}\bar{r}_{12}}{2} & 0 & 0 & -1 & 0 & \frac{\bar{r}_{12}}{4} & 0 & \frac{\bar{r}_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}}{4} & 0 & \frac{r_{13}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}}{4} & 0 & \frac{r_{23}}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{\bar{r}_{12}\bar{r}_{24}}{2} & 0 & \frac{\bar{r}_{12}}{4} & \frac{\bar{r}_{12}\bar{r}_{24}}{2}-1 & 0 & 0 & \frac{\bar{r}_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{\bar{r}_{24}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{\bar{r}_{24}}{4} & 0 \\
0 & 0 & 0 & \frac{r_{24}\bar{r}_{12}}{2} & 0 & \frac{\bar{r}_{12}}{4} & \frac{r_{24}\bar{r}_{12}}{2} & -1 & 0 & \frac{\bar{r}_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{r_{24}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{r_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & \frac{\bar{r}_{12}\bar{r}_{34}}{2} & \frac{\bar{r}_{12}}{4} & 0 & \frac{\bar{r}_{12}}{4} & \frac{\bar{r}_{12}\bar{r}_{34}}{2}-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{34}}{4} & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & \frac{\bar{r}_{34}}{4} \\
0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{12}}{2} & \frac{\bar{r}_{12}}{4} & 0 & \frac{\bar{r}_{12}}{4} & \frac{r_{34}\bar{r}_{12}}{2} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}}{4} & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & \frac{r_{34}}{4} \\
r_{23}r_{12}\bar{r}_{34} & 0 & \frac{r_{12}r_{23}}{2} & 0 & 0 & 0 & \frac{r_{12}r_{24}}{2} & 0 & \frac{r_{12}\bar{r}_{34}}{2} & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}\bar{r}_{34}}{2} & 0 & \frac{r_{24}\bar{r}_{13}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}\bar{r}_{14}}{2} & 0 & \frac{r_{23}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
r_{12}r_{23}r_{34} & 0 & \frac{r_{12}r_{23}}{2} & 0 & 0 & 0 & \frac{r_{12}\bar{r}_{24}}{2} & 0 & \frac{r_{12}r_{34}}{2} & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{13}}{2} & 0 & \frac{\bar{r}_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}r_{14}}{2} & 0 & \frac{r_{23}r_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{r_{12}r_{23}}{2} & \frac{r_{12}r_{23}}{2} & 0 & 0 & 0 & 0 & \frac{r_{12}}{4} & 0 & \frac{r_{12}}{4} & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}}{4} & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}}{4} & 0 & \frac{r_{23}}{4} & 0 & 0 & 0 & 0 \\
r_{12}\bar{r}_{23}r_{34} & 0 & \frac{r_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & \frac{r_{12}r_{24}}{2} & 0 & \frac{r_{12}r_{34}}{2} & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}r_{34}}{2} & 0 & \frac{r_{13}r_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}\bar{r}_{14}}{2} & 0 & \frac{r_{34}\bar{r}_{23}}{2} & 0 & 0 & 0 & 0 & 0 \\
r_{12}\bar{r}_{23}\bar{r}_{34} & 0 & \frac{r_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & \frac{r_{12}\bar{r}_{24}}{2} & 0 & \frac{r_{12}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}\bar{r}_{34}}{2} & 0 & \frac{r_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}\bar{r}_{23}}{2} & 0 & \frac{\bar{r}_{23}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{r_{12}\bar{r}_{23}}{2} & \frac{r_{12}\bar{r}_{23}}{2} & 0 & 0 & 0 & 0 & \frac{r_{12}}{4} & 0 & \frac{r_{12}}{4} & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}}{4} & 0 & \frac{r_{13}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}}{4} & 0 & \frac{\bar{r}_{23}}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{r_{12}r_{24}}{2} & 0 & \frac{r_{12}}{4} & \frac{r_{12}r_{24}}{2} & 0 & 0 & \frac{r_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{r_{24}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{r_{24}}{4} & 0 \\
0 & 0 & 0 & \frac{r_{12}\bar{r}_{24}}{2} & 0 & \frac{r_{12}}{4} & \frac{r_{12}\bar{r}_{24}}{2} & 0 & 0 & \frac{r_{12}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{\bar{r}_{24}}{4} & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{\bar{r}_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & \frac{r_{12}\bar{r}_{34}}{2} & \frac{r_{12}}{4} & 0 & \frac{r_{12}}{4} & \frac{r_{12}\bar{r}_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{34}}{4} & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & \frac{\bar{r}_{34}}{4} \\
0 & 0 & 0 & 0 & \frac{r_{12}r_{34}}{2} & \frac{r_{12}}{4} & 0 & \frac{r_{12}}{4} & \frac{r_{12}r_{34}}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}}{4} & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & 0 & \frac{r_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}\bar{r}_{34}}{2} & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{\bar{r}_{34}}{4} & 0 & \frac{\bar{r}_{13}\bar{r}_{34}}{2}-1 & 0 & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{\bar{r}_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{13}}{2} & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{r_{34}}{4} & 0 & \frac{r_{34}\bar{r}_{13}}{2} & -1 & 0 & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & 0 & \frac{r_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{13}\bar{r}_{24}}{2} & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & \frac{\bar{r}_{24}}{4} & 0 & \frac{1}{8} & 0 & \frac{\bar{r}_{13}}{4} & \frac{\bar{r}_{13}\bar{r}_{24}}{2}-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{\bar{r}_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{24}\bar{r}_{13}}{2} & \frac{\bar{r}_{13}}{4} & 0 & 0 & 0 & 0 & \frac{r_{24}}{4} & 0 & \frac{1}{8} & 0 & \frac{\bar{r}_{13}}{4} & \frac{r_{24}\bar{r}_{13}}{2} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{r_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}r_{34}}{2} & 0 & \frac{r_{13}}{4} & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & \frac{r_{34}}{4} & 0 & \frac{r_{13}r_{34}}{2} & 0 & 0 & \frac{r_{13}}{4} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{r_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}\bar{r}_{34}}{2} & 0 & \frac{r_{13}}{4} & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & \frac{\bar{r}_{34}}{4} & 0 & \frac{r_{13}\bar{r}_{34}}{2} & 0 & 0 & \frac{r_{13}}{4} & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & 0 & \frac{\bar{r}_{34}}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}\bar{r}_{24}}{2} & \frac{r_{13}}{4} & 0 & 0 & 0 & 0 & \frac{\bar{r}_{24}}{4} & 0 & \frac{1}{8} & 0 & \frac{r_{13}}{4} & \frac{r_{13}\bar{r}_{24}}{2} & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{\bar{r}_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{13}r_{24}}{2} & \frac{r_{13}}{4} & 0 & 0 & 0 & 0 & \frac{r_{24}}{4} & 0 & \frac{1}{8} & 0 & \frac{r_{13}}{4} & \frac{r_{13}r_{24}}{2} & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{r_{24}}{4} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}\bar{r}_{14}}{2} & 0 & \frac{\bar{r}_{23}}{4} & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{1}{8} & \frac{\bar{r}_{23}\bar{r}_{14}}{2}-1 & 0 & 0 & \frac{\bar{r}_{23}}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{14}\bar{r}_{23}}{2} & 0 & \frac{\bar{r}_{23}}{4} & \frac{r_{14}}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & 0 & \frac{1}{8} & \frac{r_{14}\bar{r}_{23}}{2} & -1 & 0 & \frac{\bar{r}_{23}}{4} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{23}\bar{r}_{34}}{2} & \frac{\bar{r}_{23}}{4} & 0 & \frac{\bar{r}_{24}}{4} & \frac{\bar{r}_{34}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\bar{r}_{34}}{4} & \frac{\bar{r}_{24}}{4} & 0 & 0 & \frac{\bar{r}_{23}}{4} & \frac{\bar{r}_{23}\bar{r}_{34}}{2}-1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}\bar{r}_{23}}{2} & \frac{\bar{r}_{23}}{4} & 0 & \frac{r_{24}}{4} & \frac{r_{34}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}}{4} & \frac{r_{24}}{4} & 0 & 0 & \frac{\bar{r}_{23}}{4} & \frac{r_{34}\bar{r}_{23}}{2} & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}\bar{r}_{14}}{2} & 0 & \frac{r_{23}}{4} & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{\bar{r}_{14}}{4} & 0 & 0 & \frac{1}{8} & \frac{r_{23}\bar{r}_{14}}{2} & 0 & 0 & \frac{r_{23}}{4} & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}r_{14}}{2} & 0 & \frac{r_{23}}{4} & \frac{r_{14}}{4} & 0 & 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & \frac{r_{14}}{4} & 0 & 0 & \frac{1}{8} & \frac{r_{23}r_{14}}{2} & 0 & 0 & \frac{r_{23}}{4} & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{23}r_{34}}{2} & \frac{r_{23}}{4} & 0 & \frac{\bar{r}_{24}}{4} & \frac{r_{34}}{4} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{r_{34}}{4} & \frac{\bar{r}_{24}}{4} & 0 & 0 & \frac{r_{23}}{4} & \frac{r_{23}r_{34}}{2} & 0 & 0 & 0 & -1 & 0
\end{smallmatrix}\right]
\begin{bmatrix}
Q(0,0,0,0) \\ Q(0,0,0,1) \\ Q(0,0,0,2) \\ Q(0,0,1,0) \\ Q(0,0,1,1) \\ Q(0,0,1,2) \\ Q(0,0,2,0) \\ Q(0,0,2,1) \\ Q(0,0,2,2) \\ Q(0,0,2,3) \\ Q(0,1,0,0) \\ Q(0,1,0,1) \\ Q(0,1,0,2) \\ Q(0,1,1,0) \\ Q(0,1,1,1) \\ Q(0,1,1,2) \\ Q(0,1,2,0) \\ Q(0,1,2,1) \\ Q(0,1,2,2) \\ Q(0,1,2,3) \\ Q(0,2,0,0) \\ Q(0,2,0,1) \\ Q(0,2,0,2) \\ Q(0,2,0,3) \\ Q(0,2,1,0) \\ Q(0,2,1,1) \\ Q(0,2,1,2) \\ Q(0,2,1,3) \\ Q(0,2,2,0) \\ Q(0,2,2,1) \\ Q(0,2,2,2) \\ Q(0,2,2,3) \\ Q(0,2,3,0) \\ Q(0,2,3,1) \\ Q(0,2,3,2) \\ Q(0,2,3,3)
\end{bmatrix}
=
\begin{bmatrix}
1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}
\qquad (24)
$$

fast with L, by about a factor 16 for each unit increase of L. The required computer memory also grows in the same way. At L = 8 the code takes about 5 min to solve the problem, and for still larger values of L it is best to use a server with large memory capacity (we have gone up to L = 10).

Application to Imputing Missing Data

Genotyping arrays can provide calls for thousands and even millions of single nucleotide polymorphisms. When dealing with such large numbers, the raw data of *some* markers will inevitably be unambiguous and so generally these cases are called as "missing data." On the other hand, some technologies such as genotyping by sequencing of low coverage in fact lead *predominantly* to missing data calls. To deal with either of these cases, one typically imputes *a posteriori* to transform the missing calls into the most plausible values, exploiting the values of the calls at neighboring loci. Such imputation is a general problem and is typically treated by machine learning approaches that attempt to infer probabilities from the data. For the current context where we are focused on SIB RIL populations, one may expect that having the exact probabilities for RIL multilocus genotypes will allow for more reliable imputation than when using algorithms which resort to statistical inference.

To test this idea, we have developed an algorithm that exploits our exact probabilities and compared it to a standard imputation algorithm. The comparison is based on applying these two algorithms to simulated RIL genotyping data, and doing so for many replicates. Specifically, for each replicate, we started with two homozygous parents having two homologous chromosomes of total length 150 cm on which we randomly positioned 100 markers. After producing from these a SIB RIL population of 100 individuals, we took the genotypes of each individual and transformed the calls by introducing missing data, selecting at random 10, 20, 30, 40, 50 or 70% of the markers for this change from a parental allele to "missing data."

For a standard imputation algorithm, we used the R package "missForest" (Stekhoven and Buelhmann, 2012; Stekhoven, 2013) that uses machine learning to estimate the most probable values underlying the missing data. In effect, it applies a hidden Markov model that adjusts its parameters to the dataset. It outputs the imputed genotypes from which one can determine an associated error rate. This error rate depends on the realization of the RIL population, which is why we perform replicates. These values are displayed on the X axis of **Figure 3** for each of the studied values of the percentage of missing data. One clearly sees that the error rate increases with that percentage, a feature that is of course expected.

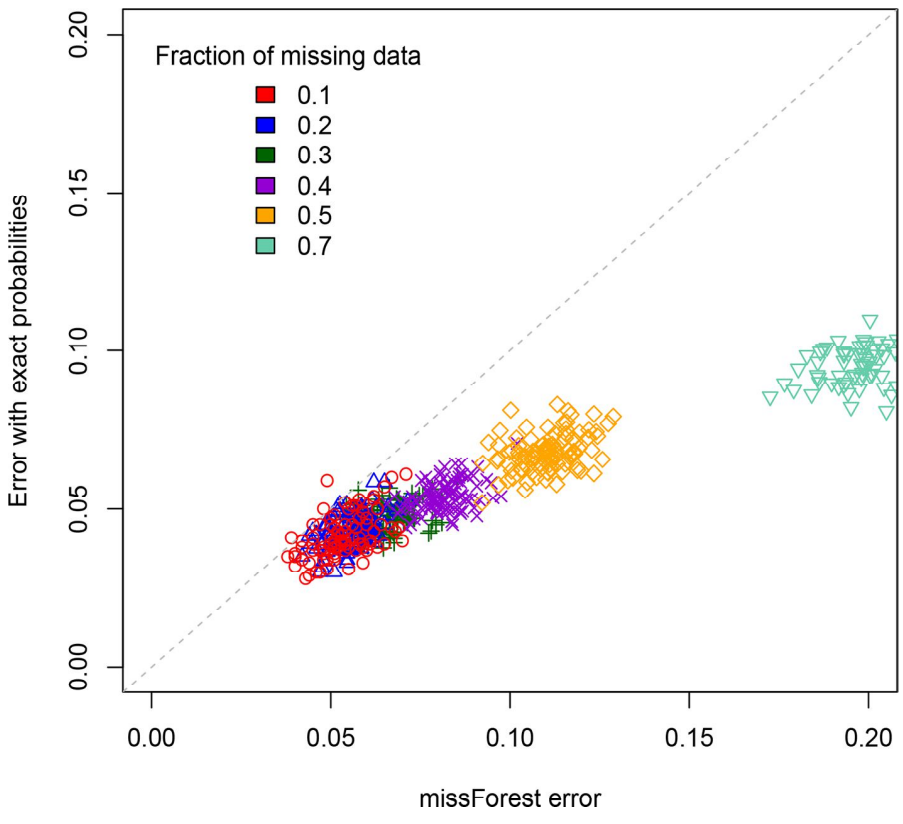


FIGURE 3 | Scatterplot comparison of imputation error rates. For each of the replicates simulated, the X axis gives the error rate for the missForest R package while the Y axis gives the error rate when using our code that exploits the *exact* multilocus genotype probabilities. For each fraction of missing data studied we display the points using a different color.

The second method (ours) is based on exploiting the fact that one can now access (cf. previous sections) the *exact* multilocus genotype probabilities in SIB RILs. For each genotyped RIL, we construct the *blocks* of adjacent markers that are called as missing data. They thus have one or two flanking markers. If there is just one such marker (the block is at the end of the chromosome), we impute the values to be that of this flanking marker. If there are instead two flanking markers (one on each side), again we just impute all the missing data to be that of these two markers if they are both of the same parental type (non-recombinant). The remaining case is where the considered individual is recombinant for those two flanking markers. To impute here, we calculate the multilocus genotype probabilities for all 2^L allelic combinations when considering these two flanking markers and the $L - 2$ markers in the missing data block. Then we select the L-locus genotype of maximum probability that is compatible with the calls at the two flanking markers, and this selected genotype specifies our imputation. The corresponding imputation errors are displayed on the Y axis of **Figure 3**. Clearly, our imputation method systematically out performs missForest, as expected since using the exact probabilities should be more reliable than using approximate ones. This is further quantified in the **Supplementary Material**. The imputation code is available online at https://github.com/olivier-c-martin/PMG_SIB_RILs.git and is internal documentation explains in greater detail the different algorithmic steps.

DISCUSSION

The construction of RILs involves successive generations of inbreeding. In realistic situations, SIB based inbreedings are performed for 10 to 20 generations and that leads to some low level of residual heterozygosity. One way to deal with such residual heterozygosity is to follow the probabilities of *all possible combinations* of allelic values for the siblings from one generation to the next. As shown by Haldane and Waddington (Haldane and Waddington, 1931), that means applying at each generation a $16^L \times 16^L$ matrix to the vector of those probabilities, where L is the number of loci considered (see also Hospital et al., 1996). Because this is very tedious and just not possible for 5 or more loci, a short cut is used whereby one considers that the statistics are given by the limiting case in which fixation should be complete and whenever a locus is in the heterozygous state one replaces it by missing data. Softwares that construct genetic maps or that perform QTL mapping then either just ignore such missing data or first perform imputation on those missing values.

That brings us to the challenge of determining RIL probabilities when fixation is indeed complete; note that there are far fewer combinations of allelic values in this situation than when one allows for residual heterozygosity, so one might hope for a simple way to obtain the corresponding multilocus genotype probabilities. But in this mathematical idealization where fixation is complete, the difficulty is that fixation formally requires an infinite number of generations. Thus, either the recursions must be taken "sufficiently far" to obtain *numerical* convergence or a mathematical trick has to be found. For $L = 2$, Haldane and Waddington succeeded in the second path thanks to much mathematical ingenuity, and interestingly, that $L = 2$ solution automatically determines the probabilities in the $L = 3$ case. However, since that founding work—going back to 1931—no solution had been proposed to tackle the problem of determining probabilities in SIB RILs with four or more loci.

Using a novel method, we have successfully overcome that long-standing challenge here. Our approach provides an algebraic solution, albeit at a computational cost that grows roughly as 16^L for L loci. That exponential growth rate is far less drastic than that of the iterative method mentioned above using $16^L \times 16^L$ matrices and even more dramatically less than applying diagonalization methods as in of the original proposition of Haldane and Waddington of 1931. As a result, not only did we break the $L = 4$ barrier but in fact we were able to rather easily treat L's up to 8. We also pointed out that our framework can deal with different female and male recombination rates, a situation that seems to have never been considered before in the context of SIB RILs, even for $L = 2$.

The ability to compute probabilities of RIL multilocus genotypes opens up to a number of applications. For instance, when building genetic maps, the ordering of markers is determined by comparing likelihoods of different orderings. That calculation can now be done using exact rather than approximate multilocus genotype frequencies, putting those mapping algorithms on a more solid footing. Similarly, when RIL genotypes must be imputed because of missing data, determining the most likely value of an allele marked as missing data requires comparing multilocus genotype probabilities. In the absence of these probabilities, imputation algorithms use approximations. We showed that it was in fact possible to avoid doing so, leading to systematically more reliable imputation results. Finally, beyond specific uses in the case of RILs, our mathematical framework that exploits self-consistency might be useful in certain population genetics problems involving an infinite number of generations.

SOFTWARE AVAILABILITY

R code implementing the methodology described in this paper as well as the study of the imputation application is available online at https://github.com/olivier-c-martin/PMG_SIB_RILs.git

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript/**Supplementary Files.**

AUTHOR CONTRIBUTIONS

OM proposed the project and with MP conceived and implemented a first approach. KJ introduced the analytic

formulation and this led to major enhancements to the algorithmic KJ and MP developed the R scripts and all authors wrote, edited, and approved the manuscript.

FUNDING

This work has benefited from a French State grant (LabEx Saclay Plant Sciences-SPS, ANR-10-LABX-0040-SPS), managed by the French National Research Agency under an "Investments for the Future" program (ANR-11-IDEX-0003-02) which funded the salary of KJ. Also, the public Ph.D. grant from the French National Research Agency (ANR) as part of the Investissement d'Avenir program, through the Initiative Doctoral Interdisciplinaire (IDI) 2015 project funded by the Initiative d'Excellence (IDEX) Paris-Saclay, ANR-11-IDEX-0003-02 funded the salary of MP.

REFERENCES

Broman, K. W. (2005). The genomes of recombinant inbred lines. *Genetics* 169 (2), 1133–1146. doi: 10.1534/genetics.104.035212

Buckler, Edward S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., et al. (2009). The genetic architecture of maize flowering time. *Science* 325 (5941), 714–718. doi: 10.1126/science.1174276

Crow, J. F. (2007). Haldane, Bailey, Taylor and recombinant-inbred lines. *Genetics* 176 (2), 729–732.

Darvasi, A., and Soller, M. (1995). Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* 141 (3), 1199–1207.

El-Din El-Assal, S., Alonso-Blanco, C., Peeters, A. J., Raz, V., and Koornneef, M. (2001). A QTL for flowering time in *Arabidopsis* reveals a novel allele of CRY2. *Nat. Genet.* 29 (4), 435–440. doi: 10.1038/ng767

Haldane, J. B. S., and Waddington, C. H. (1931). Inbreeding and linkage. *Genetics* 16 (4), 357–374.

Haldane, J. S., Meakins, J. C., and Priestley, J. G. (1919). The effects of shallow breathing. *J. Physiol.* 52 (6), 433–453. doi: 10.1113/jphysiol.1919.sp001842

Hospital, F., Dillmann, C., and Melchinger, A. E. (1996). A general algorithm to compute multilocus genotype frequencies under various mating systems. *Comput. Appl. Biosci.: CABIOS* 12 (6), 455–462. doi:10.1093/bioinformatics/12.6.455

Ihaka, R., and Gentleman, R. (1996). R: A Language for data analysis and graphics. *J. Comput. Graph. Stat.* 5 (3), 299–314. doi: 10.1080/10618600.1996.10474713

Lander, E. S., and Schork, N. J. (1994). Genetic dissection of complex traits. *Sci. (New York, N.Y.)* 265 (5181), 2037–2048. doi: 10.1126/science.8091226

Rockman, M. V., and Kruglyak, L. (2008). Breeding designs for recombinant inbred advanced intercross lines. *Genetics* 179 (2), 1069–1078. doi: 10.1534/genetics.107.083873

ACKNOWLEDGMENTS

The authors are grateful to Prof. D. de Vienne and C. Dillmann for insightful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00833/full#supplementary-material

The supplementary material contains three parts: a mathematical proof of Eq. 2 from the main text, some additional information for our imputation algorithm, and the graphical representations of the self-consistent equations for the $1L = 3$ case. The Supplementary Material for this article can be found online at: https://github.com/olivier-c-martin/PMG_SIB_RILs.git

Samal, A., and Martin, O. C. (2015). Statistical physics methods provide the exact solution to a long-standing problem of genetics. *Phys. Rev. Lett.* 114 (23), 238101. doi: 10.1103/PhysRevLett.114.238101

Stekhoven, D. J. (2013). *missForest: Nonparametric missing value imputation using random forest.* R package version 1.4.

Stekhoven, D. J., and Buehlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28 (1), 112–118. doi: 10.1093/bioinformatics/btr597

Teuscher, F., and Broman, K. W. (2007). Haplotype probabilities for multiple-strain recombinant inbred lines. *Genetics* 175 (3), 1267–1274. doi: 10.1534/genetics.106.064063

Walsh, B., and Lynch, M. (2018). *Evolution and Selection of Quantitative Traits.* Sunderland, Massachusetts, USA: Oxford University Press. doi: 10.1093/oso/9780198830870.001.0001

Weir, B. S. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data.* 2 sub edition edn. Sunderland, Mass: Sinauer Associates is an imprint of Oxford UniversityPress.

Winkler, C. R., Jensen, N. M., Cooper, Mark, P., Dean, W., and Smith, O. S. (2003). On the determination of recombination rates in intermated recombinant inbred populations. *Genetics* 164 (2), 741–745.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers

Supplementary Material

1 SUPPLEMENTARY MATHEMATICS

Here we prove the equation

$$N_Q(L) = 2^{L-2}(2^{L-1} + 1) \tag{S1}$$

Proof: If L is the number of loci, there are 4^L IBD (identical by descent) probabilities $Q(i_1, i_2, \ldots i_L)$ where $i_l = 0, 1, 2$ or 3 and furthermore these probabilities add up to 1. A number of these probabilities are equal because of two symmetries: (1) the two homologous chromosomes in each individual play identical roles, and (2) the siblings play identical roles (assuming no sex-dependence of meiosis, so that the recombination rates $r_{l,l'}$ are sex-independent). It is thus appropriate to use only one representative of each equivalence class generated by these symmetries. A way to do this is to first impose that this representative have its first index, i_1, equal to zero. Second, we can then specify exactly one element in each class by imposing that the indices of the representative Q's have either

1. $i_l \in \{0, 1\}\ \forall l \in \{2, .., L\}$, *or*
2. $i_l \in \{0, 1\}\ \forall l \in \{2, .., K-1\}$, $i_K = 2$ and $i_l \in \{0, 1, 2, 3\}\ \forall l \in \{K+1, .., L\}$

The number of equivalence classes and thus of Q's to consider is then

$$N_Q(L) = 2^{L-1} + \sum_{l=2}^{L} 2^{l-2}4^{L-l} = 2^{L-1} + 2^{2L-2}\sum_{l=2}^{L} 2^{-l} \tag{S2}$$

Given that $\sum_{l=2}^{L} 2^{-l}$ is a geometric progression of common ratio 2^{-1} from 2 to L, the sum of its terms can be expressed as:

$$\sum_{l=2}^{L} 2^{-l} = \frac{2^{-2} - 2^{-(L-1)}}{1 - 2^{-1}} = 2^{-1} - 2^{-L} \tag{S3}$$

Substituting S3 in S2, we get

$$N_Q(L) = 2^{L-1} + 2^{2L-2}(2^{-1} - 2^{-L}) = 2^{L-1} + 2^{2L-3} - 2^{L-2} \tag{S4}$$

Factorizing with respect to 2^{L-2} and after simplification, this gives

$$N_Q(L) = 2^{L-2}(1 + 2^{L-1}). \tag{S5}$$

2 IMPUTING USING THE EXACT RIL PROBABILITIES

We compared the performance of the missForest R package to our new approach that exploits the exact multilocus genotype probabilities. As mentioned in the Main, our method is based on focusing on missing data forming blocks of consecutive markers. When the block is large (this happens stochastically), it may be impossible in practice (for time and memory) to compute the needed multilocus genotype probabilities. To overcome this difficulty, we have implemented a "divide and conquer" method whereby inside the block we first focus on a subset of just 3 of those markers. After imputation is done on these 3, imputation requiring calculating multilocus probabilities involving 5 loci because of the flanking markers, we proceed to consider the remaining markers with missing data; these are now organized into one or more blocks of smaller size. The divide and conquer process can thus be repeated iteratively until there are no more markers to impute. A choice has to be made in the "divide" step for selecting the 3 most relevant markers. We do that by a bottom-up greedy approach where markers are successively removed, one step at a time. At each step, we first find the 2 markers that are closest (in this test we include the flanking markers and distances are in cM); if only one marker has missing data, we remove it; if both have missing data, we remove the one which is closest to its other adjacent marker.

For each value of the fraction of missing data (0.1, 0.2, 0.3, 0.4, 0.5 and 0.7), and for each replicate of a SIB RIL population (cf. the scatter plot of the Main), we determined the fraction of missing data that were incorrectly imputed in each method. Based on these replicates, Fig. S1 provides the box plots for each level of missing data studied. Clearly, the distributions of values hardly overlap, allowing us to conclude that using the exact multilocus RIL probabilities leads to a big improvement.

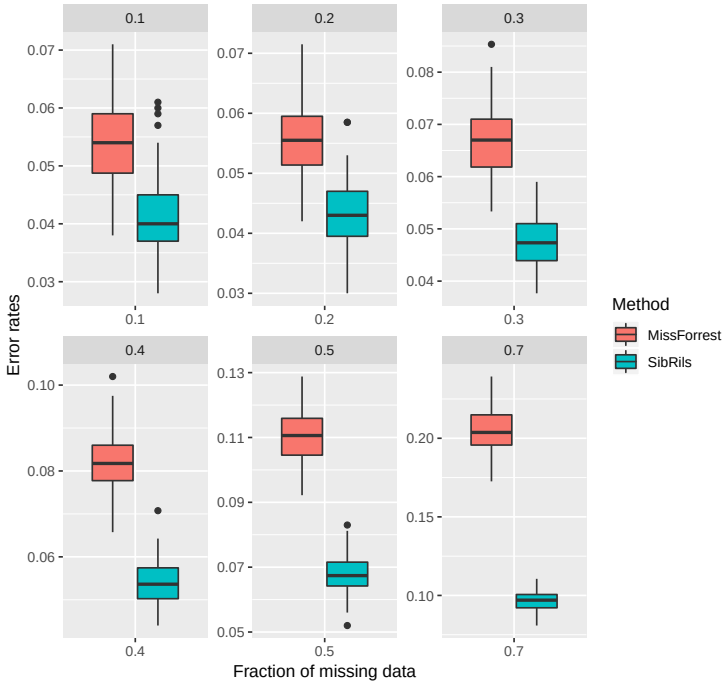


Figure S1: Box plots to compare imputation error rates between the missForest machine learning algorithm and our approach using the *exact* values of the multilocus genotype probabilities. The fraction of missing data applied to the datasets are given at the top of each plot. For almost all cases, there is hardly any overlap between the distributions of the two algorithms, the exact approach is systematically better.

3 THE SELF-CONSISTENT EQUATIONS FOR THREE LOCI

Here we provide the coefficients entering each of the $N_Q(L) = 10$ self-consistent equations for $L = 3$.

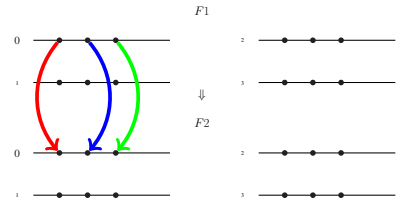
3.1 The self consistent equation for $Q(0, 0, 0)$

Figure S2 displays the 8 factors in the self-consistent equation for $Q(0, 0, 0)$:

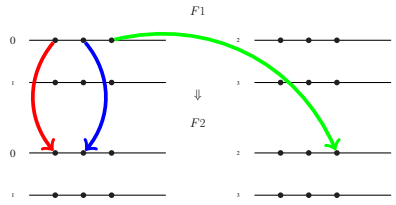
$$Q(0,0,0) = \tfrac{1}{2}(1-r_{12})(1-r_{23})[Q(0,0,0)+Q(2,2,2)] + \tfrac{1}{4}(1-r_{12})[Q(0,0,2)+Q(2,2,0)] + \tfrac{1}{4}(1-r_{13})[Q(0,2,0)+Q(2,0,2)] + \tfrac{1}{4}(1-r_{23})[Q(0,2,2)+Q(2,0,0)] \quad \text{(S6)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

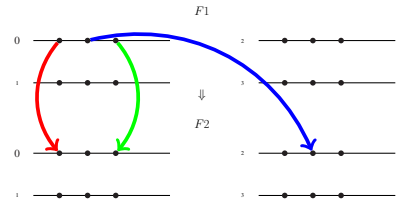
$$Q(0,0,0) = (1-r_{12})(1-r_{23})Q(0,0,0) + \tfrac{1}{2}(1-r_{12})Q(0,0,2) + \tfrac{1}{2}(1-r_{13})Q(0,2,0) + \tfrac{1}{2}(1-r_{23})Q(0,2,2)$$



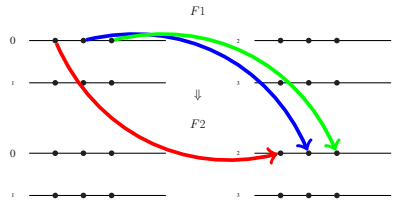
(a) $Q(0,0,0) : \frac{1}{2} \times (1-r_{12}) \times (1-r_{23}) \times Q(0,0,0)$



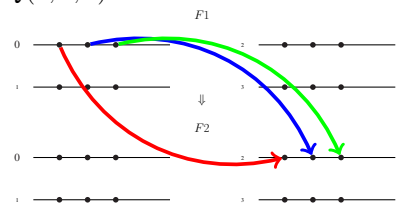
(b) $Q(0,0,0) : \frac{1}{2} \times (1-r_{12}) \times \frac{1}{2} \times Q(0,0,2)$



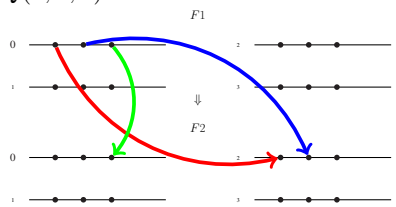
(c) $Q(0,0,0) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{13}) \times Q(0,2,0)$



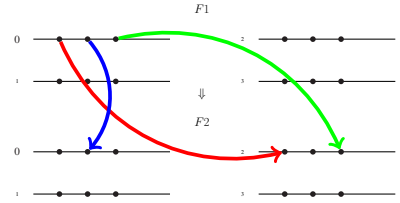
(d) $Q(0,0,0) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{23}) \times Q(0,2,2)$



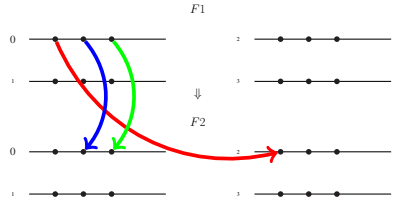
(e) $Q(0,0,0) : \frac{1}{2} \times (1-r_{12})(1-r_{23}) \times Q(2,2,2)$



(f) $Q(0,0,0) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{12}) \times Q(2,2,0)$



(g) $Q(0,0,0) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{13}) \times Q(2,0,2)$



(h) $Q(0,0,0) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{23}) \times Q(2,0,0)$

Figure S2: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S6 for $Q(0, 0, 0)$.

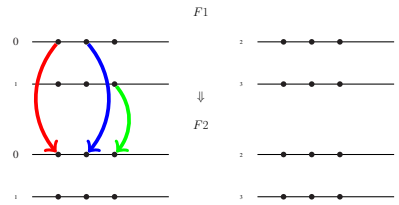
3.2 The self consistent equation for $Q(0,0,1)$

Figure S3 displays the 8 factors in the self-consistent equation for $Q(0,0,1)$:

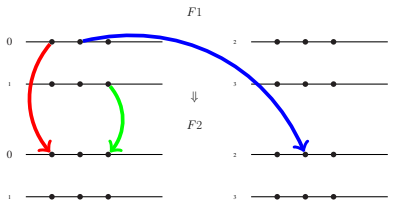
$$Q(0,0,1) = \tfrac{1}{2}(1-r_{12})r_{23}[Q(0,0,0)+Q(2,2,2)] + \tfrac{1}{4}(1-r_{12})[Q(0,0,2)+Q(2,2,0)] + \tfrac{1}{4}r_{13}[Q(0,2,0)Q(2,0,2)] + \tfrac{1}{4}r_{23}[Q(0,2,2)+Q(2,0,0)] \quad \text{(S7)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

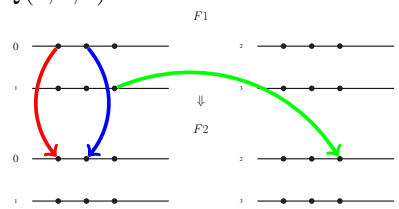
$$Q(0,0,1) = (1-r_{12})r_{23}Q(0,0,0) + \tfrac{1}{2}(1-r_{12})Q(0,0,2) + \tfrac{1}{2}r_{13}Q(0,2,0) + \tfrac{1}{2}r_{23}Q(0,2,2)$$



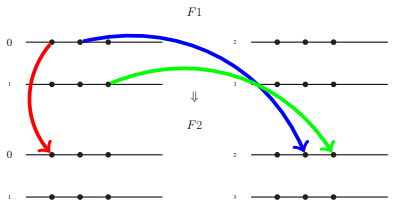
(a) $Q(0,0,1) : \frac{1}{2} \times (1-r_{12}) \times r_{23} \times Q(0,0,0)$



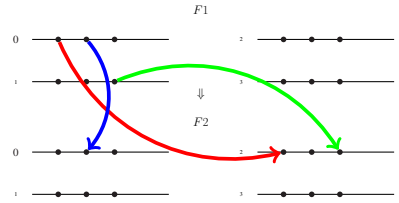
(b) $Q(0,0,1) : \frac{1}{2} \times \frac{1}{2} \times r_{13} \times Q(0,2,0)$



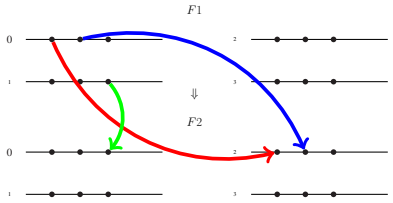
(c) $Q(0,0,1) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{12}) \times Q(0,0,2)$



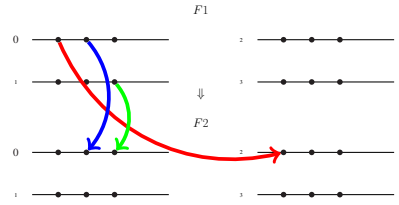
(d) $Q(0,0,1) : \frac{1}{2} \times \frac{1}{2} \times r_{23} \times Q(0,2,2)$



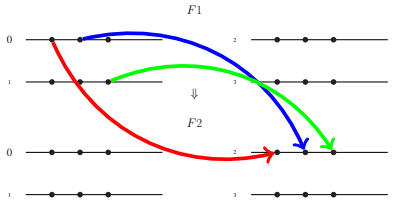
(e) $Q(0,0,1) : \frac{1}{2} \times \frac{1}{2} \times r_{13} \times Q(2,0,2)$



(f) $Q(0,0,1) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{12}) \times Q(2,2,0)$



(g) $Q(0,0,1) : \frac{1}{2} \times \frac{1}{2} \times r_{23} \times Q(2,0,0)$



(h) $Q(0,0,1) : \frac{1}{2} \times (1-r_{12}) \times r_{23} \times Q(2,2,2)$

Figure S3: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S7 for $Q(0,0,1)$.

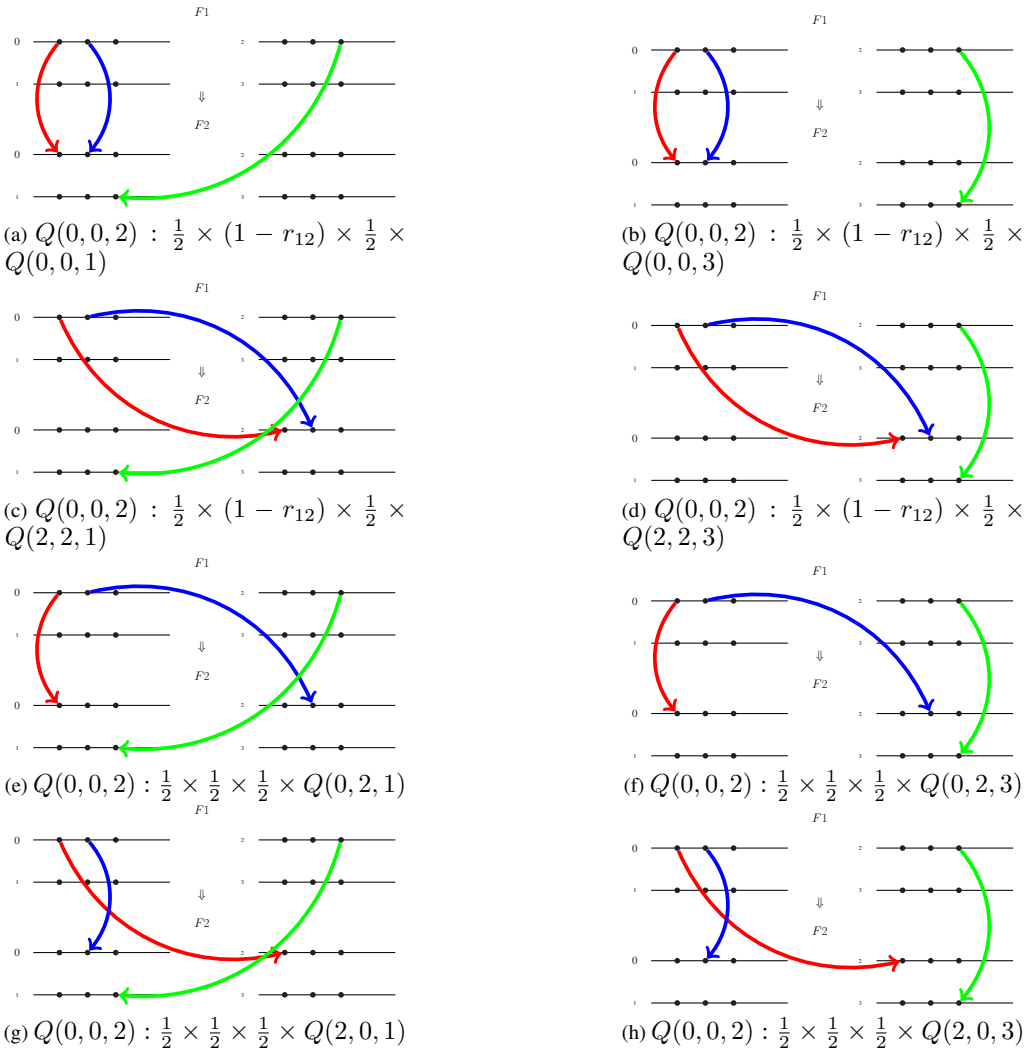
3.3 The self consistent equation for $Q(0,0,2)$

Figure S4 displays the 8 factors in the self-consistent equation for $Q(0,0,2)$:

$$Q(0,0,2) = \tfrac{1}{4}(1-r_{12})[Q(0,0,1)+Q(2,2,3)] + \tfrac{1}{4}(1-r_{12})[Q(0,0,3)+Q(2,2,1)] + \tfrac{1}{8}[Q(0,2,1)+Q(2,0,3)] + \tfrac{1}{8}[Q(0,2,3)+Q(2,0,1)] \quad \text{(S8)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

$$Q(0,0,2) = \tfrac{1}{2}(1-r_{12})Q(0,0,1) + \tfrac{1}{2}(1-r_{12})Q(0,0,2) + \tfrac{1}{4}Q(0,2,1) + \tfrac{1}{4}Q(0,2,3)$$



(a) $Q(0,0,2) : \frac{1}{2} \times (1-r_{12}) \times \frac{1}{2} \times Q(0,0,1)$

(b) $Q(0,0,2) : \frac{1}{2} \times (1-r_{12}) \times \frac{1}{2} \times Q(0,0,3)$

(c) $Q(0,0,2) : \frac{1}{2} \times (1-r_{12}) \times \frac{1}{2} \times Q(2,2,1)$

(d) $Q(0,0,2) : \frac{1}{2} \times (1-r_{12}) \times \frac{1}{2} \times Q(2,2,3)$

(e) $Q(0,0,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,2,1)$

(f) $Q(0,0,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,2,3)$

(g) $Q(0,0,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,0,1)$

(h) $Q(0,0,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,0,3)$

Figure S4: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S8 for $Q(0,0,2)$.

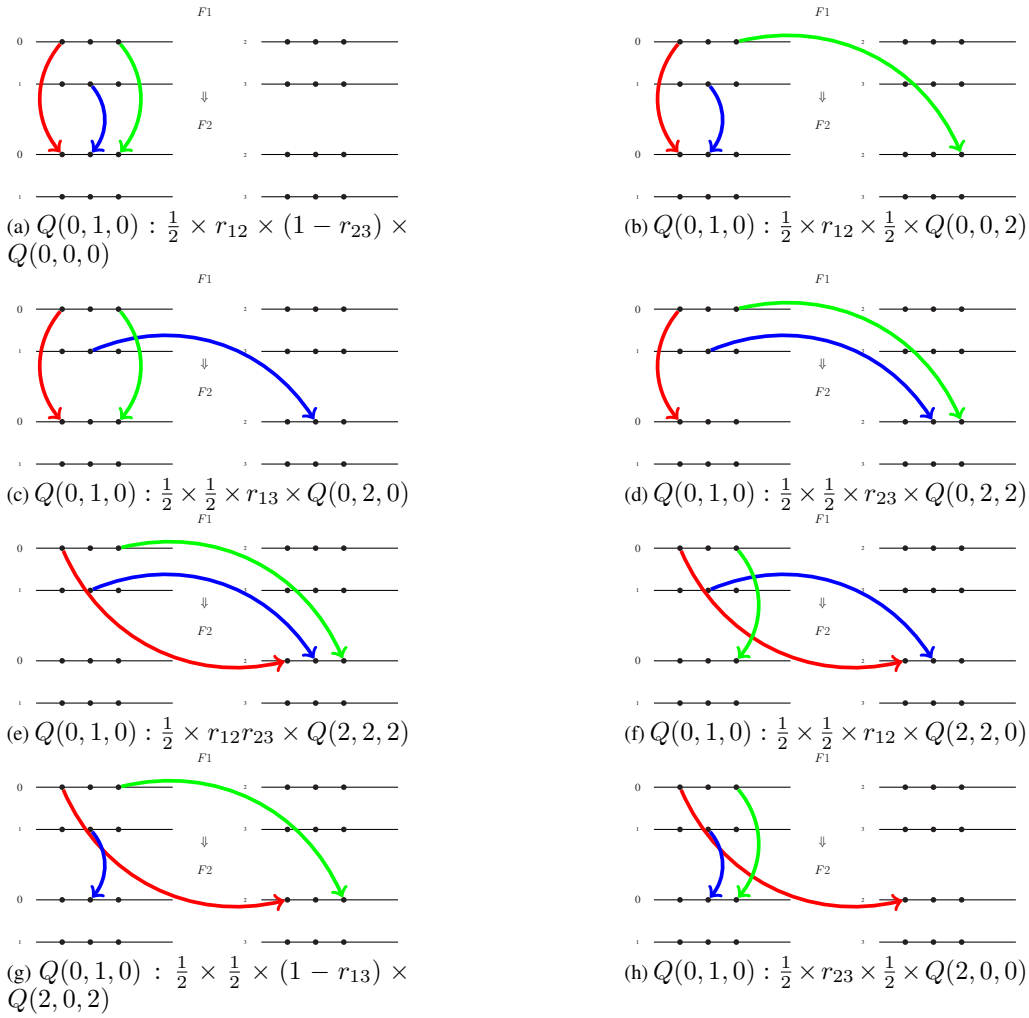
3.4 The self consistent equation for $Q(0,1,0)$

Figure S5 displays the 8 factors in the self-consistent equation for $Q(0,1,0)$:

$$Q(0,1,0) = \tfrac{1}{2}r_{12}r_{23}[Q(0,0,0)+Q(2,2,2)] + \tfrac{1}{4}r_{12}[Q(0,0,2)+Q(2,2,0)] + \tfrac{1}{4}(1-r_{13})[Q(0,2,0)+Q(2,0,2)] + \tfrac{1}{4}r_{23}[Q(0,2,2)+Q(2,0,0)] \quad \text{(S9)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

$$Q(0,1,0) = r_{12}r_{23}Q(0,0,0) + \tfrac{1}{2}r_{12}Q(0,0,2) + \tfrac{1}{2}(1-r_{13})Q(0,2,0) + \tfrac{1}{2}r_{23}Q(0,2,2)$$



(a) $Q(0,1,0) : \frac{1}{2} \times r_{12} \times (1-r_{23}) \times Q(0,0,0)$

(b) $Q(0,1,0) : \frac{1}{2} \times r_{12} \times \frac{1}{2} \times Q(0,0,2)$

(c) $Q(0,1,0) : \frac{1}{2} \times \frac{1}{2} \times r_{13} \times Q(0,2,0)$

(d) $Q(0,1,0) : \frac{1}{2} \times \frac{1}{2} \times r_{23} \times Q(0,2,2)$

(e) $Q(0,1,0) : \frac{1}{2} \times r_{12}r_{23} \times Q(2,2,2)$

(f) $Q(0,1,0) : \frac{1}{2} \times \frac{1}{2} \times r_{12} \times Q(2,2,0)$

(g) $Q(0,1,0) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{13}) \times Q(2,0,2)$

(h) $Q(0,1,0) : \frac{1}{2} \times r_{23} \times \frac{1}{2} \times Q(2,0,0)$

Figure S5: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S9 for $Q(0,1,0)$.

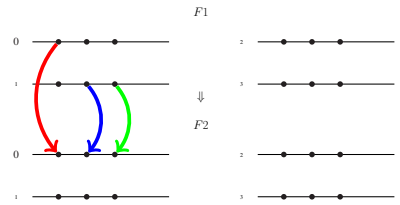
3.5 The self consistent equation for $Q(0,1,1)$

Figure S6 displays the 8 factors in the self-consistent equation for $Q(0,1,1)$:

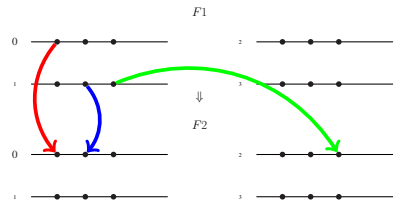
$$Q(0,1,1) = \tfrac{1}{2}r_{12}(1-r_{23})[Q(0,0,0)+Q(2,2,2)] + \tfrac{1}{4}r_{12}[Q(0,0,2)+Q(2,2,0)] + \tfrac{1}{4}r_{13}[Q(0,2,0)+Q(2,0,2)] + \tfrac{1}{4}(1-r_{23})[Q(0,2,2)+Q(2,0,0)] \quad \text{(S10)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

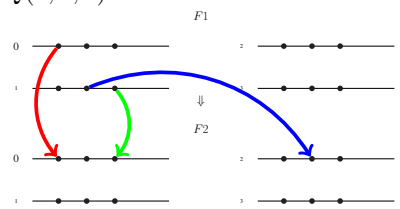
$$Q(0,1,1) = r_{12}(1-r_{23})Q(0,0,0) + \tfrac{1}{2}r_{12}Q(0,0,2) + \tfrac{1}{2}r_{13}Q(0,2,0) + \tfrac{1}{2}(1-r_{23})Q(0,2,2)$$



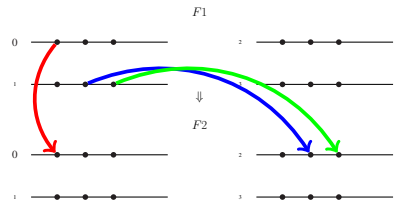
(a) $Q(0,1,1) : \frac{1}{2} \times r_{12} \times (1-r_{23}) \times Q(0,0,0)$



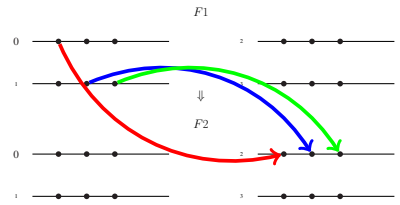
(b) $Q(0,1,1) : \frac{1}{2} \times \frac{1}{2} \times r_{12} \times Q(0,0,2)$



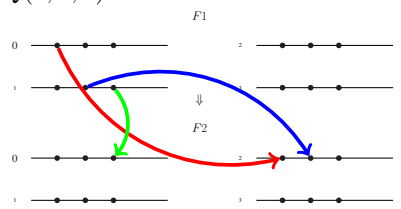
(c) $Q(0,1,1) : \frac{1}{2} \times \frac{1}{2} \times r_{13} \times Q(0,2,0)$



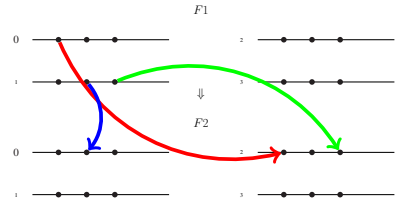
(d) $Q(0,1,1) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{23}) \times Q(0,2,2)$



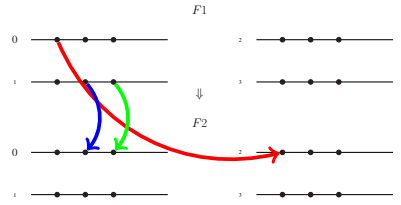
(e) $Q(0,1,1) : \frac{1}{2} \times r_{12} \times (1-r_{23}) \times Q(2,2,2)$



(f) $Q(0,1,1) : \frac{1}{2} \times r_{12} \times \frac{1}{2} \times Q(2,2,0)$



(g) $Q(0,1,1) : \frac{1}{2} \times r_{13} \times \frac{1}{2} \times Q(2,0,2)$



(h) $Q(0,1,1) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{23}) \times Q(2,0,0)$

Figure S6: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S10 for $Q(0,1,1)$.

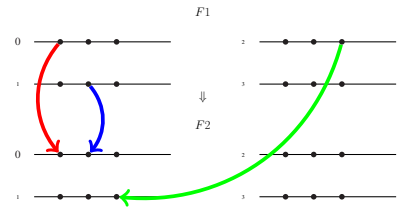
3.6 The self consistent equation for $Q(0,1,2)$

Figure S7 displays the 8 factors in the self-consistent equation for $Q(0,1,2)$:

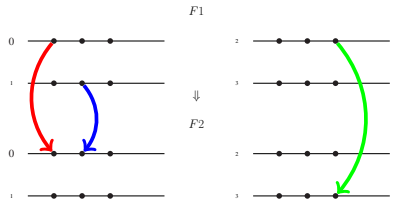
$$Q(0,1,2) = \tfrac{1}{4}r_{12}[Q(0,0,1)+Q(2,2,3)] + \tfrac{1}{4}r_{12}[Q(0,0,3)+Q(2,2,1)] + \tfrac{1}{8}[Q(0,2,1)+Q(2,0,3)] + \tfrac{1}{8}[Q(0,2,3)+Q(2,0,1)] \quad \text{(S11)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

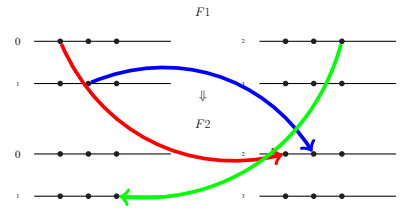
$$Q(0,1,2) = \tfrac{1}{2}r_{12}Q(0,0,1) + \tfrac{1}{2}r_{12}Q(0,0,2) + \tfrac{1}{4}Q(0,2,1) + \tfrac{1}{4}Q(0,2,3)$$



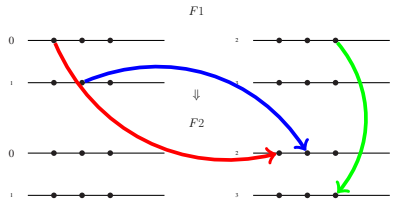
(a) $Q(0,1,2) : \frac{1}{2} \times \frac{1}{2} \times r_{12} \times Q(0,0,1)$



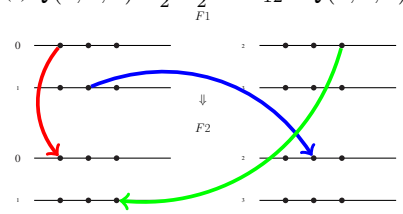
(b) $Q(0,1,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times r_{12} \times Q(0,0,3)$



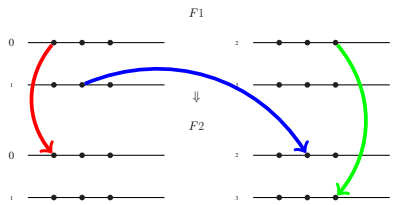
(c) $Q(0,1,2) : \frac{1}{2} \times \frac{1}{2} \times \times r_{12} \times Q(2,2,1)$



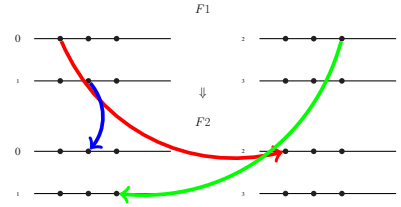
(d) $Q(0,1,2) : \frac{1}{2} \times \frac{1}{2} \times r_{12} \times Q(2,2,3)$



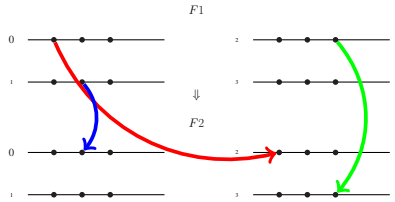
(e) $Q(0,1,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,2,1)$



(f) $Q(0,1,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,2,3)$



(g) $Q(0,1,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,0,1)$



(h) $Q(0,1,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,0,3)$

Figure S7: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S11 for $Q(0,1,2)$.

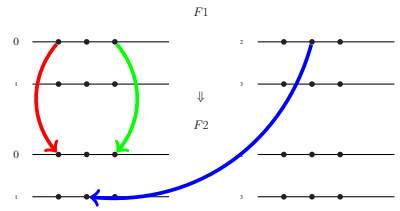
3.7 The self consistent equation for $Q(0,2,0)$

Figure S8 displays the 8 factors in the self-consistent equation for $Q(0,2,0)$:

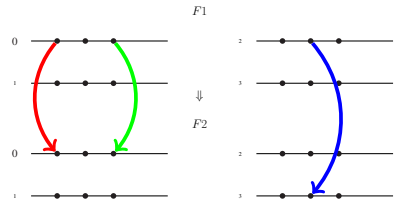
$$Q(0,2,0) = \tfrac{1}{4}(1-r_{13})[Q(0,1,0)+Q(2,3,2)] + \tfrac{1}{8}[Q(0,1,2)+Q(2,3,0)] + \tfrac{1}{4}(1-r_{13})[Q(0,3,0)+Q(2,1,2)] + \tfrac{1}{8}[Q(0,1,3)+Q(2,1,0)] \quad \text{(S12)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

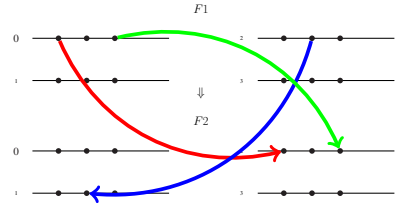
$$Q(0,2,0) = \tfrac{1}{2}(1-r_{13})Q(0,1,0) + \tfrac{1}{4}Q(0,1,2) + \tfrac{1}{2}(1-r_{13})Q(0,2,0) + \tfrac{1}{4}Q(0,1,2)$$



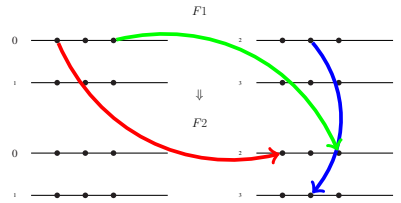
(a) $Q(0,2,0) : \frac{1}{2} \times (1-r_{13}) \times \frac{1}{2} \times Q(0,1,0)$



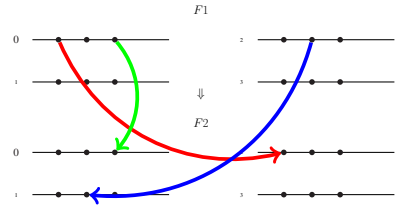
(b) $Q(0,2,0) : \frac{1}{2} \times (1-r_{13}) \times \frac{1}{2} \times Q(0,3,0)$



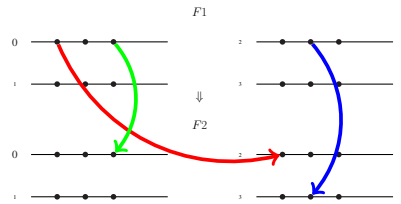
(c) $Q(0,2,0) : \frac{1}{2} \times (1-r_{13}) \times \frac{1}{2} \times Q(2,1,2)$



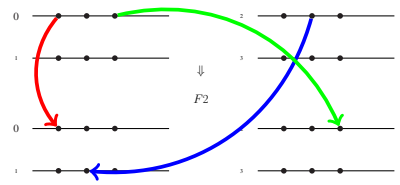
(d) $Q(0,2,0) : \frac{1}{2} \times (1-r_{13}) \times \frac{1}{2} \times Q(2,3,2)$



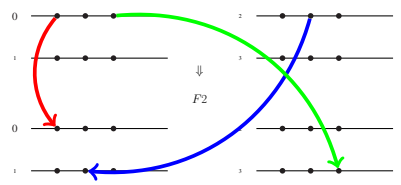
(e) $Q(0,2,0) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,1,0)$



(f) $Q(0,2,0) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,3,0)$



(g) $Q(0,2,0) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,1,2)$



(h) $Q(0,2,0) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,1,3)$

Figure S8: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S12 for $Q(0,2,0)$.

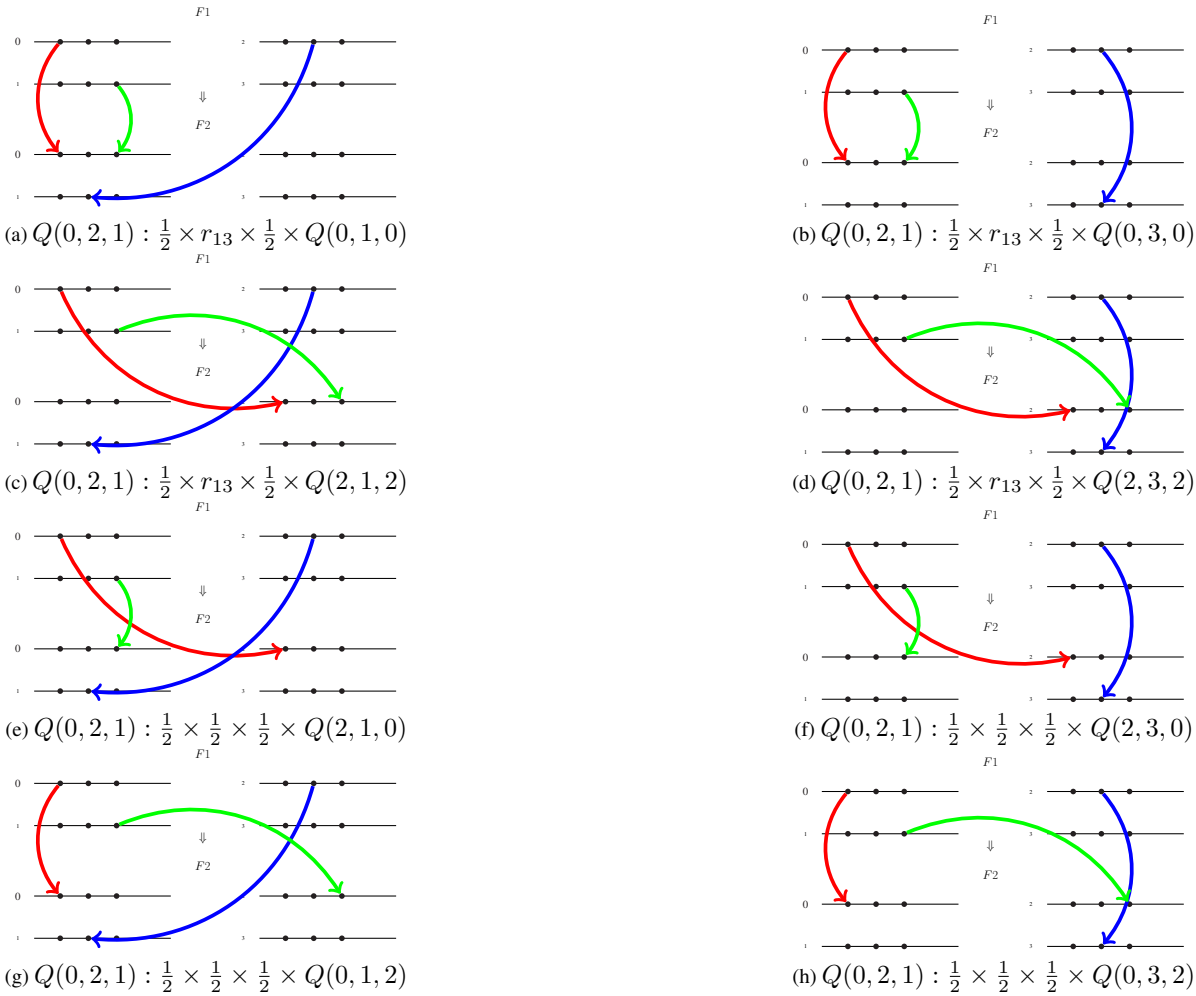
3.8 The self consistent equation for $Q(0,2,1)$

Figure S9 displays the 8 factors in the self-consistent equation for $Q(0,2,1)$:

$$Q(0,2,1) = \tfrac{1}{4}r_{13}[Q(0,1,0)+Q(2,3,2)] + \tfrac{1}{8}[Q(0,1,2)+Q(2,3,0)] + \tfrac{1}{4}r_{13}[Q(0,3,0)+Q(2,1,2)] + \tfrac{1}{8}[Q(0,3,2)+Q(2,1,0)] \quad \text{(S13)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

$$Q(0,2,1) = \tfrac{1}{2}r_{13}Q(0,1,0) + \tfrac{1}{4}Q(0,1,2) + \tfrac{1}{2}r_{13}Q(0,2,0) + \tfrac{1}{4}Q(0,2,3)$$



(a) $Q(0,2,1) : \frac{1}{2} \times r_{13} \times \frac{1}{2} \times Q(0,1,0)$

(b) $Q(0,2,1) : \frac{1}{2} \times r_{13} \times \frac{1}{2} \times Q(0,3,0)$

(c) $Q(0,2,1) : \frac{1}{2} \times r_{13} \times \frac{1}{2} \times Q(2,1,2)$

(d) $Q(0,2,1) : \frac{1}{2} \times r_{13} \times \frac{1}{2} \times Q(2,3,2)$

(e) $Q(0,2,1) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,1,0)$

(f) $Q(0,2,1) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,3,0)$

(g) $Q(0,2,1) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,1,2)$

(h) $Q(0,2,1) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,3,2)$

Figure S9: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S13 for $Q(0,2,1)$.

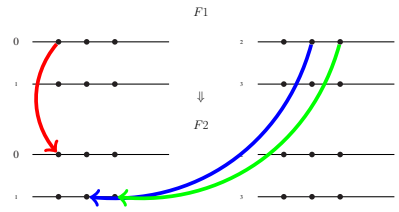
3.9 The self consistent equation for $Q(0,2,2)$

Figure S10 displays the 8 factors in the self-consistent equation for $Q(0,2,2)$:

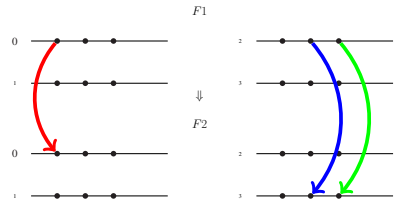
$$Q(0,2,2) = \tfrac{1}{4}(1-r_{23})[Q(0,1,1)+Q(2,3,3)] + \tfrac{1}{8}[Q(0,1,3)+Q(2,3,1)] + \tfrac{1}{8}[Q(0,3,1)+Q(2,1,3)] + \tfrac{1}{4}(1-r_{23})[Q(0,3,3)+Q(2,1,1)] \quad \text{(S14)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

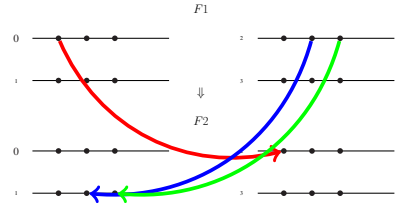
$$Q(0,2,2) = \tfrac{1}{2}(1-r_{23})Q(0,1,1) + \tfrac{1}{4}Q(0,1,2) + \tfrac{1}{4}Q(0,2,1) + \tfrac{1}{2}(1-r_{23})Q(0,2,2)$$



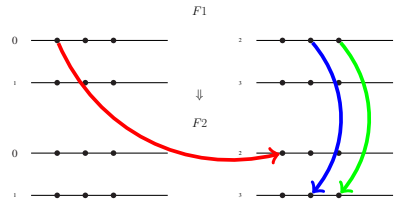
(a) $Q(0,2,2) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{23}) \times Q(0,1,1)$



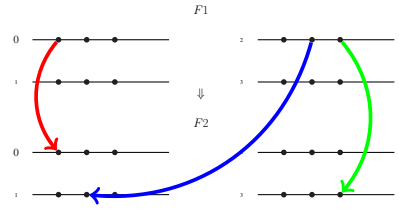
(b) $Q(0,2,2) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{23}) \times Q(0,3,3)$



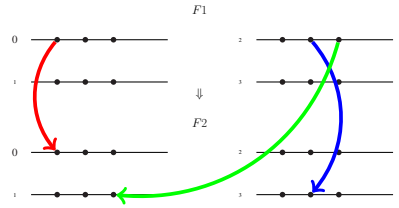
(c) $Q(0,2,2) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{23}) \times Q(2,1,1)$



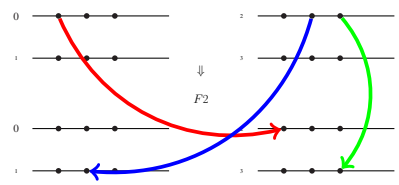
(d) $Q(0,2,2) : \frac{1}{2} \times \frac{1}{2} \times (1-r_{23}) \times Q(2,3,3)$



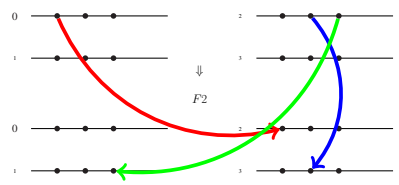
(e) $Q(0,2,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,1,3)$



(f) $Q(0,2,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,3,1)$



(g) $Q(0,2,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,1,3)$



(h) $Q(0,2,2) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,3,1)$

Figure S10: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S14 for $Q(0,2,2)$.

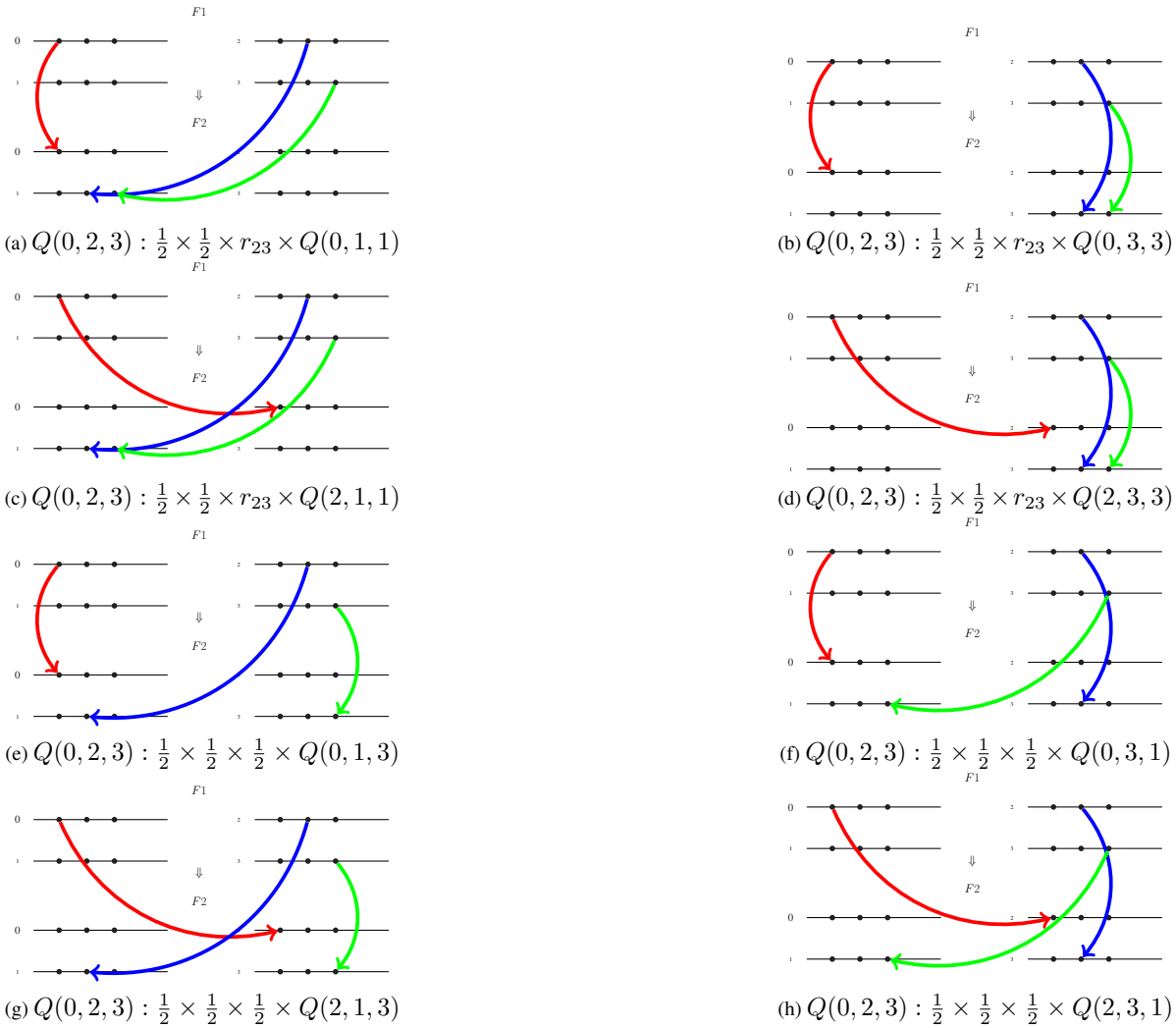
3.10 The self consistent equation for $Q(0,2,3)$

Figure S11 displays the 8 factors in the self-consistent equation for $Q(0,2,3)$:

$$Q(0,2,3) = \tfrac{1}{4}r_{23}[Q(0,1,1)+Q(2,3,3)] + \tfrac{1}{8}[Q(0,1,3)+Q(2,3,1)] + \tfrac{1}{8}[Q(0,2,1)+Q(2,1,3)] + \tfrac{1}{4}r_{23}[Q(0,3,3)+Q(2,1,1)] \quad \text{(S15)}$$

After use of symmetry to keep only non-equivalent Qs, this leads to

$$Q(0,2,3) = \tfrac{1}{2}r_{23}Q(0,1,1) + \tfrac{1}{4}Q(0,1,2) + \tfrac{1}{4}Q(0,2,1) + \tfrac{1}{2}r_{23}Q(0,2,2)$$



(a) $Q(0,2,3) : \frac{1}{2} \times \frac{1}{2} \times r_{23} \times Q(0,1,1)$

(b) $Q(0,2,3) : \frac{1}{2} \times \frac{1}{2} \times r_{23} \times Q(0,3,3)$

(c) $Q(0,2,3) : \frac{1}{2} \times \frac{1}{2} \times r_{23} \times Q(2,1,1)$

(d) $Q(0,2,3) : \frac{1}{2} \times \frac{1}{2} \times r_{23} \times Q(2,3,3)$

(e) $Q(0,2,3) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,1,3)$

(f) $Q(0,2,3) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(0,3,1)$

(g) $Q(0,2,3) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,1,3)$

(h) $Q(0,2,3) : \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times Q(2,3,1)$

Figure S11: The graphical representation of the factors multiplying each Q on the right-hand side of Eq. S15 for $Q(0,2,3)$.

Appendix D

Appendix D

Résumé en Français

Modélisation mathématique et intégration de données biologiques complexes : analyse du phénomène d'hétérosis chez la levure

Les espèces de levure du groupe phylogénétique *Saccharomyces sensu stricto*, y compris *S. cerevisiae* et *S. uvarum* étudiées ici, sont importantes dans de nombreux domaines tels que l' agriculture, la biotechnologie et la médecine. Outre son utilité pour répondre aux besoins et aux coutumes de l'homme, la levure représente un système modèle puissant pour traiter les problèmes fondamentaux en biologie. Son temps de génération réduit et sa facilité de culture et de manipulation en laboratoire ont permis de réaliser des avancées majeures. En particulier, le séquençage complet du génome de *S. cerevisiae* (Goffeau et al., 1996) a accompagné un changement de perspective partant de l'analyse individuel des gènes et ces fonctions à une vision globale d'interaction des réseaux cellulaires, ce qui a suscité un intérêt renouvelé pour le métabolisme et sa régulation. Une caractéristique frappante du métabolisme est la similitude des voies fondamentales, même entre des espèces éloignées telles que la levure et l'homme, ce qui permet par exemple d'étudier chez la levure les voies impliquées dans les maladies humaines. Cependant, l'utilisation et la régulation des voies peuvent entraîner d'énormes différences phénotypiques entre les espèces voisines, ce qui soulève la question de la relation génotype-phénotype.

Dans ce contexte, mon travail de thèse porte sur la question générale de la relation génotype-phénotype, en prêtant une attention particulière à l'étude de la vigueur hybride (ou hétérosis) chez la levure. Pour cela j'ai utilisé une approche associant biologie, mathématiques et statistiques et je me suis appuyé sur un jeu de données généré lors du projet ANR interdisciplinaire *HeterosYeast: exploitation du phénomène d'hétérosis pour l'amélioration des levures d'œnologie* (décrit dans le chapitre 2). Ce jeu de données a été recueilli sur un dispositif demi-diallèle construit en réalisant tous les croisement deux à deux entre sept souches de *S. cerevisiae* et quatre de *S. uvarum* (pour un total de 11 souches parentales et 55 hybrides) dans deux températures (18°C et 26°C) et il est composé d'un nombre de données hétérogènes correspondant à différentes niveaux d'intégration phénotypique, de la protéomique aux traits d'histoire de vie, au cours du processus de fermentation du vin blanc (figure D1).

L'ensemble des observations, organisé en types de croisements (inter et intra-spécifique) et associé à différentes niveaux de complexité cellulaire, était idéalement adapté pour la modélisation multi-échelle et pour tester des modèles de prédiction de la variation de phénotypes

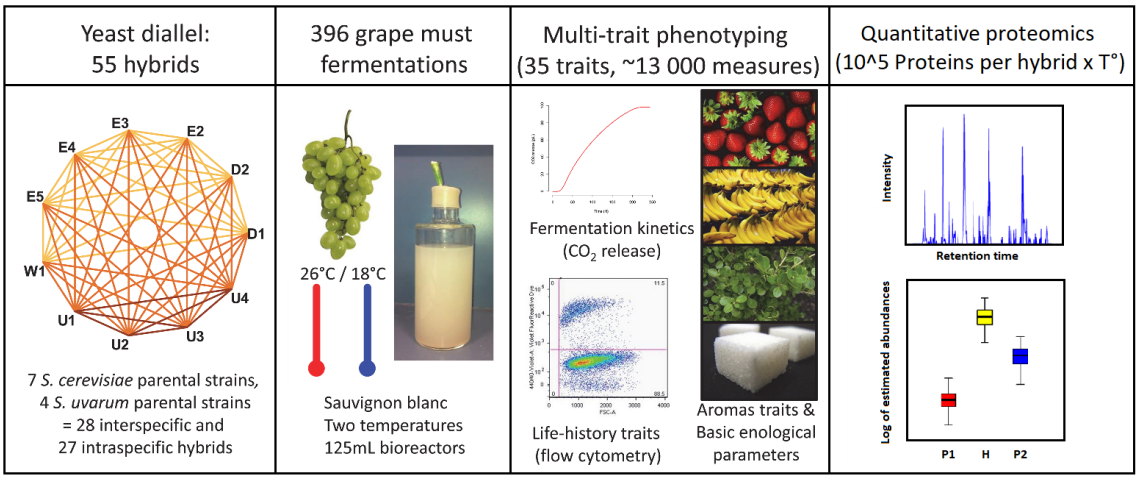


Figure D1: Protocole expérimental. Des souches diploïdes entièrement homozygotes ont été utilisées comme souches parentales selon un schéma semi-diallèle. W1, D1, D2, E2, E3, E4 et E5 sont des souches *S. cerevisiae*, U1, U2, U3 et U4 *S. uvarum*. Les fermentations ont été effectuées dans jus de raisin cépage Sauvignon blanc à 18°C et 26°C en triple exemplaire dans des fermenteurs pour un total de 396 expériences. Trente-cinq traits ont été rassemblés et regroupés en quatre classes (traits de cinétique de fermentation, traits d'histoire de vie, paramètres œnologiques de base et traits aromatiques). Les abondances de protéines ont été quantifiées pour chaque combinaison de souche × température (da Silva et al., 2015).

intégrés à partir de caractères protéiques et métaboliques (flux), ce qui m'a permis de proposer des approches de modélisation statistique et d'intégration de données biologiques afin de : (*i*) analyser la variation phénotypique du point de vue quantitatif et de la génétique des populations ; (*ii*) étudier la carte génotype-phénotype du point de vue de la biologie des systèmes évolutifs.

Cet ensemble de données a permis de questionner la relation complexe entre génotypes, phénotypes et fitness dans les populations. De plus, des développements liés à une meilleure compréhension de la structure de la diversité phénotypique des levures et du processus de fermentation du vin, ainsi que des développements méthodologiques, sont proposés dans ma thèse. Ces méthodes ont en réalité un large domaine d'applicabilité.

Ci-dessous Vous trouverez un résumé des modèles employés au cours de ma thèse, les majeurs résultats associés ainsi que des perspectives futures.

L'évolution des traits d'histoire de vie

La première approche de modélisation a été introduite dans le chapitre 1, dans lequel la variation phénotypique est présentée comme le résultat des processus d'évolution et d'adaptation. Un composant clé de l'adaptation et de l'évolutivité est la partition de la variance phénotypique en composants génétiques additifs et non additifs, et en composants environnementaux

(G × E et résiduels). Dans ce contexte, la conception du demi-diallèle du projet *HeterosYeast* présentait un intérêt particulier. Parmi toutes les approches statistiques proposées dans la littérature pour analyser les composants de variance génétique et non génétique à partir de tels modèles, j'ai décidé de adapter le modèle proposé par Lenarcic et al. (2012) à la structure particulier du semi-diallèle, qui inclut la diagonale avec les souches parentales consanguines de deux espèces. J'ai donc inclus dans mon modèle les effets additifs intra et inter-spécifiques, les effets de dépression de consanguinité et les effets d'hétérosis intra et inter-spécifiques.

Formellement, prenons y_{ijk} le phénotype observé pour le croisement entre les parents i et j dans la réplique k. Le modèle est défini par :

$$\begin{aligned} y_{ijk} = \mu + I_{s(i)=s(j)}(A_{w_i} + A_{w_j}) + I_{s(i)\neq s(j)}(A_{b_i} + A_{b_j}) + \\ + I_{i\neq j}(I_{s(i)=s(j)}H_{w_{ij}} + I_{s(i)\neq s(j)}H_{b_{ij}}) + \\ + I_{i=j}(\beta_{s(i)} + B_i) + \epsilon_{ijk}, \end{aligned} \tag{D1}$$

Où :

- μ est la moyenne globale ;
- $s(i)$ associe à chaque souche parentale i l'espèce à laquelle il appartient :

$$s(i) \in \{S.\ cerevisiae, S.\ uvarum\}$$

- A_{w_i} et A_{b_i} notent, respectivement, les contributions additives de la souche i en croisements intra-spécifiques (au sein d'une espèce, *i.e.* $s(i) = s(j)$), et inter-spécifiques (entre espèces, *i.e.* $s(i) \neq s(j)$);
- $H_{w_{ij}}$ et $H_{b_{ij}}$ désignent l'effet d'interaction entre les parents (i, j) en croisement intra-spécifique (au sein d'une espèce) et inter-spécifique (entre les espèces), respectivement. Dans notre conception demi-diallèle sans croisements réciproques, $H_{w_{ij}}$ et $H_{b_{ij}}$ sont supposées être symétriques, *i.e.* $H_{w_{ij}} = H_{w_{ji}}$ et $H_{b_{ij}} = H_{b_{ji}}$. Ces effets sont désignés sous le nom d'effets d'hétérosis intra et inter-spécifiques, respectivement ;
- $\beta s(i)$ et B_i sont respectivement l'écart par rapport à l'effet global fixé pour l'espèce $s(i)$ et la contribution de la souche i associée à la souche dans la cas des lignées pures. Dans la suite, je ferai référence à B_i en tant qu'effet de dépression de consanguinité ;
- ϵ_{ijk} est le résidu, l'écart spécifique de l'individu ijk;
- $I_{condition}$ est une variable indicatrice. Sa valeur est égale à 1 si la condition est remplie et à 0 sinon.

Tous les effets génétiques ont été considérés comme des variables aléatoires tirées d'une distribution normale. Formellement, si $\boldsymbol{q} \in \{\boldsymbol{A_w}, \boldsymbol{A_b}, \boldsymbol{B}, \boldsymbol{H_w}, \boldsymbol{H_b}\}$ est l'effet génétique sous considération:

$$\forall i \quad q_i \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{q}}). \tag{D2}$$

Le modèle génétique complet à effets mixtes est donc défini par trois effets fixes (l'interception μ et les effets de depression de consanguinité β_{Su} et β_{Sc}) et cinq variances des effets génétiques aléatoires ($\sigma^2_{\boldsymbol{A_w}}$, $\sigma^2_{\boldsymbol{A_b}}$, $\sigma^2_{\boldsymbol{B}}$, $\sigma^2_{\boldsymbol{H_w}}$, $\sigma^2_{\boldsymbol{H_b}}$).

Les analyses ont été réalisées aux deux températures indépendamment et pour chaque trait séparément (35 phénotypes et 615 protéines). Chaque combinaison trait $\times$ température a été caractérisé par ses composantes de variance et des comparaisons ont été effectuées entre eux. Avant l'interprétation des résultats, une étude de simulation a été réalisée pour s'assurer que le faible nombre de parents présents dans le plan d'expérience n'induisaient pas de biais d'estimation des paramètres.

La sortie de cette analyse est un ensemble de vecteurs de 5 coordonnées décrivant les composantes de la variance, ce qui a nécessité une nouvelle modélisation pour interpréter ces résultats. La stratégie adoptée a été d'utiliser un modèle de mélange gaussien diagonal sur les 615 vecteurs de variance estimés sur les protéines pour obtenir une classification en 9 groupes (figure D2) puis de classer les autres traits phénotypiques dans ces 9 classes en utilisant le modèle de mélange comme un modèle d'apprentissage.

Les résultats sont rapportés au chapitre 3. En bref, ce travail a révélé des interactions génotype $\times$ environnement à tous les niveaux d'organisation cellulaire (les composantes de la variance différaient entre les deux températures) et a permis de classer les combinaisons traits $\times$ température en groupes clairement distincts, ce qui excluait l'hypothèse que tous les traits ont évolué de manière neutre au cours du même processus évolutif. Une interprétation possible est que les traits partageant un profil de composant de variance similaire ont une histoire évolutive commune. De plus, au sein de certains groupes de caractères, j'ai montré que les composantes de variance de la dépression de consanguinité et de l'hétérosis étaient découplées. Ce résultat original met en évidence que dépression de consanguinité et hétérosis ont évolué indépendamment. J'ai également montré que l'épistasie est nécessairement impliquée dans ce découplage. Ces résultats (publiés en *Genetics*, Petrizzelli et al. (2019)) appellent des développements théoriques en génétique évolutive pour identifier les mécanismes et les forces motrices en jeu, et des expériences sur d'autres espèces pour évaluer si tels résultats sont communs dans les systèmes biologiques.

Du point de vue des sélectionneurs, l'analyse susmentionnée a permis d'inférer la matrice de variance-covariance entre les effets génétiques additifs pour les caractères analysés dans la population de levure du projet HeterosYeast. En utilisant l'équation bien connue de la réponse à la sélection (Chapitre 1, section 1.1.4), il est possible de prédire les résultats d'une génération de sélection. Considérons par exemple le tableau D1 dans lequel sont énumérés les caractéristiques de fermentation souhaitables pour la production de vin blanc (Philippe Marullo, communication personnelle). Il est possible de construire un indice de sélection prenant en

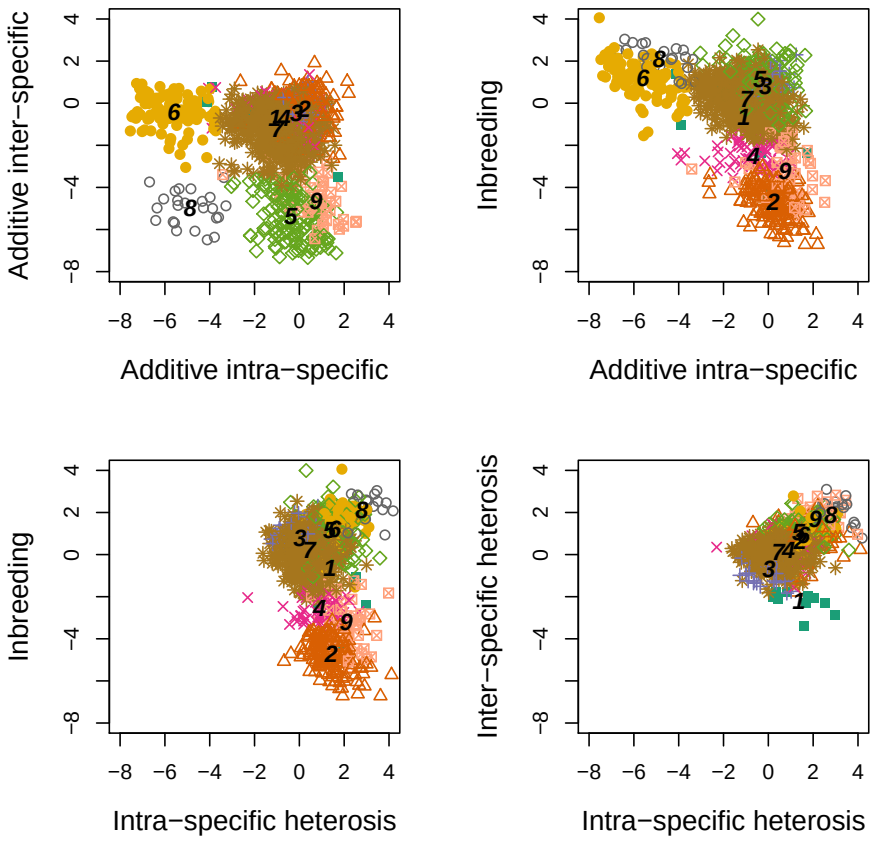


Figure D2: Patrons de corrélations entre les composantes de la variance génétique des abondances des protéines. Les points correspondent aux protéines, les combinaisons de types et de couleurs identifient les grappes obtenues par leur classification basée sur un modèle de mélange gaussien. Les nombres de 1 à 9 identifient les centres de classe pour chaque groupe.

compte la valeur observée pour les caractères sélectionnés et le coefficient de pondération associé. L'approche naïve consiste à considérer la somme pondérée entre ces deux quantités. Ainsi, la sélection peut être effectuée sur des croisements présentant une valeur d'indice supérieure à un certain seuil et le calcul du gradient de sélection est simple. La réponse à l'équation de sélection renverrait la valeur phénotypique moyenne attendue de la progéniture à la génération suivante. Cela renverrait également la réponse attendue à la sélection pour les traits qui ne sont pas sélectionnés directement. Par conséquent, en utilisant la méthode que j'ai proposée pour l'estimation des composantes de la variance, c'est possible de prédire l'évolution de caractères non sélectionnés, y compris l'abondance de protéines, après une génération de sélection.

De plus, les composants additifs de la matrice de covariance de variance associée à la population *HeterosYeast* correspondent à la fameuse matrice G des paysages de fitness. Les vecteurs propres associés à la matrice G révéleraient les directions possibles de l'évolution et pourraient aider à comprendre la géométrie du paysage de remise en forme multi-trait chez la levure.

	Objectifs		Coefficient	
Trait	**Blanc 18 garde**	**Blanc 18 primeur**	**Blanc 18 garde**	**Blanc 18 primeur**
AFtime	min	min	1	1
t.lag	min	min	1	1
Hexanol	low	low	0,25	0,25
Octanoic acid	low		0,1	
Phenylethanol-acetate	low	high	0,5	0,5
Isoamyl-acetate	low	high	0,5	0,5
Residual sugar	<2	<2		
Phenylethanol	low	high<400	0,5	0,5
4MMP	high	high	1	1
Decanoic acid	low		0,1	
SO_2L/SO_2T	high	high	0,5	0,5

Table D1: Objectifs de la fermentation des moûts de raisins blancs. Objectifs pour les caractères présentant un intérêt œnologique pour la fermentation des moûts de raisins à 18 degrés pour les vins *garde* et *primeur*. Un coefficient de pondération basé sur les intérêts œnologiques est attribué à chaque objectif. Les objectifs peuvent changer avec le type de vin souhaité.

En général, le modèle statistique proposé dans cette première approche de modélisation peut être utilisé pour tout problème concernant les interactions par paires entre entités physiques ou biologiques. En écologie, le même modèle pourrait être utilisé pour étudier la compétition entre individus pour des ressources dans un environnement donné, pour accéder aux performances en mélanges et pour quantifier la capacité de mélange de groupes de génotypes, de populations ou d'espèces.

Biologie intégrative

Dans la seconde approche de modélisation, les phénotypes au niveau plus intégré sont considérés comme résultant des processus d'intégration de multiples échelles cellulaires. Dans ce contexte, j'ai proposé de prédire un niveau intermédiaire d'organisation cellulaire : les flux métaboliques.

Le cadre mathématique général de la modélisation du métabolisme (au moyen de modèles basés sur des contraintes) et les méthodes classiquement utilisées pour l'inférence des flux métaboliques sont présentés au chapitre 4. L'approche de modélisation proposée est illustrée dans les chapitres 4 et 5, et a consisté à interfacer des données protéomiques avec un modèle métabolique à base de contraintes reposant sur :

- l'annotations du génome permettant l'association de réactions enzymatiques à l'expression génique/abondance de protéines à l'échelle du génome ;
- l'hypothèse selon laquelle l'abondance des protéines conditionne l'utilisation des voies et qu'il devrait exister, à l'échelle du génome, une corrélation entre les abondances et les flux de protéines.

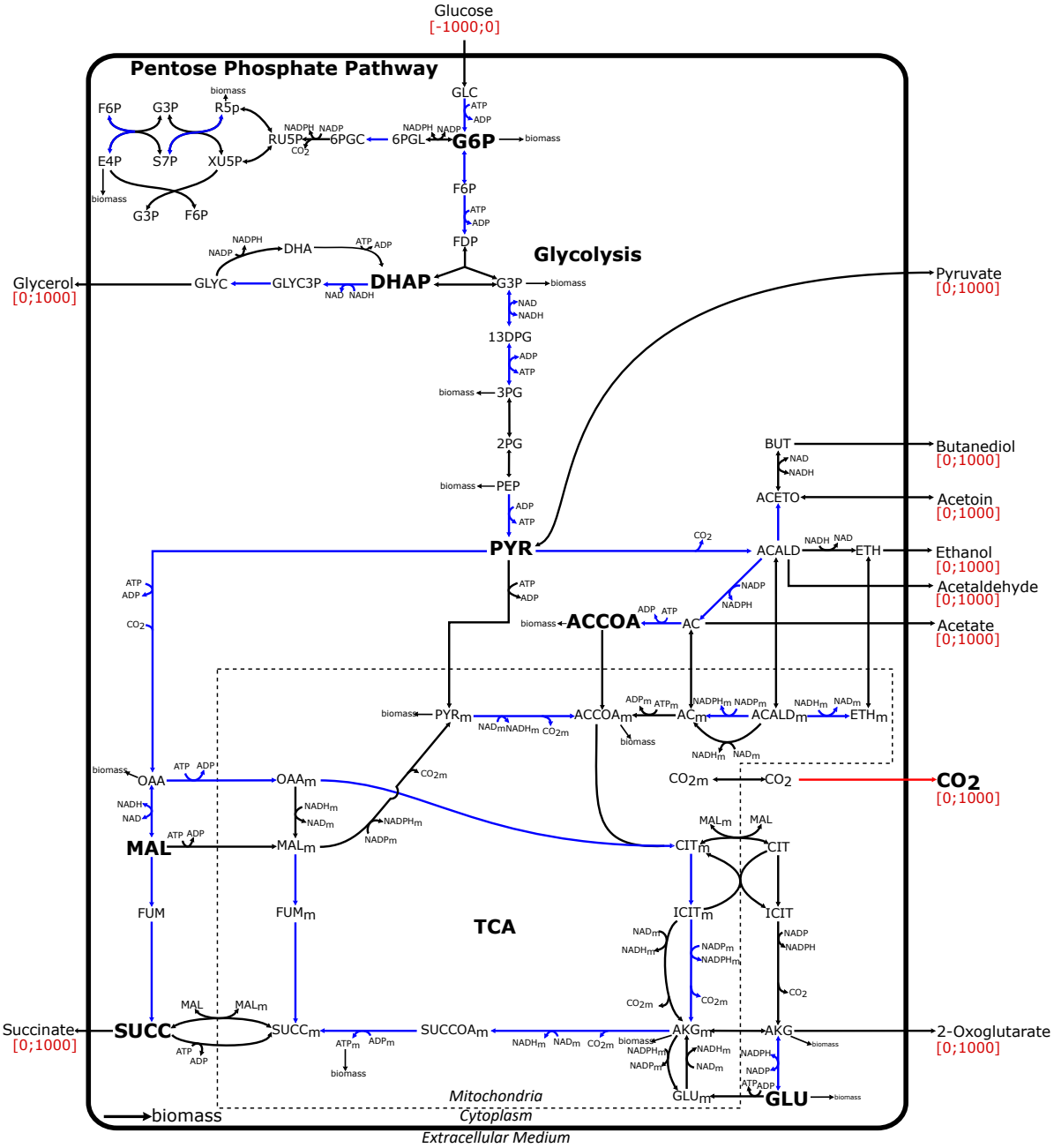


Figure D3: Représentation du modèle DynamosYeast du métabolisme carboné central chez *S. cerevisiae*. Les métabolites sont notés en noir. Les contraintes sur les flux d'échange sont en rouge entre crochets et correspondent à la fermentation, avec le glucose comme flux d'entrée unique. Les flèches bleues dénotent les réactions pour lesquels l'abondance de protéines/complexe de protéines enzymatique associée a été mesurée. La flèche rouge dénote le seul flux de sortie mesuré lors du projet HeterosYeast.

Contrairement aux approches classiques et basée sur les mêmes principes de Lee et al. (2012), mon approche est entièrement pilotée par les données et ne repose sur aucune hypothèse sur les principes d'optimisation du métabolisme cellulaire, qui sont discutables du point de vue de l'évolution. Il s'appuie sur une description probabiliste de l'espace faisable des flux, étant

donné les contraintes stœchiométriques et thermodynamiques (Braunstein et al., 2017), ainsi que sur une réduction additionnelle de l'espace faisable des flux par la valeur des flux cellulaires observés. Ensuite, parmi toutes les solutions possibles, j'ai choisi celle qui correspond le mieux à la distribution observée de l'abondance des protéines.

En utilisant un modèle stœchiométrique réduit du métabolisme central carboné de la levure, le modèle DynamoYeast, et les données HeterosYeast (figure D3), j'ai ainsi pu prédire un ensemble de flux pour chaque combinaison souche × température. Puis, j'ai comparé les patrons de corrélations entre les caractères à plusieurs niveaux d'intégration. Pour ce faire, j'ai utilisé des méthodes statistiques de pointe conçues pour des jeux de données hétérogènes de grande dimensionnalité, qui se sont révélés efficaces.

Les résultats sont rapportés au chapitre 5. En bref, une analyse canonique régularisée des corrélations a été utilisés pour identifier des liens entre les flux et les phénotypes. Les données révèlent deux grandes familles de caractères de fermentation ou de traits d'histoire de vie dont l'interprétation biochimique est cohérente en termes de trade-off, et qui n'avaient pas été mises en évidence à partir des seules données de protéomique quantitative. En particulier, la corrélation négative entre r et K s'explique par une utilisation différente du métabolisme carboné central. Un r élevé et un K faible sont associés à la glycolyse et à la fermentation, tandis que les r bas et les K élevés sont associés au cycle de Krebs et à la respiration.

Une analyse discriminante linéaire sur la matrice de correlation entre proteines (variables) et flux metaboliques (individues) a confirmé ce le lien entre la variation des traits et le flux du métabolisme carboné central. En fait, les protéines qui coïncident avec les groupes de traits associés à r et avec les flux glycolytiques et de fermentation sont enrichies en protéines impliquées dans la glycolyse et la fermentation, mais également dans la synthèse et la dégradation des protéines et le cytosquelette, qui peuvent être associés aux division cellulaire. Les protéines qui coïncident avec le groupe de traits associées à K et avec les flux de TCA et de respiration sont enrichies en protéines impliquées dans le TCA et la respiration, mais également dans le transport d'électrons, la conversion d'énergie et le métabolisme de l'azote et du soufre.

En fin, j'ai pu montrer que l'introduction d'un niveau d'intégration phénotypique supplémentaire et intermédiaire, les flux métaboliques, entre les traits protéomiques et les traits observables, permet de mieux comprendre le bien connu compromis écologique $r - K$ en tant que compromis entre les utilisations de la voie métabolique.

Le compromis $r - K$ pourrait ainsi être associé à différents modes de taux de consommation de glucose (élevé ou faible). La stratégie "ant" rappelée au chapitre 2 était associée à une reproduction rapide, à une capacité de charge élevée et à une petite cellule lors de la fermentation et à un faible taux de reproduction lors de la respiration (chapitre 2 section 2.2.1), mais aussi à un faible taux de consommation de glucose, éventuellement associé à des flux plus importants

dans les voies du pentose-phosphate.

Les choix métaboliques des espèces vivantes sont une sorte de casse-tête loin d'être parfaitement compris. Les analyses préliminaires effectuées pour étudier la stratégie FBA d'un taux de consommation de glucose inférieur ont été revues dans ce travail au chapitre 2 section 4.3.3. En comparant la solution FBA à l'espace réalisable réduit par les observations expérimentales, j'ai montré que l'utilisation de la voie du pentose-phosphate était un moyen d'économiser des ressources, en produisant de l'énergie à un prix inférieur, en termes de consommation de glucose. Il serait intéressant de faire d'autres comparaisons avec d'autres fonctions objectives pour mieux comprendre les bases métaboliques sous-jacentes de la variation des traits phénotypiques.

Les perspectives futures seraient d'appliquer la méthode à un modèle à l'échelle du génome de levure (Heavner et al., 2013), mais également à d'autres systèmes biologiques. Par exemple, il existe dans notre laboratoire une vaste collection de données protéomiques et phénotypiques recueillies sur la feuille de maïs à différents stades de développement, ainsi qu'un modèle métabolique à l'échelle du génome pour la feuille de maïs (Simons et al., 2014A,B). En combinant les données, le modèle à l'échelle du génome et la méthode proposée, je suis convaincu que cela contribuerait à renforcer les bases moléculaires de la variation du développement des feuilles.

Au-delà du développement méthodologique qui pourrait être utile à la communauté scientifique (espérons-le!), Ma thèse montre que la modélisation mathématique et statistique alliée au cadre évolutif aide à comprendre la diversité du monde vivant.

Bibliography

Braunstein, A., Muntoni, A. P. and Pagnani, A. (2017). An analytic approximation of the feasible space of metabolic networks, *Nature Communications* **8**: 14915.
URL: *http://www.nature.com/doifinder/10.1038/ncomms14915*

da Silva, T., Albertin, W., Dillmann, C., Bely, M., la Guerche, S., Giraud, C., Huet, S., Sicard, D., Masneuf-Pomarede, I., de Vienne, D. and Marullo, P. (2015). Hybridization within Saccharomyces Genus Results in Homoeostasis and Phenotypic Novelty in Winemaking Conditions, *PLoS ONE* **10**(5).
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4422614/*

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S. G. (1996). Life with 6000 Genes, *Science* **274**(5287): 546–567.
URL: *https://science.sciencemag.org/content/274/5287/546*

Heavner, B. D., Smallbone, K., Price, N. D. and Walker, L. P. (2013). Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance, *Database: The Journal of Biological Databases and Curation* **2013**.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3739857/*

Lee, D., Smallbone, K., Dunn, W. B., Murabito, E., Winder, C. L., Kell, D. B., Mendes, P. and Swainston, N. (2012). Improving metabolic flux predictions using absolute gene expression data, *BMC Systems Biology* **6**(1): 73.
URL: *https://doi.org/10.1186/1752-0509-6-73*

Lenarcic, A. B., Svenson, K. L., Churchill, G. A. and Valdar, W. (2012). A general bayesian approach to analyzing diallel crosses of inbred strains, *Genetics* **190**(2): 413–435.

Petrizzelli, M., Vienne, D. d. and Dillmann, C. (2019). Decoupling the Variances of Heterosis and Inbreeding Effects Is Evidenced in Yeast's Life-History and Proteomic Traits, *Genetics* **211**(2): 741–756.
URL: *https://www.genetics.org/content/211/2/741*

Simons, M., Saha, R., Amiour, N., Kumar, A., Guillard, L., Clément, G., Miquel, M., Li, Z., Mouille, G., Lea, P. J., Hirel, B. and Maranas, C. D. (2014B). Assessing the Metabolic Impact of Nitrogen Availability Using a Compartmentalized Maize Leaf Genome-Scale Model, *Plant Physiology* **166**(3): 1659–1674.
URL: *http://www.plantphysiol.org/content/166/3/1659*

Simons, M., Saha, R., Guillard, L., Clément, G., Armengaud, P., Cañas, R., Maranas, C. D., Lea, P. J. and Hirel, B. (2014A). Nitrogen-use efficiency in maize (Zea mays L.): from 'omics' studies to metabolic modelling, *Journal of Experimental Botany* **65**(19): 5657–5671.

Titre : Modélisation mathématique et intégration de données biologiques complexes : analyse du phénomène d'hétérosis chez la levure

Mots clés : Vigueur hybride, consanguinité, dispositif diallèle, intégration de données, métabolisme, levure

Résumé : Le cadre général de cette thèse est la question de la relation génotype-phénotype, abordée à travers l'analyse du phénomène d'hétérosis chez la levure, dans une approche associant biologie, mathématiques et statistiques. Antérieurement à ce travail, un très gros jeu de données hétérogènes, correspondant à différents niveaux d'organisation (protéomique, caractères de fermentation et traits d'histoire de vie), avait été recueilli sur un dispositif demi-diallèle entre 11 souches appartenant à deux espèces. Ce type de données est idéalement adapté pour la modélisation multi-échelle et pour tester des modèles de prédiction de la variation de phénotypes intégrés à partir de caractères protéiques et métaboliques (flux), tout en tenant compte des structures de dépendance entre variables et entre observations. J'ai d'abord décomposé, pour chaque caractère, la variance génétique totale en variances des effets additifs, de consanguinité et d'hétérosis, et j'ai montré que la distribution de ces composantes permettait de définir des groupes bien tranchés de protéines dans lesquels se plaçaient la plupart des caractères de fermentation et de traits d'histoire de vie. Au sein de ces groupes, les corrélations entre les variances des effets d'hétérosis et de consanguinité pouvaient être positives, négatives ou nulles, ce qui a constitué la première mise en évidence expérimentale d'un découplage possible entre les deux phénomènes. Le second volet de la thèse a consisté à interfacer les données de protéomique quantitative avec un modèle stoichiométrique du métabolisme carboné central de la levure, en utilisant une approche de modélisation à base de contraintes. M'appuyant sur un algorithme récent, j'ai cherché, dans l'espace des solutions possibles, celle qui minimisait la distance entre le vecteur de flux et le vecteur des abondances observées des protéines. J'ai ainsi pu prédire un ensemble de flux et comparer les patrons de corrélations entre caractères à plusieurs niveaux d'intégration. Les données révèlent deux grandes familles de caractères de fermentation ou de traits d'histoire de vie dont l'interprétation biochimique est cohérente en termes de trade-off, et qui n'avaient pas été mises en évidence à partir des seules données de protéomique quantitative. L'ensemble de mes travaux permettent de mieux comprendre l'évolution de la relation entre génotype et phénotype

Title : Mathematical modelling and integration of complex biological data: analysis of the heterosis phenomenon in yeast

Keywords : Hybrid vigor, inbreeding, diallel design, constraint-based model, metabolism, yeast

Abstract : The general framework of this thesis is the issue of the genotype-phenotype relationship, through the analysis of the heterosis phenomenon in yeast, in an approach combining biology, mathematics and statistics. Prior to this work, a very large set of heterogeneous data, corresponding to different levels of organization (proteomics, fermentation and life history traits), had been collected on a semi-diallel design involving 11 strains belonging to two species. This type of data is ideally suited for multi-scale modelling and for testing models for predicting the variation of integrated phenotypes from protein and metabolic (flux) traits, taking into account dependence patterns between variables and between observations. I first decomposed, for each trait, the total genetic variance into variances of additive, inbreeding and heterosis effects, and showed that the distribution of these components made it possible to define well-defined groups of proteins in which most of the characters of fermentation and life history traits took place. Within these groups, the correlations between the variances of heterosis and inbreeding effects could be positive, negative or null, which was the first experimental demonstration of a possible decoupling between the two phenomena. The second part of the thesis consisted of interfacing quantitative proteomic data with the yeast genome-scale metabolic model using a constraint-based modelling approach. Using a recent algorithm, I looked, in the space of possible solutions, for the one that minimized the distance between the flux vector and the vector of the observed abundances of proteins. I was able to predict unobserved fluxes, and to compare correlation patterns at different integration levels. Data allowed to distinguish between two major types of fermentation or life history traits whose biochemical interpretation is consistent in terms of trade-off, and which had not been highlighted from quantitative proteomic data alone. Altogether, my thesis work allow for a better understanding of the evolution of the genotype-phenotype map.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France