



UNIVERSITY OF KIEL

## **Dissertation**

in fulfillment of the requirements for the degree

*Doctor rerum naturalium*

of the Faculty of Mathematics and Natural Sciences

at the Christian Albrechts University of Kiel

# **DARWIN THROWS DICE: MODELLING STOCHASTIC PROCESSES OF MOLECULAR EVOLUTION**

---

Submitted by

**Gustavo Valadares Barroso**

from the Group Molecular Systems Evolution

at the Max Planck Institute for Evolutionary Biology



Max-Planck-Institut für  
Evolutionsbiologie, Plön

Kiel, February 2019

**First referee:** Dr. Julien Yann Dutheil (MPI for Evolutionary Biology)

**Second referee:** Prof. Dr. Tal Dagan (University of Kiel)

**External examiner:** Prof. Dr. Ellen Baake (University of Bielefeld)

**Additional examiner:** Prof. Dr. Diethard Tautz (University of Kiel)

**Chairperson:** Prof. Dr. Matthias Leippe (University of Kiel)

Date of oral examination: 29/04/2019

Approved for publication: 29/04/2019

In memory of my beloved Grandfather

**Walter J. Wiele**

02 September 1933 – 25 November 2018





## ZUSAMMENFASSUNG

Die Verfügbarkeit von Protein- und DNA-Sequenzen hat seit Mitte des zwanzigsten Jahrhunderts die Evolutionsbiologie revolutioniert. Erstmals war es möglich die genetische Variation zwischen Individuen auf molekularer Ebene zu beschreiben und sogar zu quantifizieren. Diese neue Art von Daten trafen auf eine immense Vielzahl von Theorien zur Evolution die über die Jahrhunderte entstanden waren. Sewall Wright, Ronald A. Fisher und John B. S. Haldane, den Pionieren auf dem Gebiet der Populationsgenetik, gelang es erstmals die Evolution von Genen in idealen Populationen zu modellieren.

Es ist nun möglich ihre theoretischen Vorhersagen zu überprüfen und unter Verwendung molekularer Daten die ursprüngliche Populationsgröße und die natürliche Selektion zu begreifen.

In einer Zeit in der es möglich ist vollständige Genome zu sequenzieren stellt sich die Überprüfung der theoretischen Annahmen zunächst als Herausforderung heraus. Die Zufälligkeit, mit der unterschiedliche evolutionäre Prozesse auf die DNA-Sequenzen einwirken, macht es schwer die eigentlichen Signale von Hintergrundrauschen abzugrenzen. Auch wenn bereits beträchtliche Fortschritte auf diesem Gebiet erzielt wurden, stehen nur wenige verfügbare evolutionäre Modelle den immensen Mengen von verfügbaren Sequenzdaten gegenüber.

In dieser Thesis mache ich einen Schritt um diese Lücke zu verkleinern. Meine wichtigste Entwicklung ist die der integrierten sequenziellen Markov Koalenz (iSMC) – ein neuartiges System, das gleichzeitig die Effekte der Demographie und der molekularen Heterogenität der genetischen Diversität moduliert. Dieses Verfahren ermöglicht es realitätsgetreuere Modelle der Populationsgenetik zu erstellen als zuvor.

Das Schicksal der DNA über mehrere Generationen hinweg wird durch stochastische Prozesse innerhalb der Zellen geprägt. Besonders hervorzuheben ist das unvorhersehbare Verhalten der Moleküle während der meiotischen Rekombination zwischen homologen Chromosomen sowie die Entstehung von Fehlern während der DNA-Replikation (Mutation). Beide Mechanismen werden durch komplexe Dynamiken gesteuert.

Weiterhin treten Rekombination und Mutationen in bestimmten Regionen des Genoms häufiger auf als in anderen. Obwohl solche Heterogenität einen wichtigen Einfluss auf die Evolution hat, wurde sie in bisherigen Populations-Genetischen-Modellen vernachlässigt. Hier zeige ich, dass die Einbeziehung dieser Heterogenität in die Koaleszenz es ermöglicht die genomweite Variation der Rekombinationsrate (Kapitel 1) und Mutationsrate (Kapitel 2) zu inferieren. Auf diese Weise bringt das vorgestellte Modell die Populationsgenetik der tatsächlichen Biologie von Genomen einen Schritt näher.

Die Auswirkungen intrazellulärer stochastischer Prozesse erstreckt sich deutlich über den Einfluss auf die DNA-Sequenz hinaus. Aufgrund der zufälligen Diffusion verschiedener Moleküle können sich isogene Zellen auch in homogener Umgebung stark in ihren Expressionsmustern und somit Phänotypen unterscheiden. Um Chaos zu vermeiden ist es notwendig die intrazellulären stochastischen Prozesse durch natürliche Selektion zu dämpfen. Im dritten Kapitel verwende ich Daten aus Einzelzell- Transkriptomanalysen um zu entschlüsseln, welche Faktoren das Hintergrundrauschen der Genexpression verringern. Obwohl die Selektion gegen gesteigertes Hintergrundrauschen auf unterschiedlichen Organisationsniveaus wirkt, zeige ich, dass sie vor allem durch die Architektur molekularer Netzwerke beeinflusst wird. Dieses verändern unser Verständnis der Genotyp-Phänotyp-Fitness Interaktionen grundlegend.

## SUMMARY

The availability of protein and DNA sequences in the second half of the 20<sup>th</sup> century revolutionised evolutionary biology. For the first time, it was possible to quantify genetic variation among individuals at the molecular level. These data immediately met a large body of theory that had been accumulated in the previous decades. Pioneered by Sewall Wright, Ronald A. Fisher and John B. S. Haldane, the field of Population Genetics had been modelling the evolution of genes within idealised populations. Now, their theoretical predictions could finally be confronted. Using molecular data to understand past demography and natural selection became an attainable goal.

In the current era of whole-genome sequences, application of these early theoretical results proved to be challenging. The stochastic nature of evolutionary processes acting on DNA sequences makes it hard to distinguish signal from noise. Although progress has been made in this direction, models of molecular evolution are still lagging behind the huge availability of sequence data. In this thesis I contribute to bridging this gap, even if slightly. My main result is the development of the integrated sequentially Markovian coalescent (iSMC) – a novel framework that jointly models the effects of ancestral demography and molecular heterogeneity in shaping genetic diversity. This principled approach represents a step towards more realistic models of Population Genetics.

The fate of DNA over generations is driven by stochastic processes inside the cell. Of particular relevance here is that the erratic behaviour of molecules results in both chromosomal recombination during meiosis and copy errors during DNA replication (mutation). Both mechanisms exhibit complex dynamics, and some regions of the genome are more likely to experience recombination or mutation events than others. Although such heterogeneity impacts evolution, it has largely been neglected by Population Genetic models. Here I show that its incorporation into the Coalescent leads to accurate inference of spatial variation in the recombination rate (chapter 1) and the mutation rate (chapter 2). The ensuing model brings Population Genetics closer to the biology of genomes.

The consequences of intracellular stochasticity extend beyond DNA sequences, however. Due to randomness in the diffusion of key molecules, isogenic cells differ in their gene expression patterns – hence in their phenotypes – even in homogeneous environments. To avoid chaos, intracellular stochasticity must be tamed by natural selection. In the third chapter, I leverage single-cell transcriptomics data to disentangle the factors that constrain gene expression noise. Although selection against elevated noise acts at different levels of organisation, I show that it responds primarily to the architecture of molecular networks. This result may impact our understanding of the genotype-phenotype-fitness map.

# TABLE OF CONTENTS

<b>I. INTRODUCTION</b>	<b>1</b>
1. The importance of stochasticity to evolution.....	1
2. The nature of stochastic processes in evolutionary biology.....	3
2.1 Mutations and life-histories.....	3
2.2 Stochastic changes in environmental conditions.....	12
<b>II. CONCLUDING REMARKS</b>	<b>15</b>
<b>III. CHAPTER 1</b>	<b>17</b>
1. Introduction.....	18
2. Results.....	19
3. Discussion.....	27
4. Methods.....	29
<b>IV. CHAPTER 2</b>	<b>39</b>
1. Introduction.....	40
2. Results.....	41
3. Discussion.....	51
4. Methods.....	53
<b>V. CHAPTER 3</b>	<b>57</b>
1. Introduction.....	58
2. Results.....	60
3. Discussion.....	77
4. Conclusion.....	80
5. Methods.....	80

<b>VI. REFERENCES</b>	<b>88</b>
<b>VII. APPENDIX 1</b>	<b>107</b>
<b>VIII. APPENDIX 2</b>	<b>108</b>
<b>IX. APPENDIX 3</b>	<b>133</b>
<b>X. LIST OF FIGURES</b>	<b>151</b>
<b>XI. AUTHOR CONTRIBUTIONS</b>	<b>152</b>
<b>XII. ACKNOWLEDGEMENTS</b>	<b>153</b>
<b>XIII. AFFIDAVIT</b>	<b>155</b>
<b>XIV. CURRICULUM VITAE</b>	<b>156</b>

*“Science is more than a body of knowledge. It's a way of thinking.”*

– Carl Sagan





# INTRODUCTION

Whether the natural world is fundamentally deterministic or stochastic remains an open question [1,2]. The uncertainty arises because deterministic processes can appear to be random, as in the amplification of minor differences in the initial conditions of dynamical systems known as deterministic chaos [3,4]. This phenomenon happens at all levels of organisation – from the interaction of elementary particles [1] to intracellular reactions [5,6] and eco-evolutionary dynamics between species [7,8] – and pertains to the predictability of evolution [9]: would replaying the tape of life lead to the same outcome [10,11]? According to deterministic chaos, the answer would be yes, although full prediction of the events unfolding would require staggeringly complex models as well as infinite precision in their variables [1]. As a result, the distinction between deterministic chaos and “true” stochasticity is not only unlikely to be settled, but also empirically trivial: since measurements are bound to finite precision, the universe through the lens of science is inevitably stochastic. In other words, even if we knew the present position of all atoms in the universe, we would neither be able to deduce the future nor make perfect inference about the past. This realisation has direct consequences for the study of evolution. Since the long time-scale where evolution operates precludes direct observation of events, we rely on modelling to understand how ancestral processes shaped life on Earth. In order to account for a meaningful proportion of the possible evolutionary trajectories, our models must be stochastic. In this thesis, I will describe novel stochastic models of molecular evolution.

## 1. The importance of stochasticity to evolution

Broadly speaking, evolution can be defined as change in biological diversity over time. From its initial formulation as a purely adaptive process whereby variation is sorted deterministically by means of natural selection, Darwinian evolution embraced the inheritance principles of Mendelian genetics and was reformulated during the Modern Synthesis in the first half of the 20th century [12,13]. At that time, the emerging field of population genetics sought to formally describe the forces that act on genetic variation to effectively promote changes in allele frequencies; in doing so, it provided solid theoretical ground to explain how variation at the population level is eventually transformed into the phenotypic variation at the species level that had inspired Darwin [12]. Early theoretical results showed that the mechanisms of evolution are plenty. Besides natural selection [14],

stochastic processes also play a role [13]. Attempts to quantify the relative importance of both in affecting allele frequencies have established a long standing debate in the field [15,16].

Throughout the development of population genetics theory in the 1940's and 50's, most believed that natural selection was the dominant force and that little genetic diversity would be present in natural populations [12,17]. As molecular data became available [18], it was clear that variation is ubiquitous [19]. The pan-selectionist view where the majority of mutations has substantial fitness effects was not able to explain the observed levels of molecular polymorphism due to the predicted high cost associated with segregation of non-optimal variants [17]. To solve the apparent paradox, Kimura posited that most, if not all, such variation would behave neutrally [20]. In this scenario, strongly deleterious mutations are quickly purged away and are not observed in the data; conversely, strongly advantageous mutations are very rare and quickly rise to fixation; the vast majority of single nucleotide polymorphisms (SNPs) segregating at the population level are invisible to selection and experience a random walk towards either loss or fixation [12]. Subsequent comparative studies among different species largely provided evidence in favor of the Neutral Theory. The observation that divergence tends to accumulate linearly with time formed the basis for the molecular clock hypothesis [21], which states that the branches of a phylogenetic tree can be dated if one has knowledge about the rate at which mutations happen per nucleotide per unit of time. This model was later refined to incorporate a distribution of fitness effects that allows for a small fraction of mutations being slightly deleterious and an even smaller fraction being slightly advantageous – the Nearly Neutral Theory of molecular evolution [22]. Recently, however, in light of large-scale whole-genome sequencing and powerful inference tools, the old debate has resurfaced [23–25]. As compelling stories of episodic positive selection accumulate, together with an increasing appreciation of the role of linkage and recombination in modulating shared evolutionary histories among sites [26,27], the post-modern neutralist-selectionist debate focuses mostly around the indirect effect of selection on genome-wide diversity. It is not the purpose of this thesis to settle this debate. Rather, I will argue that stochasticity pervades molecular evolution regardless of the magnitude of genetic drift; other key processes such as mutation, recombination, migration and gene expression are deeply rooted in chance.

## 2. The nature of stochastic processes in evolutionary biology

Lenormand, Roze and Rousset (2008) [28] have classified stochastic processes in evolutionary biology in three major groups. First, random errors during DNA replication result in mutations at the nucleotide level, creating variation among individuals. Second, chance events in life histories at the individual level lead to random reproductive output and consequently to “drifting” of molecular variants. Third, unpredictable fluctuations of environmental conditions at the population level constantly re-define the fittest types by changing selective pressures. Each of these types of stochasticity influences evolution in a different manner. In section 1.2.1 I will describe how stochasticity in reproduction leads to the backbone population genetics: a genealogy of the population. Combined with random mutations, it results in shared patterns of polymorphism that can be exploited to study past evolutionary events. In section 1.2.2 I will briefly introduce the potential consequences of random environmental changes, and bring the attention to a fourth type of stochastic process that has so far been overlooked in evolutionary biology: noise in the genotype-phenotype map.

### 2.1 Mutations and life-histories

Stochasticity impacts the life-cycle of individuals. Birth, reproduction and death are events of the highest evolutionary relevance that entail a considerable amount of chance: it is easy to imagine that otherwise fit individuals can accidentally die or fail to find food or mates. Let birth, reproduction and death be combined such that the object of study becomes a random variable that represents the net reproductive output of each individual. The Wright-Fisher model [29] can be used to follow their lines of descent over generations. The model considers a diploid, panmictic population of constant size  $N$  that evolves neutrally. Individuals from generation  $t$  produce a very large pool of gametes and die immediately afterwards, hence establishing non-overlapping generations. Their gametes are then paired at random to give rise to generation  $t + 1$ . Crucially, because the gamete pool is not limiting, the same individual can have its gametes chosen to form offspring multiple times (i.e., individuals from generation  $t + 1$  can be viewed as a random sample *with replacement* of the individuals from generation  $t$ ). Since population sizes are typically large ( $N > 10,000$ ), the number of offspring of each individual can therefore be approximated by a Poisson distribution with mean and variance both equal to 1. That the variance is greater than 0 implies that some

individuals will produce more than one offspring per generation, while others will produce none. Indeed, under the Poisson distribution, the probability that an individual leaves no descendants in the next generation is  $P(X = 0) = e^{-1} \approx 0.37$ , meaning that only about 63% of individuals are expected to successfully reproduce. As the process is iterated over generations, the entire population descends from a small fraction of the population that lived  $t$  generations ago, eventually sharing a most recent common ancestor (MRCA) – an otherwise ordinary individual that accidentally became the only one from its generation to have descendants in present time. Tracing its lines of descent will paint the genealogy of the population. Implicit in the genealogy are birth, reproduction and death, incorporating stochasticity in life histories.

Reproduction in such finite populations has consequences for the evolution of diversity [30,31]. Consider a population of diploid size  $N$  in which there are two allelic types,  $A$  and  $a$ . Our goal is to track the number of these alleles over future generations, thus we further approximate by letting go of individual boundaries and treat it as a population of  $2N$  gene copies that segregate independently. Therefore, instead of looking at the reproductive output of each individual, we are now concerned with the total number of offspring each allelic type collectively leaves at each generation. Let the frequency of alleles  $A$  and  $a$  at generation  $t$  be  $p$  and  $q = 1 - p$ , respectively, and assume there are no mutations transforming  $A$  individuals into type  $a$  and vice-versa. Since under neutrality the probability of successful reproduction is independent of the type, the frequency of individuals carrying allele  $A$  in generation  $t + 1$  is binomially distributed with mean equal to  $p$  and variance given by  $pq / (2N)$  [32]. Since the variance is inversely proportional to  $N$ , the stochastic fluctuation in allele frequencies known as genetic drift will be stronger when the population size is small. In the absence of new mutations, drift will eventually lead to fixation of one of the alleles.

*Backwards-in-time.* An alternative way of looking at the forwards-in-time loss of diversity by drift is to look at the probability ( $I$ ) that two randomly chosen alleles in present time are identical by descent. Since in a diploid population each individual carries two copies of each locus, the probability that two randomly chosen alleles share a parental allele (coalesce) in the previous generation is  $1 / (2N)$  [33]. However, even if they do not coalesce in the immediately previous generation, they may do so in the generation before that. The same

logic extends deeper into the past such that the probability of identity by descent can be written down as a recursion [17]:

$$I_t = 1/(2N) + (1 - 1/(2N)) \times I_{(t-1)} \quad (1)$$

where  $t$  denotes discrete generations steps. The first term on the right-hand side describes the probability that the two alleles coalesce in the immediately previous generation. The second term describes the probability that the two alleles have distinct parents in the previous generation, which are in turn identical by descent with probability  $I_{t-1}$ . If we now turn things around and measure diversity by the heterozygosity coefficient  $H = 1 - I$  (i.e., the probability that two randomly drawn alleles are *not* identical), we can express the above recursion as:

$$1 - H_t = 1/(2N) + (1 - 1/(2N)) \times (1 - H_{(t-1)}) \quad (2)$$

$$H_t = (1 - 1/(2N)) \times H_{(t-1)}$$

reaching the classical result that diversity is lost by a rate of  $1 / (2N)$  per generation [32]. To illustrate the concept of genealogical variance I have so far considered a population of abstract individuals. To bring the model closer to reality, I now assume that they carry (non-recombining) sequences of DNA, thus, each generation is an opportunity for mutations. In this case, identity by descent can only occur if no mutations happened along the branches connecting our focal pair of alleles. Representing the mutation rate *per* generation per locus by  $\mu$ , we have [17]:

$$I_t = (1 - \mu)^2 \times [1/(2N) + (1 - 1/(2N)) \times I_{(t-1)}] \quad (3)$$

As expected, since the factor of  $(1 - \mu)^2$  reduces the probability of identity by descent, mutations oppose the effect of drift. We can finally ask how much diversity is maintained in a population over time (the so-called mutation-drift balance). Such equilibrium state is found by letting  $I_t = I_{t-1} = I$ . Ignoring terms of order  $\mu$  and rearranging as a function of the expected heterozygosity, we find that  $H = 4N\mu / (1 + 4N\mu)$ . (The term  $4N\mu$  is often represented by  $\theta$  and designated the “population-scaled mutation rate”.) Therefore, if we have an estimate of  $\mu$  (e.g. from experimental studies), we can use levels of genetic diversity as a measure of the

population size [34]. This is because the mutation-drift balance is nothing more than a tug-of-war between stochasticity in individual life histories and stochasticity in DNA replication.

*The Coalescent.* The Wright-Fisher model just presented describes the evolution of tens of thousands of individuals for tens of thousands of generations. In many cases, dealing with this entire pedigree is impractical – and also unnecessary since we have seen that the vast majority of these individuals do not contribute to present-day diversity. Instead, we can further develop the backwards-in-time perspective to gain insight on molecular evolution by focusing on the statistical properties of the genealogy of a sample of size two (I illustrate here the concept for  $n = 2$  because the approaches that I develop on chapters 1 and 2 focus on pairs of genomes). Rather than keeping track of the ancestry of our sample through every generation on the way to their MRCA, we can simply ask how long it takes for their MRCA to be found (and denote this time the  $T_{\text{MRCA}}$ ). This framework is known as the Coalescent [33,35]. It was independently derived by Kingman [36], Tajima [37] and Hudson [38], who were motivated to understand the evolution of molecular diversity within populations.

The statistical property we are looking for is the average number of generations that it takes for two randomly sampled alleles to find a common ancestor. We have seen that the probability that they descend from the same allele (coalesce) in the immediately previous generation is  $1 / (2N)$ . Similarly, the probability that they coalesce exactly two generations ago is the product of the probability that they do *not* coalesce in the previous generation ( $1 - 1 / (2N)$ ) and the probability that they coalesce immediately after that ( $1 / (2N)$ ). Generalising this process, the probability that two alleles coalesce exactly  $t$  generations ago is

$(1 - 1/(2N))^{t-1} \times 1/(2N)$  [17]. That is, they must *not* coalesce for  $t - 1$  generations and then immediately do it. The  $T_{\text{MRCA}}$  of our sample is therefore geometrically distributed with mean  $2N$  and variance  $4N^2 - 2N$  [17]. Due to neutrality, the genealogical process is independent of the mutational process generating polymorphism. Thus, simulating a short DNA fragment under the Coalescent can be done in two steps [35]. First, draw a geometrically-distributed random variable to represent the  $T_{\text{MRCA}}$ . Once the  $T_{\text{MRCA}}$  of our sample has been found, we have a binary tree with two branches of the same length ( $L$ ). Second, add mutations on each branch by drawing a Poisson-distributed random variable (representing the number of mutations in the locus) with rate proportional to  $L \times \mu$  [35].

Since mutations on either branch will generate SNPs in our sample, its expected level of polymorphism is  $2 \times 2N \times \mu = \theta$  as in the Wright-Fisher model. Therefore,  $\theta$  obtained from a sample of two haploid individuals (or alternatively, one diploid) is an unbiased estimator of the diversity of the population [34]. Here lies the computational efficiency of the Coalescent: it concerns only the historical mutations that are pertinent to the sampled sequences.

The large genealogical variance in the Coalescent ( $4N^2 - 2N$ ) implies that the genealogy of a single locus is very noisy. This is an unwelcome property if we wish to accurately estimate diversity in a population, because the genealogy of the locus will likely be shorter or taller than the average just by chance. Importantly, genealogical variance reflects stochasticity in the ancestral process and it is independent of the sampling variance (which is relatively small since the probability that the sample has the same MRCA as the entire population increases rapidly with sample size [35]). In other words, the true  $T_{\text{MRCA}}$  of our locus may deviate substantially from its expectation given  $N_e$ , hence our estimate of  $\theta$  will be inaccurate regardless of the sample size. An obvious solution is to sample multiple independent loci (e.g., from different chromosomes) and combine information from their genealogies. In fact, it is now possible to sequence entire genomes. Although this approach offers a wealth of information, statistical analysis of whole genomes bears an additional challenge: the genealogies at neighbouring sites are not independent from each other [39–41]. To account for this effect, we need to incorporate another layer of biological complexity: meiotic recombination [42,43].

*Coalescent with recombination.* The nucleotides in a chromosome are physically linked to each other [44]. Consequently, in the absence of recombination, chromosomes would travel along generations as units, all their sites sharing an identical MRCA. In sexually reproducing species, meiotic recombination prevents these shared histories by introducing breakpoints where chromosomal blocks part ways in the ancestral process [39]. The crossing-over of sister chromatids prior to the formation of gametes shuffles genetic variation within parents and results in offspring that has a re-arrangement of the haplotypes from their grandparents [35]. In coalescence terms, recombination causes chromosomes to have two parents in the previous generation and to be a mosaic of chromosomes from the ancestral population. Recombination is itself a stochastic process and the probability that a cross-over event

happens between two sites is a function of their distance [45]. Therefore, loci in the same chromosome that are sufficiently apart from each other have almost independent genealogies. As their physical proximity increases, so does the overlap between their evolutionary histories. Even if they do not have the same MRCA, the genealogies might share a sizeable proportion of their branches because a cross-over event that happened between them in the past will only unlink them from that time backwards. This idea of non-independence of genealogies due to physical linkage is referred to as Linkage Disequilibrium (LD) and is central to population genomics [44,46].

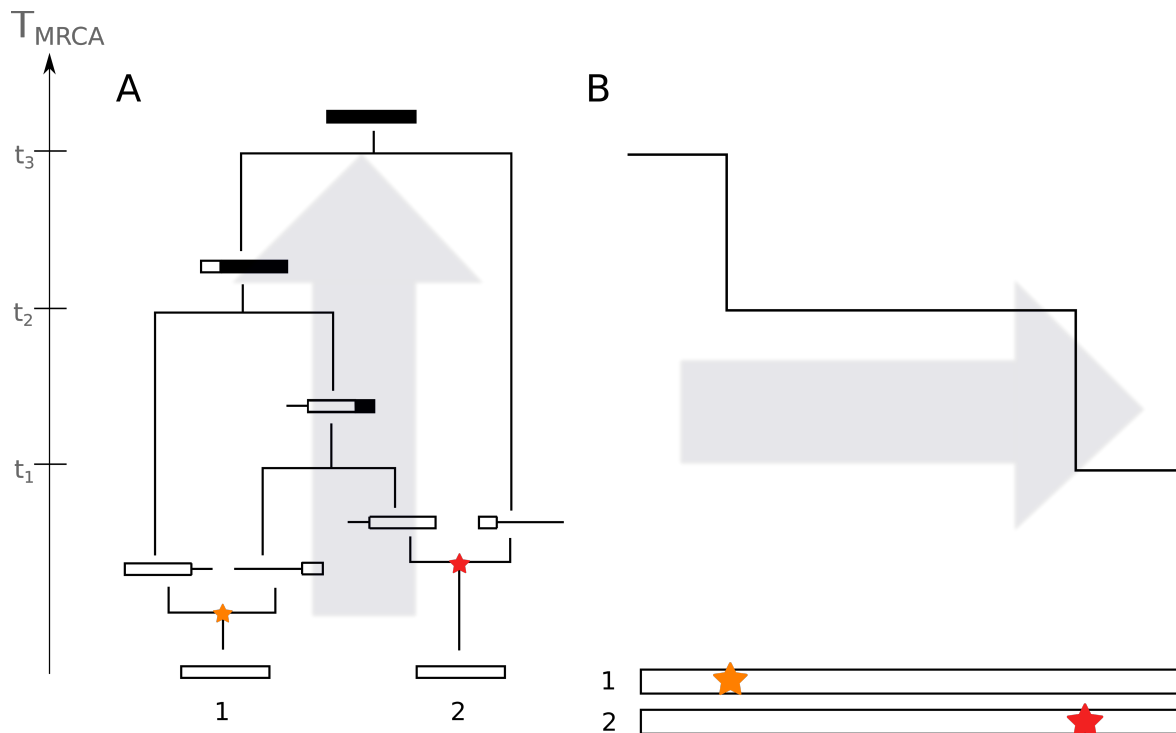
The existence of LD implies that the genealogy under the Coalescent with recombination can no longer be described as a simple bifurcating tree. Instead, it requires a structure that records both coalescence and recombination events. This structure is known as the Ancestral Recombination Graph (ARG) [47,48], and is best introduced by the following thought experiment where we simulate the ancestry of our sample. Once more, we let go of the concept of individuals, and approximate by treating the population as a collection of  $2N$  orthologous chromosomes. Tracing the occurrence of events backwards-in-time, we wait exponentially distributed times for *either* a recombination or a coalescence event. Recombination in a chromosome will split it between two parental chromosomes, whereas coalescence will merge them into a common ancestral chromosome. The process is iterated until all sites find their MRCA (**Figure 1A**). While the backwards-in-time assembly of the ARG offers a complete representation of the ancestral process, its complexity grows with the number of sites in the genome (more precisely, with the number of recombination events) [35]. Since in practice we do not know the true ARG that underlies our sample, we need to integrate over many possible ARGs to have a meaningful representation of the stochastic process. Hence, using the Coalescent framework to analyse whole-genome sequences requires an alternative interpretation of the ARG.

The key simplification of the Coalescent with recombination – which led to the development of important inference tools – happened in two steps, six years apart, and underscores the value of creative work that does not necessarily have an immediate application. In 1999, Carsten Wiuf and Jotun Hein described the Coalescent as a process unfolding spatially along chromosomes [49]. Starting from the  $T_{\text{MRCA}}$  that describes the ancestry of the first site in an



alignment, the spatial Coalescent moves along the chromosome (**Figure 1B**). Due to linkage, neighbouring sites will share their MRCA unless a recombination event has happened between them at some point  $< T_{\text{MRCA}}$ . In this case, one of the chromosomes is split between two parents. While its left side follows the original ancestry path, its right side is now segregating independently in the ancestral population; it will either coalesce back to the original MRCA, or find another one, in which case the  $T_{\text{MRCA}}$  will change at that particular position. The  $T_{\text{MRCA}}$  will only change again at the next position that experiences a recombination event within its time-frame. The spatial process is iterated sequentially until the end of the chromosome, adding recombination events and re-grafting the ARG to accommodate new coalescence events as it proceeds. The result is a distribution of  $T_{\text{MRCA}}$ 's that can be viewed as a serial collection of genealogies depicting the LD structure. Mutations happen in each genomic region with a rate proportional to the local  $T_{\text{MRCA}}$ .

This change in perspective brought the Coalescent with recombination closer to data analysis. The difference lies in how the two approaches resolve the multiple MRCAs scattered across the chromosome (**Figure 1**). In the backwards-in-time assembly of the ARG, a new recombination event is always allowed to happen at any position, and we constrain coalescence times as we move deeper into the past. We thus need to unravel the ancestry of the entire chromosome at once – a daunting task. (So complicated, in fact, that the existing analytical model can only compute the likelihood between two polymorphic sites [50–52].) On the other hand, in the spatial assembly of the ARG, a new coalescence event is always allowed to happen at any time point, and we constrain recombination breakpoints as we move towards the end of the chromosome. Thus we can expand the ARG site-by-site, greatly reducing complexity. Such convenience, however, comes at a cost. Whereas assembly of the ARG is a Markovian process in time (i.e., the next event only depends on the current sample configuration), the same is not true for its spatial counterpart. Due to the presence of so-called trapped non-ancestral material (see **Appendix 2**), the probability of transitioning to a new  $T_{\text{MRCA}}$  does not depend only on the current  $T_{\text{MRCA}}$ , but potentially on all others previously visited. In 2005, Gil McVean proposed that discarding such rare events would be a robust approximation, therefore getting rid of long-range correlations and rendering the process Markovian. Such approximation was dubbed the Sequentially Markovian Coalescent (SMC) [53,54]. The stage was set for Coalescent inference using whole genomes.



**Figure 1. A very simple Ancestral Recombination Graph.** **A**, the backwards-in-time assembly of the ARG for a pair of orthologous chromosomes. Two recombination events (stars) happen early in the ancestry of the sample. Thick bars represent nucleotides that are found in the present-day sample (ancestral material) whereas lines represent variation that is lost by drift forwards-in-time (non-ancestral material). Coalescence events are marked at  $t_1$ ,  $t_2$  and  $t_3$ , where black shading indicates merged ancestral material. **B**, the equivalent representation of the same evolutionary history unfolding spatially along the chromosome.

*Inference using the Coalescent.* We have seen that stochasticity in population genetics comes in two main flavours: random reproductive success of individuals and random occurrence of mutations during DNA replication. The first results in a pedigree of the whole population; the latter creates variation among individuals that on average reflects such genealogical history. To introduce the Coalescent, I took a simulation-like approach and focused on the standard model based on the Wright-Fisher population (notably assuming panmixia, constant size, and neutrality). These models, however, has been extended in multiple directions and can currently incorporate a wide range of evolutionary scenarios [55,56]. The days when the Coalescent served merely as a null model of evolution are gone [38]; it has become a powerful analytical framework where models of increasing complexity are being developed to understand the past [57–59].

We now turn to data analyses. In this setting, we have a sample of DNA sequences and wish

to learn about the evolutionary processes that shaped them. For example, we may be interested in inferring the demographic history of the population to which our sample belongs. Broadly speaking, there are two basic approaches to inference. On the one hand, one can perform model-based simulations with varying parameter values and contrast summary statistics obtained from these simulations with those obtained from the data [60–62]. The simulations where statistics are more similar to data are then selected to paint a posterior distribution of the parameters we wish to estimate (e.g., the growth rate of the population). On the other hand, one can fit an explicit model to data. In this case, our model is a somewhat parsimonious abstraction of reality that describes how its parameters contribute to the data-generating process (evolution). We study how varying parameter values affect the likelihood – the probability of the model, given the observed data – and use optimisation procedures to look for those that best explain the data [63] (e.g., that maximise the likelihood). In chapters 1 and 2, I take this approach to estimate recombination and mutation rates along the genome using Coalescent-based models.

As an exercise of reasoning from observed variables (data) to unobserved variables (parameters), inference is a difficult endeavour in general [64]. Within the context of population genomics, it is further complicated for two reasons. First, the low levels of nucleotide diversity in non-structured populations ( $\theta$  typically ranges from about  $10^{-4}$  to  $10^{-2}$  per site [65]) limits the amount of information that is recorded in sequence data. Second, evolutionary processes such as genetic drift (modulated by demographic history), natural selection and gene flow all affect patterns of polymorphism, but they can leave similar footprints in the data. For this reason, the same pattern can often be equally well explained by multiple scenarios. Disentangling between them is a major goal of statistical models.

A critical challenge with population genomic inference is modelling multiple evolutionary factors simultaneously. Available methods simplify by focusing on the factors they propose to study and assuming all others are negligible. For example, models for demographic inference [57,66] assume neutrality, whereas models for inference of selection typically assume constant population size [67]. The existence of “ghost” factors that substantially influence the data but are unaccounted for can lead to biased estimates. To illustrate why, consider the classic burglars and earthquakes problem [68]. Imagine you have an alarm in your house that

goes off when there is a burglar invasion. The same alarm can also be triggered by unnoticeable earthquakes. While at work, you receive a call by your neighbour saying the alarm is ringing. Your goal is to infer what triggered the alarm. Since earthquakes are fairly rare in your region, you assign higher probability to a burglary event. If, however, you turn on the TV and there is news of a small earthquake, you may decide not to call the police. Now consider what would happen if you did not know that small earthquakes could trigger your alarm. Every time it went off, you would infer that you have been robbed, leading to false positives. The same principle holds for population genomic inference. To better extract signal from the data, models must incorporate the joint effect of multiple evolutionary factors on shaping DNA sequences.

In chapters 1 and 2 I describe iSMC – a novel modelling framework that extends the SMC to incorporate spatial heterogeneity of evolutionary processes along the genome. By jointly inferring demographic histories and variation in recombination and mutation rates, the current implementation of iSMC relieves some of the strong simplifying assumptions made by available inference tools based on the Coalescent [58,59]. As a result, not only can iSMC infer a wider set of parameters than other methods, but its estimates should be more accurate because the joint-inference approach can disentangle the effects that different evolutionary factors leave on sequence data.

## **2.2 Stochastic changes in environmental conditions**

Environmental conditions are dynamic. Whether biotic or abiotic, most environmental variables fluctuate over time. The time-scale of such fluctuations determines how natural selection shapes the way individuals deal with them [28]. On one extreme, cyclic fluctuations that have a duration shorter than the species' generation time (e.g., in human terms, day-night cycles and climate seasons) can be dealt with modulation of gene expression levels. These short environmental changes trigger biochemical reactions which result in activation or deactivation of particular sets of genes. Thus, in the presence of short and cyclic environmental fluctuations, natural selection can favour molecular networks capable to respond when necessary [69].

On the other extreme, stochastic changes that establish new conditions for undetermined

periods of time are more challenging to cope with. Since the new conditions are unexpected, natural selection is unlikely to promote the evolution of a well-adapted molecular network that can be switched on upon stimulation. One possibility is that selection constantly acts on variants better fit to standing environmental conditions. Regardless of whether such recurrent episodes of positive selection act on protein sequences or gene expression levels, if the resulting sweeps are strong, it is possible that alleles favoured under particular conditions rise to fixation. In this case, the next change in the environment can threaten extinction – unless a new beneficial mutation arises, a situation known as evolutionary rescue [70]. An alternative possibility is that selection promotes plasticity in molecular networks such that an individual can explore a range of possible phenotypes [71,72]. For example, a scenario where the environmental conditions change often enough (although unpredictably) to impose a selective pressure can favour the evolution of so-called bet-hedging strategies: when fitness is reduced under “regular” conditions, but increased under stress. One way to achieve this flexibility is by means of Stochastic Gene Expression (SGE).

A fundamental goal of biology is to understand the flow of information from DNA sequences (genotype) to the effect of its encoded proteins on organismal traits (phenotype), and the consequences of trait variation on reproductive output (fitness). Much effort has been placed into deciphering the layers of the so-called genotype-phenotype-fitness map [73]. For example, it is now widely recognised that most traits have complex architectures: their variation is influenced by hundreds of genes, even the entire genome (as described by the omnigenic model [74]). Thus the effect of one gene can be compensated by another, such that models of trait evolution allow the mapping of multiple genotypes onto the same phenotype. So far, however, a standard assumption has been that each genotype maps to a single phenotype. The increasing recognition of gene expression as an inherently stochastic process challenges this view [75–77].

To appreciate the impact of SGE on the genotype-phenotype-fitness map, we must understand an organism as a complex adaptive system. We can describe its architecture by a collection of networks, where nodes represent individual parts and edges represent interactions among them [78,79]. Inside the cell, smaller networks connect to each other if the product of one is a node of the other, thereby establishing a higher level of organisation [80]. Modularity is an

important property here: the ability to isolate the components of a system to minimise external interference, and also to integrate and recombine different modules in order to achieve novel biological function [81]. Within this context, it is tempting to picture the intracellular environment as a perfectly orchestrated clockwork machine. However, the random diffusion of molecules imposes a challenge to faithful execution of biochemical functions. Network edges are not permanent and their establishment is a stochastic process, raising the question of how network nodes are put together exactly when necessary.

A partial explanation is provided by the law of large numbers. Many essential ions and small molecules are present in vast quantities such that stochasticity in their individual movements is averaged out collectively. The law of large numbers, however, does not hold everywhere in the cell [72,76]. In particular, key processes such as gene expression are performed by molecules that are maintained at low copy number relative to their targets. Considering the typical number of genes in an organism (of the order of  $10^4$ ), the low numbers of players such as transcription factors and RNA polymerases implies that stochasticity in their diffusion and binding will not be perfectly compensated [82]. Consequently, the spatio-temporal distribution of proteins is actually heterogeneous. Since their distribution determines which networks are active, stochasticity is an inherent property of cells. This idea was confirmed by an experiment showing that isogenic bacteria behave differently even in homogeneous environments [72].

Although SGE is conceptually well-understood [75,83], its evolutionary consequences have not been thoroughly explored. Intuitively, SGE can be either advantageous or deleterious. In the presence of fluctuating environmental conditions, selection may favour genotypes with noisy expression as a way to explore the phenotypic space without committing to particular DNA variants [84,85]. Thus, SGE can be viewed as a way to cope with an uncertain environment, increasing the probability that at any point at least a fraction of cells is fit. On the other hand, cells must avoid noise propagation across the networks [86,87]: because the distribution of proteins influences gene expression and vice-versa, SGE is self-reinforcing. Therefore, it is expected that selection acts to reduce stochasticity in the expression of core genes, i.e., those that code for highly connected proteins in central networks [88]. In chapter 3, I put SGE in a evolutionary systems biology framework, investigating how natural

selection shapes gene-specific transcriptional noise given the constraints imposed by intracellular networks, and discuss the major determinants of the evolution of SGE.

## **CONCLUDING REMARKS**

A model is an abstract description of reality. It should be simple enough that it can be formulated rigorously and precisely, but also complex enough to capture key components of the process we are trying to understand. The Wright-Fisher is a mathematical model of life histories in a population where individuals do not feed or flee from predators. Still, it laid the foundation of population genetics and fostered the development of more complex models of molecular evolution [56]. Any laboratory experiment is also a model. For example, a set-up to study the dynamics of predator-prey systems ignores a myriad of other species and biochemical molecules that would be present in the wild, as well as fluctuations in environmental conditions. Yet it can answer important questions, which will in turn motivate the development of more realistic experiments [89]. In this process of refinement, where model predictions are confronted with data, we re-evaluate previous assumptions and progressively achieve a better understanding of the world [90].

As biologists dig deeper into extracting signal from sequence data, there is an increasing demand for more realistic models in population genomics. The parsimony principle itself has been challenged in the context of eco-evolutionary models [91]. In the first two chapters of this thesis, I develop the integrated sequentially Markovian Coalescent (iSMC) – which extends the SMC to incorporate heterogeneity in molecular rates along the genome – and demonstrate its accuracy with case studies in three different species. Broadly speaking, the resulting model jointly accounts for the effect of time and space in the Coalescent. By incorporating biological complexity in a principled fashion, it helps bridging the gap between theoretical population genetics (what we know about the evolutionary forces shaping population dynamics) and data analyses (how much of this knowledge is implemented in computational tools that can be used for inference in real datasets). Further, application of the iSMC framework is not restricted to heterogeneity in the recombination rate (chapter 1) or in the mutation rate (chapter 2). Promising avenues of research include modelling spatial variation in the migration rate (as a result of differential permissiveness to gene-flow along

the genome [92]) and in the effective population size (as a result of gene-flow itself or natural selection modulating the number of individuals that contribute to diversity around focal loci [93]). I therefore expect iSMC to have a positive impact on the field in the years to come.

More than an inference tool, iSMC represents a defense of the view of science championed by Judea Pearl [64] and David Deutsch [90]. Its approach to extract biological signal from DNA sequences using conceptualised relations of how evolution (as represented by model parameters) is expected to influence polymorphism data contrasts with the instrumentalist view of so-called data-mining techniques. While I agree that “black box” inference is valuable when the study system is too complicated to be formulated concisely, I strongly subscribe to the idea that science is ultimately an exercise of explaining the natural world – which can only be achieved by a patient quest for causality.

Finally, I show in chapter 3 how selection at different levels of organisation can drive the evolution of SGE. Understanding this phenomenon may influence the way we model the genotype-phenotype-fitness map. For example, an intriguing question in human evolution is how to concile the extensive phenotypic differences between humans and chimpanzees with the striking similarity in their proteome. Fraser (2013) [94] suggested that selection along the human lineage has primarily acted on gene expression levels, which promoted phenotypic changes while keeping protein structures intact (which presumably are close to their fitness peaks). The results I present on chapter 3 suggest that SGE is an important component of the phenotype and a constant target of selection. With the increasing recognition that the architecture of complex traits [95] as well as the dominance effects of mutations [96] are deeply entrenched into gene expression and molecular networks, incorporating SGE into these models can be an important step towards solving the so-called mystery of the missing heritability [97,98].



# CHAPTER 1

## Inference of recombination maps from a single pair of genomes and its application to ancient samples

**Authors:** Gustavo V. Barroso\*<sup>1</sup>, Natasa Puzovic<sup>1</sup> and Julien Y. Dutheil<sup>1</sup>

**Affiliations:** 1) Max Planck Institute for Evolutionary Biology, Department of Evolutionary Genetics, August-Thienemann-Straße 2 24306 Plön – GERMANY

### **ABSTRACT:**

Understanding the causes and consequences of recombination rate evolution is a fundamental goal in genetics that requires recombination maps from across the tree of life. Since statistical inference of recombination maps typically depends on large samples, research in non-model organisms requires alternative tools. Here we extend the sequentially Markovian coalescent model to jointly infer demography and the variation in recombination along a pair of genomes. Using extensive simulations and sequence data from humans, fruit-flies and a fungal pathogen, we demonstrate that iSMC accurately infers recombination maps under a wide range of scenarios – remarkably, even from a single pair of unphased genomes. We exploit this possibility and reconstruct the recombination maps of ancient hominins. We report that the ancient and modern maps are highly correlated, in a manner that reflects the established phylogeny of Neanderthals, Denisovans and modern human populations.

# 1. INTRODUCTION:

Meiotic recombination is a major driver of the evolution of sexually-reproducing species [99]. The crossing-over of homologous chromosomes creates new haplotypes and breaks down linkage between neighbouring loci, thereby impacting natural selection [100,101] and consequently the genome-wide distribution of diversity [65]. The distribution of such cross-over events is heterogeneous within and among chromosomes [102,103], and commonly referred to as the recombination landscape – a picture of how often genetic variation is shuffled in different parts of the genome. Interestingly, this picture is not static, but instead is an evolving trait that varies between populations [104,105] and species [106]. Moreover, the proximate mechanisms responsible for shaping the recombination landscape vary among *taxa*. For example, among primates (where the *PRDM9* gene is a key player determining the location of so-called recombination hotspots [107]) the landscape is conserved at the mega-base (Mb) scale, but not at the kilo-base (kb) scale [108]. In birds, which lack *PRDM9*, the hotspots are found near transcription start sites in the species that have been studied so far [105,109]. In *Drosophila* (where clear hotspots appear to be absent [110]), inter-specific changes are associated with mei-218 variants [111], a gene involved in the positioning of double-strand breaks [112]. The molecular machinery influencing the distribution of cross-over events is still poorly understood in many other groups, where estimates of the recombination landscape in closely related species are lacking.

Aside from their intrinsic value in genetics, accurate recombination maps are needed to interpret the distribution of diversity along the genome. Since the rate of recombination determines the extent to which linked loci share a common evolutionary history [27], inferring selection [113–115], introgression [114,116] and identifying causal loci in association studies requires knowledge of the degree of linkage between sites [117]. Furthermore, recombination can cause GC-biased gene conversion [118,119], which can mimic the effect of selection [120] or interfere with it [121]. Obtaining recombination maps, however, remains a challenging task. Due to the typically low density of markers, experimental approaches provide broad-scale estimates and are limited in the number of amenable *taxa*. Conversely, population genomic approaches based on coalescent theory [122,123] have proved instrumental in inferring recombination rates from polymorphism data.

Traditionally, population genomic methods infer recombination maps from variation in linkage disequilibrium (LD) between pairs of single nucleotide polymorphisms (SNPs) [124–126]. However, since “LD-based” methods typically require large sample sizes per population (from a dozen haplotypes [127]), their application is restricted to a few model organisms where such sequencing effort could be afforded. Here we introduce a new modeling framework (iSMC) to infer the variation in the recombination rate along the genome (as reflected by cross-over events), using a single pair of unphased genomes. Using simulations, we show that iSMC is able to accurately recover the recombination landscape under diverse scenarios. We further demonstrate its efficacy with case studies in Humans, Fruit-flies and the fungal pathogen *Zyoseptoria tritici*, where experimental genetic maps are available. Finally, we exploit our new method to investigate the recombination landscape of ancient hominin samples: Ust’Ishim, the Vindija Neandertal, the Altai Neandertal and the Denisovan. Because it allows inference from datasets for which sample size is intrinsically limited, such as ancient DNA samples, our method opens a new window in the study of the recombination landscape evolution.

## 2. RESULTS:

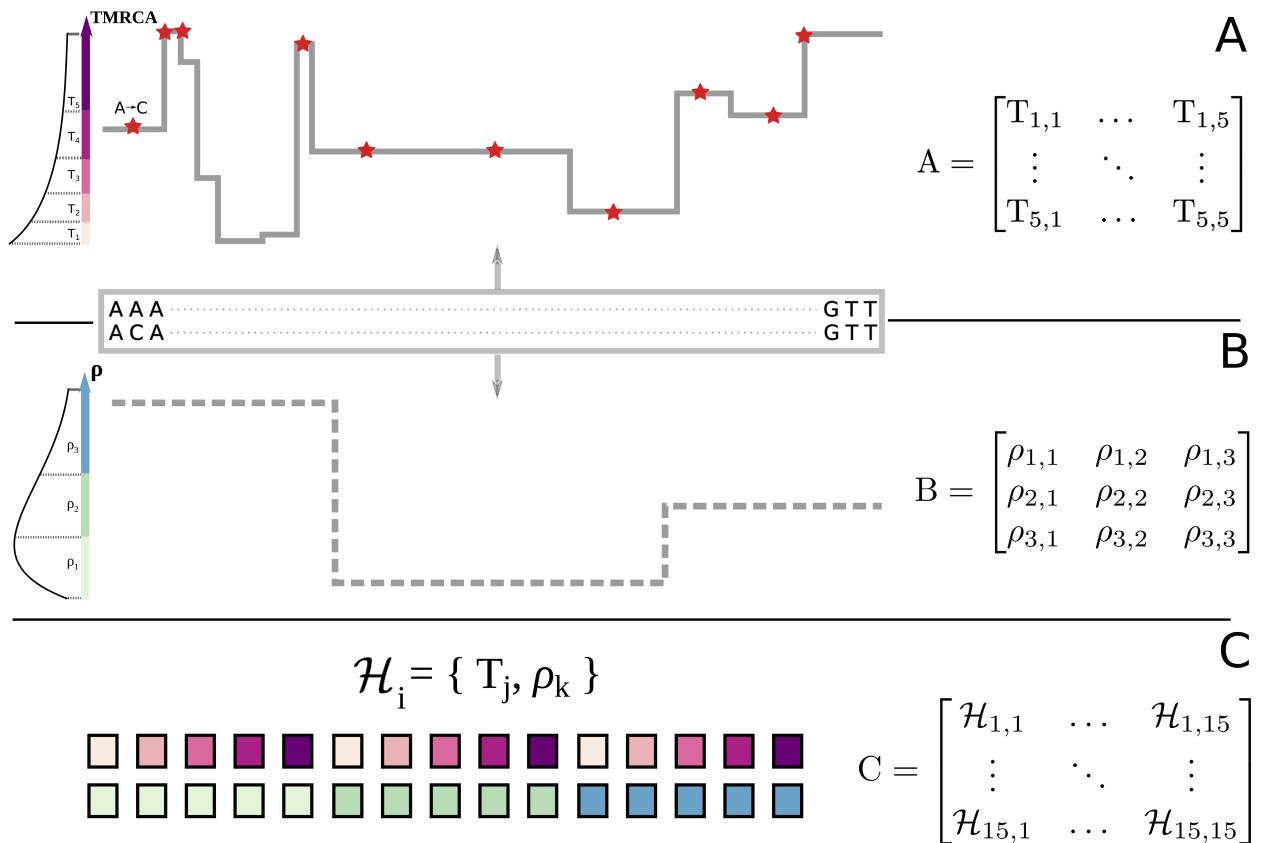
### Overview of iSMC

Besides its common interpretation as a backwards-in-time process, the coalescent with recombination [41,128] can also be modelled as unfolding spatially along chromosomes [49]. Starting from a genealogy at the first position of the alignment, the process moves along the chromosome sequence, adding recombination and coalescence events to the ensuing ancestral recombination graph (ARG) [47,48] (**Figure 1A**). Due to long-range correlations imposed by rare recombination events that happen outside the ancestry of the sample (in so-called trapped non-ancestral material [35]), the genealogy after a recombination event cannot be entirely deduced from the genealogy before, rendering the process non-Markovian. The sequentially Markovian coalescent process (SMC) [53,54] ignores such recombination events, but captures most of the properties of the original coalescent [129] while being computationally tractable. This model is the foundation of recent tools for demographic inference [57,58,130] and has been used to infer the broad-scale recombination map of the human-chimpanzee ancestor based on patterns of incomplete lineage sorting [131,132].

In the SMC, transition probabilities between genealogies are functions of ancestral coalescence rates and – of key relevance to this study – the population recombination rate ( $\rho = 4.Ne.r$ ) [57,58]. Thus, heterogeneous recombination landscapes affect the SMC by modulating the frequency of genealogy transitions: genomic regions with higher recombination rate are expected to harbour relatively more genealogies than regions with smaller recombination rate (**Figure 1**). We leverage this information by extending the SMC to accommodate spatial heterogeneity in  $\rho$  (see Methods). In brief, our new model combines the discretised distribution of times to the most recent common ancestor (TMRCA) of the pairwise SMC [57] with a discretised distribution of  $\rho$  to jointly model their variation along the genome. Since we model the transition between discretised  $\rho$  categories as a spatially Markovian process along the genome, combining the SMC with the Markov model of recombination variation leads to a Markov-modulated Markov model. We cast it as a hidden Markov model [133,134] (HMM) to generate a likelihood function, where the observed states are orthologous nucleotides and the hidden states are {TMRCA,  $\rho$ -category} pairs (**Figure 1C**). We name our new approach “integrative sequentially Markov coalescent (iSMC)”, as it enables jointly capturing the effect of time and space in the Coalescent. This framework explicitly connects the genealogical process with the classical definition of LD as the non-random association of alleles at different loci [44], which has been formulated in terms of covariances in coalescence times [135]. Henceforth, we restrict the use of the term LD to its “topological” interpretation [136].

The SMC is a neutral model where time is re-scaled and measured in units of the effective population size ( $Ne$ ). Thus, information about the recombination rate is obtained in the form of the compound parameter  $\rho = 4.Ne.r$ . Since under neutrality and panmixia  $Ne$  is constant along autosomes, we use the inferred  $\rho$  landscape as a proxy for the spatial variation in the molecular rate  $r$ . (Importantly, local variation in TMRCA primarily reflects genealogical and sampling variance and cannot, on its own, be used to tease apart  $Ne$  and  $r$ .) Our approach is to model spatial variation in  $r$  using a single discrete distribution (**Figure 1B**), which can be accommodated to various models of recombination rate variation (see Methods). After fitting the alternative distributions to sequence data, Akaike's Information Criterion (AIC) [137] is employed as a mean of model selection. If AIC favours a spatially heterogeneous model over the null model where  $\rho$  is constant along the genome, iSMC then estimates a recombination

landscape of single-nucleotide resolution by weighting the discretised values of the favoured distribution of  $\rho$  with their local posterior probabilities. In the following section, we benchmark our model on different simulated scenarios. Therein, we computed the proportion of variance ( $R^2$ ) in simulated maps that is explained by inferred maps after binning the landscapes into windows of 50 kb, 200 kb, 500 kb and 1 Mb.



**Figure 1. Schematic representation of iSMC for one pair of genomes, with five time intervals and three recombination rate categories.** **A**, In the SMC process, the spatial distribution of TMRCA can be described by a matrix of transition probabilities that depend on the population recombination rate  $\rho$  and the ancestral coalescence rates. **B**, variation in  $\rho$  along the genome, modelled as a Markovian process and described by a matrix of transition probabilities. **C**, the combination of both Markovian processes leads to a Markov-modulated Markovian process. The hidden states of the resulting hidden Markov model are all pairwise combinations of discretized classes in **A** and **B**.

### Simulation study

To assess iSMC's overall performance, we simulated five recombination landscapes corresponding to different patterns of magnitude and frequency of change in  $\rho$  and a “null” scenario with constant recombination rate along the genome (see Methods). For each

landscape, we simulated 10 ARGs using SCRM [138], each describing the ancestry of 2 haploid chromosomes. We tested two discretisation schemes for the joint distribution of TMRCAs and recombination rates: the first with 40 time intervals, five  $\rho$  categories; the second with 20 time intervals, 10  $\rho$  categories, leading to a total of 200 hidden states in both configurations. Model selection based on AIC favours the correct model in 45 of the 50 datasets (**Table S1**), with the five exceptions belonging to the scenario where changes are frequent and of small magnitude. In this regime, transitions to regions of slightly different recombination rates do not significantly skew the distribution of genealogies, and the short length of blocks with constant  $\rho$  leaves little signal in the data. Accordingly,  $R^2$  ranges from 8.1% to 70.5% in the five identifiable replicates with frequent changes of low magnitude, and from 60.2% to 98.8% in the other three scenarios (**Figure 2A, Table S2**). Overall, the results are consistent between replicates and robust to the choice of discretisation, although the 40x5 configuration performs better in the scenario with a challenging parameter combination (**Figure 2A**). Therefore, in the following we focus on the 40x5 configuration, noting that it implements a finer discretisation of time that is more adequate to capture the effect of ancestral demography. As we introduce new simulated scenarios, we focus on the recombination landscape with frequent changes of large magnitude.

*Demographic history.* The random sampling of haplotypes during population bottlenecks and expansions affects LD between SNPs, thus creating spurious signals of variation in  $\rho$  [52,139,140]. To test whether iSMC could capture the effect of demography on the inference of recombination maps, we simulated a heterogeneous recombination landscape coupled with either a recent 20-fold increase, or ancient 20-fold decrease in population size. We then fit our model twice for each scenario: first, erroneously assuming a flat demographic history; second, allowing iSMC to infer piecewise constant coalescence rates in order to accommodate population size changes. Overall,  $R^2$  is high (ranging from 46.2% to 91.9%, **Figure 3**), showing that the inferred recombination landscape is relatively robust to misspecification of the demographic scenario, but is systematically higher when demography is jointly inferred (**Figure 2B-C, Table S3**). The difference is stronger at the fine scale, where, in the presence of complex demography the distribution of genealogies can get locally confined to a time period, and ignorance about differential coalescence rates reflects poorly on local  $\rho$  estimates. We conclude that the joint-inference approach of iSMC can disentangle

the signal that variable recombination and fluctuating population sizes leave on the distribution of SNPs.

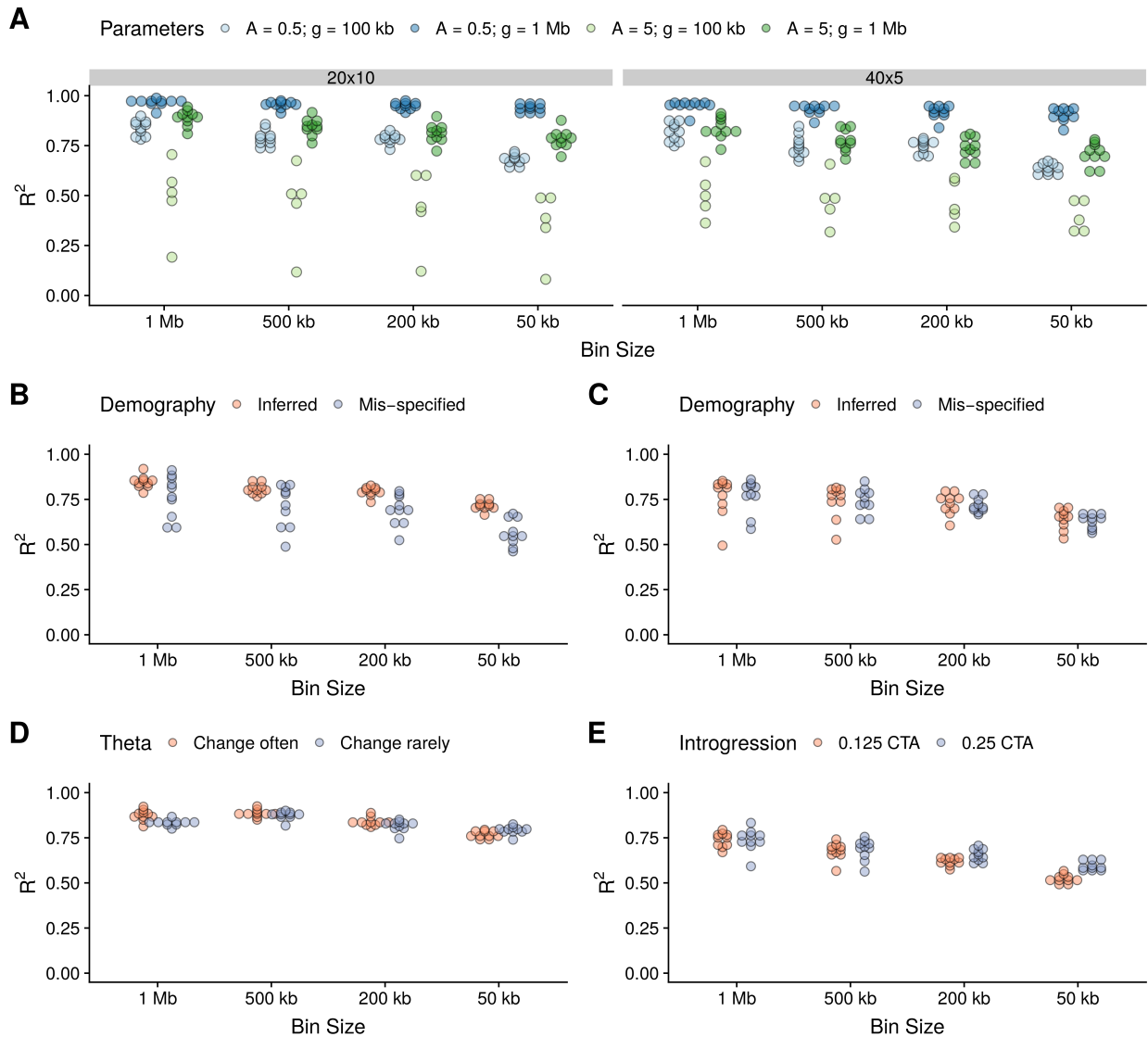
*Introgression events.* Recent studies suggested that introgression is a frequent phenomenon in nature [141,142]. The influx of a subset of chromosomes from a “source” into a “target” population (in a process analogous to a genetic bottleneck) introduces long stretches of SNPs in strong LD. Past introgression events will thus affect runs of homozygosity, biasing the distribution of genealogies. To test the robustness of iSMC to the confounding effect of introgression, we simulated two scenarios of admixture which differ in their time of secondary contact between populations (see Methods). The proportion of variance explained remains high (ranging from 49.1% to 83.3%, **Figure 2D, Table S4**) and depends on the time when introgression occurred. Recombination maps are less accurately recovered in case of recent introgression, because in such case there has been less time for recombination to break SNP associations that do not reflect local  $\rho$  in the target, sampled population.

*Variation in mutation rate.* The rate of *de novo* mutations varies along the genome of many species. For example, CpG di-nucleotides experience an increase in mutation rate ( $\mu$ ) as a result of methylation followed by deamination into thymine, whereas the efficiency of the molecular repair machinery is negatively correlated with the distance from the DNA replication origin, causing  $\mu$  to vary accordingly [143]. Such heterogeneity could bias iSMC's estimates because the transition into a region of higher  $\mu$  mimics the transition to a genealogy with a more ancient common ancestor, since in both cases the outcome is locally increased genetic diversity. To assess the impact of variation of mutation rate on the estimation of recombination rate, we simulated two scenarios of variation of  $\theta = 4.Ne.\mu$  along the genome, corresponding to low and high frequency of change, relative to the frequency of change in the recombination rate. We report that transitions to different mutation rates along the genome globally do not introduce substantial biases in our estimates ( $R^2$  ranges from 73.8% to 92.4%, **Figure 2E, Table S5**).

### **Application to a fungal pathogen, fruit-flies and humans**

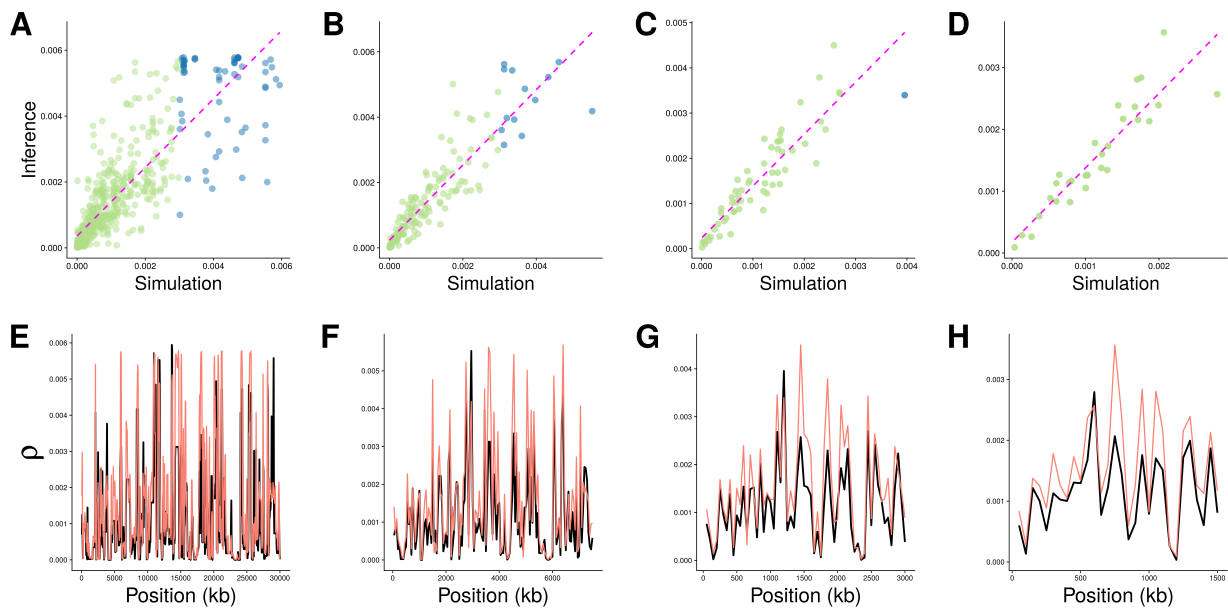
Next, we benchmarked iSMC on model organisms with contrasting genomic architectures and evolutionary histories. We used the proportion of variance in genetic maps that is

explained by corresponding iSMC-inferred maps as a proxy for iSMC's accuracy, noting, however, that these maps are not expected to be perfectly correlated due to evolution of the recombination landscape (see [110] and Discussion). To estimate 95% confidence intervals (CI) for each  $R^2$ , we performed 100 bootstrap replicates of the bins in the recombination maps. In all three species,  $R^2$  values are significant (CI does not include zero) and robust to the choice of model (**Table 1**).



**Figure 2. Recombination map recovery under various simulated scenarios according to bin size.** Dot plots show the distribution of squared Pearson correlation coefficients ( $R^2$ ) between the simulated and inferred recombination maps. **A**, four scenarios of spatial variation in the recombination rate, corresponding to different combinations of parameters (colour), and comparison between two discretisation schemes (panels). **B-C**, comparison between a model where demography is mis-specified and another where it is jointly inferred (colour), in scenarios of recent growth (**B**) or ancient bottleneck (**C**). **D**, two scenarios of spatial variation in the mutation rate, varying its frequency of change (colour). **E**, two scenarios of introgression, varying the time of gene-flow (colour). **Legend:**  $A$  is the shape of the Gamma distribution;  $g$  is the average length of blocks.





**Figure 3. Inference of recombination maps in the presence of recent population growth.** Each column represents a bin size (50 kb, 200 kb, 500 kb and 1 Mb). **A-D**, scatter-plots of inferred versus simulated maps, coloured according to the  $\theta / \rho$  ratio (green  $\geq 1$ , blue  $< 1$ ). Dashed magenta line represents ordinary least squares regression. **E-H**, corresponding simulated (black) and inferred (orange) maps.

The leaf blotch *Zymoseptoria tritici* is a highly polymorphic fungal pathogen with a compact genome (40 Mb) that is under widespread selection [144,145] and exhibits an extremely rugged recombination landscape [146,147]. In this species, AIC favours a heterogeneous model with the presence of recombination hotspots in all three pairs of genomes analysed (**Table 1**, see Methods).  $R^2$  ranges from 24% to 38% at the 20 kb scale and from 27% to 36% at the 100 kb scale. In sharp contrast to *Z. tritici*, the recombination landscape in *Drosophila* is notably smooth [110], and AIC favours a heterogeneous model based on a Gamma distribution (**Table 1**). In this species,  $R^2$  ranges from 14% to 28% at the 100 kb scale and from 44% to 78% at the 1 Mb scale. Like in *Z. tritici*, model fitting in humans favours a heterogeneous distribution of recombination rates with the presence of hotspots (**Table 1**). We inferred recombination maps under this model for each of the three Yoruban (African), three Dai Chinese (Asian) and three Finnish (European) genomes available in the Simons Genome Diversity Project [148], and compared them to the sex-averaged genetic map from DECODE [104]. The proportion of variance in the DECODE map explained by iSMC maps inferred from African individuals (3%, 2% and 2% at the 50 kb scale; 30%, 20% and 20% at the 1 Mb scale) are lower than when  $R^2$  is computed using individual maps from either Asia (6%, 6% and 5%; 46%, 39% and 38%) or Europe (7% 5% and 4%; 39%, 36% and 40%). This is

expected since the DECODE map was estimated from a pedigree study of a non-African population, which has a different present-day distribution of cross-over events than African populations due to evolution of the recombination landscape since their split. Taken together these results show that iSMC can infer recombination maps from species with extremely different recombination profiles.

**Table 1: Performance of iSMC in three distantly related species.** AIC and proportion of variance in genetic maps from each species that is explained by iSMC-inferred maps ( $R^2$  +/- 95% Confidence Interval), according to different models of spatial variation in the recombination rate.

Configuration	Zymoseptoria (chr 1) 20-kb / 100-kb			Drosophila (chr 2L) 100-kb / 1-Mb			European Humans (chr 10) 50-kb / 1-Mb			
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3	
AIC	40x5	573962.4	590963.9	590982.2	<b>2240750.1</b>	<b>2248567.9</b>	<b>2207897.3</b>	1404461.8	1380086.0	1439539.1
	40x(4+1)	573970.1	590899.1	590902.4	2242199.2	2249872.2	2209533.9	1404531.0	1380123.9	1439580.1
	40x2	573953.6	590894.4	590899.1	2242195.2	2249868.2	2209529.7	1404526.1	1380119.8	1439576.3
	100x2	<b>572721.5</b>	<b>589636.2</b>	<b>589695.5</b>	2242103.8	2249770.3	2209444.2	<b>1404297.4</b>	<b>1379833.2</b>	<b>1439380.1</b>
R <sup>2</sup> fine-scale	40x5	0.281 +/- 0.13	0.302 +/- 0.12	0.379 +/- 0.16	0.241 +/- 0.8	0.173 +/- 0.07	0.213 +/- 0.09	0.062 +/- 0.022	0.056 +/- 0.02	0.048 +/- 0.02
	40x(4+1)	0.243 +/- 0.1	0.244 +/- 0.12	0.329 +/- 0.13	0.259 +/- 0.07	0.137 +/- 0.05	0.284 +/- 0.09	0.062 +/- 0.02	0.052 +/- 0.02	0.044 +/- 0.02
	40x2	0.242 +/- 0.12	0.244 +/- 0.12	0.329 +/- 0.13	0.258 +/- 0.08	0.226 +/- 0.08	0.283 +/- 0.08	0.062 +/- 0.022	0.051 +/- 0.02	0.044 +/- 0.02
	100x2	0.261 +/- 0.12	0.245 +/- 0.14	0.325 +/- 0.13	0.26 +/- 0.07	0.208 +/- 0.08	0.282 +/- 0.08	0.066 +/- 0.024	0.05 +/- 0.02	0.044 +/- 0.02
R <sup>2</sup> large-scale	40x5	0.302 +/- 0.19	0.355 +/- 0.18	0.341 +/- 0.20	0.714 +/- 0.27	0.555 +/- 0.19	0.727 +/- 0.22	0.426 +/- 0.13	0.355 +/- 0.13	0.411 +/- 0.13
	40x(4+1)	0.326 +/- 0.20	0.334 +/- 0.19	0.314 +/- 0.19	0.764 +/- 0.16	0.435 +/- 0.18	0.783 +/- 0.17	0.409 +/- 0.13	0.368 +/- 0.12	0.407 +/- 0.15
	40x2	0.326 +/- 0.19	0.334 +/- 0.17	0.314 +/- 0.18	0.764 +/- 0.14	0.637 +/- 0.19	0.782 +/- 0.14	0.41 +/- 0.15	0.368 +/- 0.13	0.407 +/- 0.13
	100x2	0.348 +/- 0.22	0.324 +/- 0.15	0.268 +/- 0.16	0.765 +/- 0.17	0.595 +/- 0.18	0.783 +/- 0.14	0.397 +/- 0.13	0.364 +/- 0.13	0.405 +/- 0.12

## Application to ancient samples

At the fine scale ( $\sim 2$  kb), the location of cross-over events in great apes is strongly influenced by the sequence of the *PRDM9* gene [103,107,149]. Such recombination hotspots tend to erode over time, being replaced somewhere else in the genome with the rise of new *PRDM9* alleles [150,151]. Therefore, recombination maps should become more dissimilar with increasing divergence between populations and species. This hypothesis has been corroborated by two lines of evidence. First, comparisons between recombination maps of extant great ape species show no overlap of hotspots at the fine scale but correlations increase with window size, suggesting that molecular players other than *PRDM9* shape the landscape at the large scale. Second, *in silico* prediction of *PRDM9* binding sites in the Denisovan genome has shown no overlap of hotspots with modern humans [152]. iSMC's unique ability to extract information from single diploids allowed for an alternative test of this hypothesis through the analyses of four ancient samples [153]: the Altai Neanderthal [154], the Vindija Neanderthal [155], the Denisovan [156] and the Ust'Ishim individual [157], a 45,000-year-

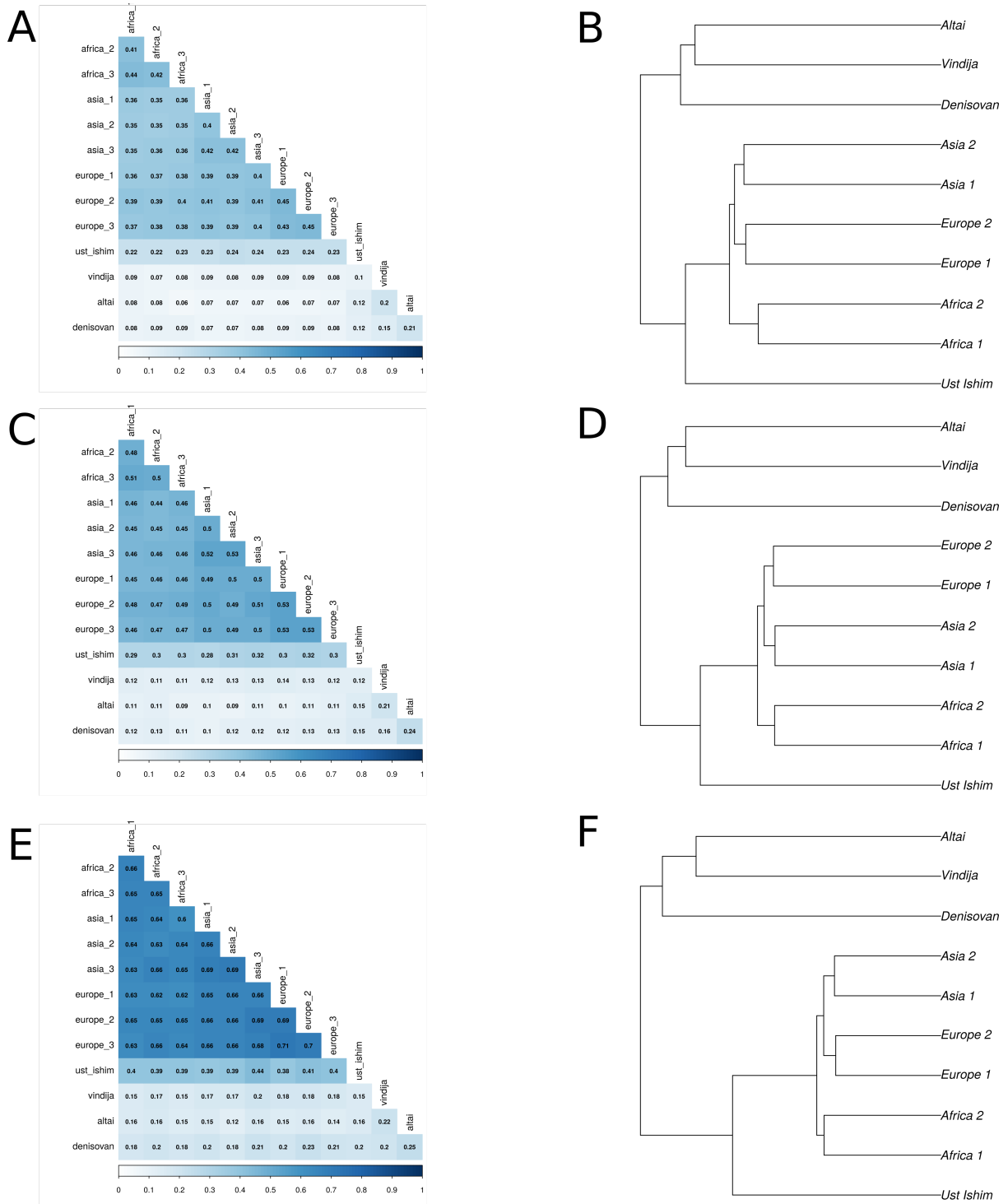
old modern human from Siberia. Since the low density of polymorphic sites in humans makes estimates of  $\rho$  excessively noisy at the 2 kb scale (the scale where *PRMD9* acts), we used 50 kb maps as a proxy for the fine-scale recombination landscape undergoing rapid evolution. Pair-wise correlations between individual maps at this scale reveal that the evolution of the recombination landscape reflects the evolutionary history of hominins: Asians and Europeans form a distinct cluster; the 45,000 year-old Ust’Ishim is sister to the modern humans clade, depicting similarities in the recombination landscape that have been frozen by his demise soon after the out-of-Africa migration; and all modern humans are diverged from the monophyletic Neanderthal-Denisovan group (**Figure 4A, B**). Overall, the topology is consistent at larger scales (**Figure 4D, F**), with a few notable exceptions. First, the differentiation among clades of modern human populations becomes blurrier with increasing window size, as expected due to slower evolution of the recombination landscape at larger scales. Second, the pair-wise correlations within Africans is lower than within non-Africans (**Figure 4A, C, E**). Under panmixia, individual maps from the same population should be highly similar since chromosomes are expected to spend the same number of generations in the different genomic backgrounds during the ancestral process. However, if there is high polymorphism in genes that modulate the position of cross-over events, it is possible that stochasticity in meiotic segregation leads to chromosomes being more often associated with particular alleles – leading to differences among individual recombination maps. Therefore, since African populations carry as much as 50 times more *PRDM9* alleles than non-Africans [158,159], we hypothesise that allelic diversity in trans-acting modifiers of recombination has led to differences in the degree of within-population similarities observed in our correlograms. Third, at the 1 Mb scale, the two Neanderthals (Altai and Vindija) no longer form a monophyletic group; instead, Altai clusters with the Denisovan. While it is plausible that this association is driven by biological signal (both the Altai Neanderthal and the Denisovan come from the same cave in Russia – where a first generation hybrid between the two populations has been recently found [160] – while the Vindija Neanderthal comes from Croatia), it is also possible that their recombination maps are similar enough at this scale that the observed clustering is driven by the lower sequence quality of the Vindija genome. Further investigation is needed to sort out these hypotheses – ideally, as the number of high-quality ancient genomes continues to increase in the next years. Concretely, these results show that 1) iSMC can extract information from ancient genomes, and 2) in hominins, the

divergence of the recombination landscape mirrors the divergence of the species.

### 3. DISCUSSION:

Our analyses show that iSMC is able to infer accurate recombination maps from high-coverage single pairs of genomes. Nevertheless, the proportion of variance explained in experimental maps (**Table 1**) is consistently lower than that obtained in simulations. While this difference can be partly explained by technical noise (e.g., sequencing or SNP calling errors), there are alternative explanations for it. First, biological processes that affect  $N_e$  locally but are unaccounted for by our model will affect LD without reflecting the recombination rate. Among these, introgression and natural selection [59] can introduce a bias if prevalent along the genome. Second, the distinct data types used by experimental and statistical methods imply that they measure different facets of recombination [110]. While experimental maps are a snapshot of the landscape at present-day generation, the historical map estimated by iSMC reflects the time-average cross-over rate at each position of the genome because ancient recombination events also influence the TMRCA distribution. As a result of this contrast between experimental and statistical approaches, the ensuing maps are not expected to be perfectly correlated. Since evolution of the recombination landscape occurs more rapidly at the fine scale, the similarity between experimental and statistical maps should increase with window size, in accordance with our results (**Table 1**). While this observation could be driven by a reduction in estimation noise when maps are averaged within larger windows, in the simulation study – where the recombination landscape is static over time – accuracy increases only slightly with increasing window sizes (**Figure 2**). This suggests that the differences in  $R^2$  between simulation and case studies are not only driven by noise, but also by evolution of the recombination landscape.

Evolution of the recombination landscape implies that the present-day distribution of cross-over events may carry little information about linkage that influenced long-term processes such as linked selection and introgression. Therefore, historical maps are more meaningful than present-day maps in the context of assessing the evolutionary consequences of recombination rate variation [141]. Due to its power with restricted sample sizes, iSMC is well suited to extract LD information from population genomic datasets with high quality whole-genome sequences from a relatively small number of individuals [158–161].



**Figure 4.** Evolution of the recombination landscape in hominins. **Left:** pair-wise Spearman correlations between individual recombination maps (average over 22 chromosomes). Shades of blue indicate the strength of each correlation. **Right:** corresponding dendrograms, estimated using  $1 - \text{correlation}$  as a measure of distance. A-B, 50 kb windows. C-D, 200 kb windows. E-F, 1 Mb windows.

We have demonstrated its accuracy in species with contrasting levels of diversity, demographic histories and selective pressures, and posit that it will be useful for investigation in other species. Not only will such maps aid the interpretation of diversity in non-model organisms, but a picture of the recombination landscape in different groups will tell us about the nature of recombination itself [162]. Open questions include whether the recombination landscape is associated with large-scale genome architecture and how variation in the recombination landscape relates to life history and ecological traits. Finally, as ancient DNA samples become more common (including species other than humans [163]), it will be possible to obtain maps from extinct *taxa*, granting the opportunity to study the evolution of the recombination landscape with unprecedented resolution [131,164].

## 4. METHODS:

### The Markov-modulated Hidden Markov Model framework

SMC models discretise a distribution of coalescence times into  $t$  intervals to implement a discrete space Hidden Markov Model (HMM) with  $t \times t$  transition matrix:

$$\mathbf{Q}(\rho)_{smc} = \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1t} \\ G_{21} & & & G_{2t} \\ \vdots & & & \vdots \\ G_{t1} & G_{t2} & \cdots & G_{tt} \end{bmatrix} \quad (1)$$

where  $G_{ij}$  (the transition probabilities between genealogies  $i$  and  $j$ ) is a function of ancestral coalescence rates and the global parameter  $\rho$ , which is assumed to be constant along the genome [54,58]. The key innovation in iSMC is to relieve this assumption by letting  $\rho$  vary along the genome, following its own Markov process, where values drawn from an a priori distribution are used to compute the transition probabilities between genealogies. Let  $R$  be any strictly positive probability distribution with mean 1.0 describing the variation in recombination along the genome. If  $R$  is discretised into  $k$  categories of equal density, the possible values that  $\rho$  can assume in the Markov-modulated process are all  $r_j * \rho_0$ , where  $r_j$  is the  $j$ th  $R$  category and  $\rho_0$  is the genome-wide average population recombination rate. Our Markov model (inspired by the observation that the distribution of cross-over events is not

random, but clustered in regions of similar values) states that the probability distribution of  $R$  at position  $i + 1$  only depends on the distribution at position  $i$ . We consider the case where the transition probability between any two  $R$  categories ( $P_{ij}$ ) is identical and equivalent to one auto-correlation parameter ( $\delta$ ). The transition matrix of this Markovian process is simply:

$$\mathbf{Q}_\rho = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & & & P_{2k} \\ \vdots & & & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix} = \begin{bmatrix} 1-\delta & \frac{\delta}{k-1} & \cdots & \frac{\delta}{k-1} \\ \frac{\delta}{k-1} & 1-\delta & & \frac{\delta}{k-1} \\ \vdots & & & \vdots \\ \frac{\delta}{k-1} & \frac{\delta}{k-1} & \cdots & 1-\delta \end{bmatrix} \quad (2)$$

Because  $\rho$  is a parameter of the SMC, variation in the recombination rate affects the transition probabilities between genealogies ( $\mathbf{Q}_{SMC}$ ). Since spatial variation in  $\rho$  is modeled as a Markovian process, the combined process is said to be Markov-modulated by  $\rho$ , leading to a Markov-modulated HMM. If  $t$  is the number of discrete genealogies of the SMC, and  $k$  is the number of discretised  $\rho$  categories, then the Markov-modulated HMM is a HMM with  $n = t \times k$  hidden states (**Figure 1**). The transition matrix of the Markov-modulated process,  $\mathbf{Q}_{iSMC}$ , is given by the Kronecker product of  $\mathbf{Q}_\rho$  and  $\mathbf{Q}_{SMC}$ :

$$\mathbf{Q}_{iSMC} = \mathbf{Q}_\rho \otimes \mathbf{Q}_{SMC} = \begin{bmatrix} P_{11} \cdot Q(\rho_1)_{SMC} & \cdots & P_{1k} \cdot Q(\rho_k)_{SMC} \\ \vdots & & \vdots \\ P_{k1} \cdot Q(\rho_1)_{SMC} & \cdots & P_{kk} \cdot Q(\rho_k)_{SMC} \end{bmatrix} \quad (3)$$

In brief,  $\mathbf{Q}_{iSMC}$  is a composition of  $k^2$  sub-matrices of dimension  $t \times t$ , each being a  $\mathbf{Q}_{SMC}$  assembled using  $\rho_0$  scaled by the corresponding category of  $R$ . The main diagonal sub-matrices are further scaled by  $1 - \delta$ , and the off-diagonal sub-matrices by  $\delta / (k - 1)$ .

### Modelling spatial variation in recombination rates

iSMC implements three models of spatial variation in the recombination rate. We first consider a Gamma probability density function with a single parameter ( $\alpha = \beta$ ), which constrains it to have a mean equal to 1.0. After discretisation into  $k$  categories of equal density, the mean value inside each category is drawn to scale  $\rho_0$  during integration over all

recombination rates in the forward recursion (**equation 4**). In our simulation study, since we used a continuous Gamma distribution to draw values of the recombination landscape, we used this model to infer recombination maps. In the second model, we extend the Gamma distribution by adding a category that represents the intensity of the recombination rate in sharp hotspots (parameter  $H$ ). Since hotspots are narrow relative to the extension of the background recombination rate, we use extra parameters to accommodate this effect. As before,  $\delta$  is the transition probability between gamma categories, and we introduce  $w$  as the transition probability from any gamma category to  $H$ , and  $z$  as the transition probability from  $H$  to any gamma category. The third model is a particular case of the second one, obtained by letting the number of discretised categories of the Gamma distribution equal to 1, such that it becomes a probability mass function of only two categories. Importantly, because of the scaling factor provided by the posterior probabilities, reconstruction of the recombination landscape by the posterior average of  $\rho$  allows for a much wider range of values than the sole categories of such discrete distributions. In other words, the posterior average naturally smooths the landscape to fit values that are intermediate between categories, such that even a binary background-hotspot model can infer gradual changes in the landscape.

### **Model selection and computation of the posterior recombination landscape**

iSMC works in two steps: (1) fitting models of recombination rate variation and (2) inferring recombination maps based on the selected model. During step 1, the model parameters are optimized by maximizing the likelihood using the Powell multi-dimensions procedure [165], which is computed for the entire sequence by applying the forward recursion of the HMM [134] as implemented in the zipHMMLib [166] at every position  $i$  of the alignment:

$$F_{i,G_v}(\rho_m) = \left( \sum_{l=1}^k \left( \sum_{u=1}^t F_{i-1,G_u}(\rho_l) \times Pr(G_u \rightarrow G_v | \rho_l) \times Pr(\rho_l \rightarrow \rho_m) \right) \right) \times Pr(G_v \rightarrow S_i) \quad (4)$$

where we integrate over all  $k$  discretized values of  $\rho$  and over all  $t$  TMRCA intervals. The transition between genealogies ( $G_u \rightarrow G_v$ ) is a function of both the focal recombination rate ( $\rho_l$ ) and the ancestral coalescence rates, and  $G_v \rightarrow S_i$  represents the emission probability from  $G_s$  to the observed state at position  $i$ . In case AIC favours one of the heterogeneous models, in step 2 iSMC uses the estimated parameters to estimate the posterior average  $\rho$  for



all sites in the genome. To this end, it first uses the so-called posterior decoding method [134] as implemented in zipHMMlib [166] to compute the posterior probability of every hidden state at each position in the sequence. Since in the Markov-modulated HMM the hidden states are pairs of  $\rho$  categories and TMRCA intervals, this results in joint probability distributions  $P_i(x, y)$  of recombination values  $x$  and coalescence times  $y$ , for all sites  $i$  in the genome. Thus, if  $r_l$  is the value of  $R$  inside discretised category  $l$  and  $\rho_0$  the genome-wide average recombination rate, the posterior average  $\rho$  at position  $i$  is given by

$$(\bar{\rho}_i) = \frac{1}{k} \times \left( \sum_{l=1}^k \left( \sum_{j=1}^t P_i(x_l, y_j) \right) \times r_l \times \rho_0 \right) \quad (5)$$

### Testing hidden state configurations

The hidden states of iSMC are pairs of genealogies and recombination rates where both elements are drawn from discretised distributions. Since the complexity of the forward algorithm (which computes the likelihood) is quadratic in the number of hidden states, there is a limit to the discretisation scheme that can be adopted, as too fine a discretisation would lead to impractical execution times. We set the number of hidden states to 200, and used this limit to run iSMC in all simulated and real datasets. Within this maximum, however, it is possible to devise several combinations of hidden states by changing the way in which we discretise the TMRCA and  $\rho$  distributions. The goal of reconstructing the recombination landscape would in principle make natural the choice of investing in a fine-grained discretisation of the distribution of  $\rho$ . However, this would mean a coarse-grained discretisation of time and, since the signal for fitting the distribution of  $\rho$  comes from the expected number of TMRCA transitions, this strategy could reduce iSMC's power to detect such changes. Therefore, in the simulation study, we tested the performance of two configurations: 20 time intervals x 10  $\rho$  categories and 40 time intervals x 5  $\rho$  categories (Table S2). When fitting the ‘‘Hotspot model’’ to humans, fruit-flies and *Z. tritici*, we tested a configuration with 40 time intervals x 2  $\rho$  categories and another with 100 time intervals x 2  $\rho$  categories.

### Modelling complex demographic histories

The original HMM implementation of the SMC [57] uses the expectation-maximisation

algorithm to optimise transition probabilities, where the actual targets of inference – the coalescence rates at each time interval – are latent variables of the model. Here we use cubic spline interpolation [130] to map coalescence rates at time boundaries, which are then assumed to be piecewise constant for the duration of each interval. Because we use three internal splines knots (i.e., the demographic history is divided into four epochs wherein a cubic curve is fitted), the number of parameters is substantially reduced in our model – in particular when a fine discretisation of TMRCA is employed. Importantly, in the spline implementation, the number of model parameters is independent of the number of classes in the discretization scheme.

### **Computational resources and performance**

The limiting computing resources are different between the optimisation and decoding steps of iSMC. During optimisation, execution time is key: for human chromosome 10, the program uses around 2 Gb of RAM, and runs for about 18 h to fit the hotspot model with 100 time intervals on chromosome 10 from an African individual, using a 2.6 Ghz machine. On the other hand, the limiting resource during computation of the posterior average is memory. On the same 2.6 Ghz machine, it takes around 15 minutes and 20 Gb of RAM to decode a 5 Mb fragment.

### **Simulation study**

*Four scenarios of spatial variation in  $\rho$ .* We simulated a piecewise constant recombination rate along the genome by drawing values from a Gamma distribution with parameters  $\alpha$  and  $\beta$ , and segment lengths from a geometric distribution with mean length  $g$ . We considered four possible scenarios where  $\alpha = \beta = 0.5$  or  $5.0$ , and  $g = 100$  kb or  $1$  Mb. For each of the four combinations, we simulated 10 independent pairs of two 30 Mb haploid chromosomes under a constant population size model, assuming  $\theta = 0.003$  and  $\rho = 0.0012$ . For each of the following simulated scenarios, we focus on the landscape with  $\alpha = 0.5$  and  $g = 100$  kb. All scenarios share the same sequence length, sample size, as well as  $\theta$  and  $\rho$  parameter values,.

*Demographic history.* We simulated two demographic scenarios. First, a 20-fold population expansion 0.01 coalescent time units ago; second, a 20-fold population bottleneck 0.5

coalescent time units ago. Assuming an effective population size of 30,000, these coalescent times correspond to 1,200 and 60,000 generations ago, for the expansion and bottleneck events, respectively.

*Introgression events.* We simulated two introgression scenarios where a source population introduces a pulse of genetic material into a target population. In both scenarios, the split between source and target populations happened 2.0 coalescent time units ago, and the source replaces 10% of the genetic pool of the target. In the first scenario, secondary contact happened 0.125 coalescent time units ago; in the second, it happened 0.25 coalescent time units ago. Assuming an effective population size of 30,000, the split between population happened 240,000 generations ago, and the introgression events happened 15,000 and 30,000 generations ago, respectively.

*Variation in the mutation rate.* We simulated a piecewise constant mutation rate along the genome by drawing rate values from a uniform distribution and segment lengths from a geometric distribution with mean length  $f$ , where  $f$  is either 20 kb or 500 kb. The uniform distribution generating scaling factors of  $\theta$  has mean = 5.05 instead of 1.0. In this case, the expected genome-wide average  $\theta = 0.015$ . The reason for that is our focus on the spatial distribution of  $\theta$  itself. If the landscape had mean = 0.003, its highly heterogeneous nature would scale  $\theta$  down to values well below  $\rho$  (0.0012) too often along the 30 Mb sequence. The ensuing loss of signal (due to low SNP density) would result in poorly inferred maps that display low correlations with the simulations, not because of spatial *heterogeneity* in  $\theta$  (local transitions), but instead because the ratio  $\theta / \rho$  would be too low in many windows across the chromosome. In all the above scenarios, the proportion of variance ( $R^2$ ) in simulated maps that is explained by inferred maps was computed after binning the landscapes into non-overlapping windows of 50 kb, 200 kb, 500 kb and 1 Mb, that is, the analysis is agnostic to the true breakpoints of the simulated landscapes.

## **Data analysis**

Model selection followed by inference of recombination maps in the three species studied (**Table 1**) was performed using publicly available sequences (chromosome 2L from haploid pairs ZI161 / ZI170, ZI179 / ZI191 and ZI129 / ZI138 in the *Drosophila* Population

Genomics Project Phase 3 [167]; chromosome 1 from haploid pairs Zt09 / Zt150, Zt154 / Zt155 and Zt05 / Zt07 for *Z. tritici* [146]; chromosome 10 from three Finnish individuals (LP6005442-DNA\_C10, LP6005442-DNA\_D10, LP6005592-DNA\_A02) available in the Simons Genome Diversity Project [148] for humans). In the first two species, gaps and unknown nucleotides in the sequences (in FASTA format) were assigned as missing data, whereas in Humans the available strict mask for the dataset was applied after parsing the VCF files.

iSMC was fitted four times to each pair of genomes of the three species: 1) with 40 discretised time intervals and a model of variation in  $\rho$  based on a Gamma distribution with five discretised categories; 2) with 40 discretised time intervals and a model of variation in  $\rho$  based on an extended Gamma distribution with four discretised categories and an additional “Hotspot” category; 3) with 40 discretised time intervals and a model of variation in  $\rho$  based on a probability mass function of two categories; 4) with 100 discretised time intervals and a model of variation in  $\rho$  based on a probability mass function of two categories. In each case,  $R^2$  was computed as the square of the Pearson correlation coefficient between the resulting recombination landscape and available genetic maps both at the fine scale (100 kb for *Drosophila*, 20 kb for *Z. tritici* and 50 kb for Humans) and at the large scale (1 Mb for *Drosophila*, 100 kb for *Z. tritici* and 1 Mb for Humans). For each  $R^2$ , its 95% Confidence Interval was computed from a distribution obtained by performing 100 bootstrap replicates of the binned recombination maps.

Whole-genome sequence data were used for in-depth analyses of the recombination landscape in the hominin clade. Model fit (based on the “Hotspot” distribution) and inference of recombination maps was performed independently on each chromosome. The individual IDs within the Simons Genome Diversity Project [148] and corresponding population of origin of the six contemporary modern humans are as follows: African (Yoruban): LP6005442-DNA\_A02, LP6005442-DNA\_B02 and SS6004475; Asian (Dai Chinese): LP6005441-DNA\_D04, LP6005443-DNA\_B01 and LP6005592-DNA\_D03; European (Finnish): LP6005442-DNA\_C10, LP6005442-DNA\_D10, LP6005592-DNA\_A02. The available strict mask for the dataset was applied to assign low-quality positions as missing data. The four ancient DNA samples were downloaded from the server at the Max Planck

Institute for Evolutionary Anthropology in Leipzig

(<http://cdna.eva.mpg.de/neandertal/Vindija/VCF/>) in May 2018. Since these are complete VCF files where all callable positions are reported, no mask was used and absent positions were assigned as missing data.

The analyses of modern and ancient datasets were performed considering only positions present in the DECODE genetic map. The correlograms and dendograms presented in **Figure 4** were obtained by hierarchical clustering (using UPGMA) of pair-wise distances computed from  $1 - r_s$ , where  $r_s$  is the Spearman correlation of ranks between two individual recombination maps. Correlation matrices were computed separately for each chromosome after discarding bins with more than 50% missing data in any of the diploid sequences, and the average correlation matrix over all chromosomes was used to compute the pair-wise distances. Recombination maps for all samples are available as a resource in FigShare: [https://figshare.com/projects/Archaic\\_Recombination\\_maps/44354](https://figshare.com/projects/Archaic_Recombination_maps/44354)

### **Acknowledgements**

The authors thank Alice Feurtey, Asger Hobolth, Bernhard Haubold, Eva Stukenbrock, Fabian Klötzl, Kai Zeng, Pier Palamara and Stephan Schiffels for fruitful discussions about this work. JYD acknowledges funding from the Max Planck Society. This work was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft) attributed to JYD, within the priority program (SPP) 1590 “probabilistic structures in evolution”.

### **DECLARATION:**

This chapter is published in the biorxiv preprint server with DOI:  
<https://doi.org/10.1101/452268>

## **SUPPLEMENTAL MATERIAL:**

The following supplemental material is available in the digital archive (**Appendix 1**):

**Table S1:** AIC value for each replicate of each of the simulated landscapes.

**Table S2:**  $R^2$  for each replicate of each parameter of the simulated landscape, according to discretisation scheme.

**Table S3:**  $R^2$  for each replicate of each simulated demographic history, according to whether coalescence rates were jointly-inferred or not.

**Table S4:**  $R^2$  for each replicate of each scenario of introgression.

**Table S5:**  $R^2$  for each replicate of each scenario of mutation rate variation.

## CHAPTER 2

# Inferring the genomic landscape of mutation rates from polymorphism data

**Authors:** Gustavo V. Barroso\*<sup>1</sup> and Julien Y. Dutheil<sup>1</sup>

**Affiliations:** 1) Max Planck Institute for Evolutionary Biology. Department of Evolutionary Genetics. August-Thienemann-Straße 2 24306 Plön – GERMANY

### ABSTRACT:

In sexually-reproducing species, the distribution of diversity along the genome is influenced by variation in (1) branch lengths of genealogies as a result of genetic drift; (2) effective population size as a result of natural selection and gene-flow; (3) recombination rate, via GC-biased gene conversion and modulation of linked selection; and (4) the rate of de novo mutations generating polymorphism. Quantifying the relative importance of these factors in shaping patterns of diversity is a major goal of population genomics, however, spatial variation in the mutation rate has largely been neglected by empirical studies, in part due to its difficult estimation. This is problematic because regions of the genome with differential mutation rates will either mimic or dilute the signal of selection, leading to false positives and negatives in genome-wide scans. Here we present a new statistical model (iSMC) that infers the genomic landscape of mutation rates from polymorphism data while accounting for the effect of demography and recombination rate variation. Our simulation study demonstrates that it has high accuracy in diverse scenarios. We find that spatial variation in the mutation rate is a significant explanatory factor of the distribution of diversity in a fungal pathogen. This result suggests that mutation rate heterogeneity should be more often incorporated in data analysis. Our explicit model of the mutation landscape allows parametric inference from polymorphism data, thus fostering research in species where large-scale sequencing of pedigrees is not feasible.

# 1. INTRODUCTION:

In sexually-reproducing species, levels of genetic diversity vary along the genome [65] according to four major determinants. First, stochasticity in the ancestral process changes the total branch lengths of genealogies in different parts of the genome, causing the number of mutations they undergo to differ proportionally [123]. This variation is further enlarged by demography [57]. Second, natural selection disturbs local genealogies away from their distribution under neutrality [128,168]. Both negative and positive selection shorten branch lengths in the vicinity around the selected locus (linked selection) [27,59,169], while balancing selection has the opposite effect [170]. Third, recombination impacts diversity either neutrally through GC-biased gene conversion [121,171] or indirectly by modulating the extent to which neighbouring loci share their evolutionary histories with selected sites (i.e., the breadth of linked selection) [27]. Finally, spatial variation in the rate of *de novo* mutations ( $\mu$ ) depends on a number of genomic features such as base composition (e.g., CpG dinucleotides) and distance from the DNA replication origin [143]. A fundamental goal in population genomics is to quantify the relative importance of each of these factors in contributing to genome-wide levels of diversity in natural populations [65].

Indeed, a major challenge in population genomics inference is to disentangle the effects of different evolutionary forces shaping DNA sequences. On the one hand, methods that simultaneously model demography and local variation in  $\mu$  and the effective population size ( $N_e$ ) are restricted to unlinked loci [172–174]. While this simplification avoids the confounding effect of linkage disequilibrium, it leaves out most of the genome, where local variation in  $\mu$  and  $N_e$  may be highly relevant. On the other hand, models that incorporate linkage information typically focus on characterising a single process (e.g., *either* demography [58,130] *or* the recombination landscape [124] *or* selection [59]) that alone is expected to explain the observed patterns of polymorphism. This approach is problematic because spatial variation in branch lengths, mutation and recombination rates, as well as  $N_e$ , can leave similar footprints on sequence data [175]. For example, a genomic region with increased diversity could be explained either by a more ancient common ancestor, balancing selection or higher mutation rate, but its level of polymorphism can only inform on the compound parameter  $\theta = 4.N_e.\mu$ . Therefore, neglecting the existence of important factors that play a role in shaping diversity leads to inference that is both biased and statistically



confident (since under-parametrised models have less uncertainty about which parameter is driving the signal). A prominent such “ghost” factor in population genomic inference is the spatial heterogeneity of  $\mu$ . Since its incorporation leads to more accurate inference from unlinked loci [174], it is expected to have a similar effect with whole-genome data.

Obtaining unbiased estimates of the mutation landscape is challenging. Naive observation from sequence data – either through divergence with a closely related species or diversity within populations – is susceptible to the confounding effects outlined above. An alternative approach is large-scale sequencing of family trios to pinpoint *de novo* mutation events in the germ-line by contrasting the sequences of parents and offspring [176,177]. However, not only this method requires huge resources, it also relies on pedigree information that can only be obtained from a handful of species. Nevertheless, a recent study based on human pedigrees suggest that the impact of spatial variation of  $\mu$  on polymorphism data may be greater than previously recognised: up to 46% of the human-chimp divergence, and up to 69% of human diversity, can potentially be explained by variation in *de novo* mutation rates at the 100 kb scale [178]. Bearing these results in mind, spatial variation in  $\mu$  may have been overlooked as an explanation for the genome-wide distribution of diversity [179], and a thorough assessment of its importance in other species is of interest.

We have previously described a modelling framework (iSMC) that jointly infers the demographic history of the sample and variation in the population recombination rate  $\rho = 4.Ne.r$  along the genome [180]. Here we further extend this framework to account for spatial variation of the population mutation rate  $\theta = 4.Ne.\mu$ . Since it explicitly accounts for the distribution of genealogies, this integration allows statistical inference of the mutation landscape along genomes using polymorphism data. We demonstrate via simulations that our model can accurately recover the mutation landscape using a single pair of genomes. Results from the fungal pathogen *Zymoseptoria tritici* show that the rate of *de novo* mutations is an important factor shaping genetic diversity in this species.

## **2. RESULTS:**

### **Overview of the model**

The first implementation of the sequentially Markovian Coalescent (SMC) describes how genealogies change along a diploid genome as a function of ancestral population sizes and the global recombination rate  $\rho$  [57]. Model fitting is achieved by casting the SMC as a hidden Markov model (HMM) [133] and assuming that the probability of observing a heterozygous site is a function of the underlying TMRCA and the global mutation rate  $\theta$  (see Methods). We previously showed how this process can be framed in a more general model called iSMC, where  $\rho$  is allowed to vary along the genome by following its own Markov process that modulates the frequency of transitions of local genealogies [180].

In the general case, iSMC is a HMM where the observed states are the configurations of nucleotides at each orthologous position of the alignment, and the hidden states are  $n$ -tuples storing all possible combinations of genealogies and discretised values of each parameter of interest that is allowed to vary along the genome. If one such parameter affects either the transition or emission probabilities of the HMM, then the parameters that control its degree of variation can be optimised, e.g., by maximum likelihood. In the  $\rho$ -modulated iSMC ( $\rho$ -iSMC) the hidden states are 2-tuples containing pairs of TMRCA intervals and recombination rates. We now allow  $\theta$  to also vary along the genome (**Figure 1**), following its own Markov process, i.e., letting the hidden states be {TMRCA,  $\theta$ -category,  $\rho$ -category} triplets. Crucially, the signal of variation in  $\rho$  and  $\theta$  left on the distribution of SNPs is discernible because their contributions to the likelihood are orthogonal: the recombination and mutation rates affect transition and emission probabilities of the HMM, respectively and exclusively. We also note that information about the mutation rate is obtained in the form of the compound parameter  $\theta = 4.Ne.\mu$ . By assuming neutrality and panmixia (hence homogeneous  $Ne$  along the genome), we use the inferred  $\theta$  landscape as a proxy for the spatial variation in the molecular rate  $\mu$ . Importantly, local variation in TMRCA primarily reflects genealogical and sampling variance and cannot, on its own, be used to tease apart  $Ne$  and  $\mu$ .

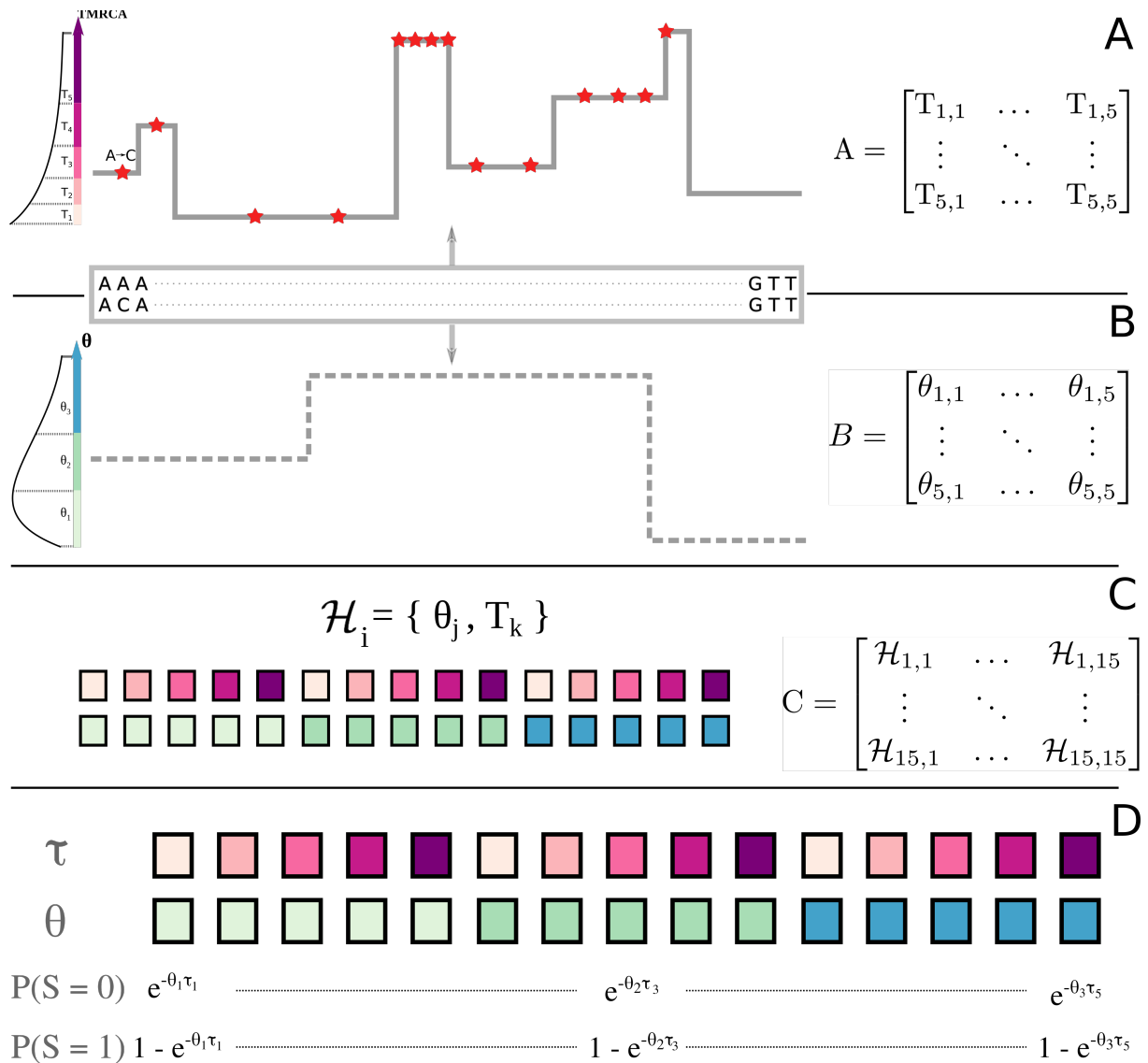
### **Simulation study**

We conducted a simulation study designed to test iSMC's ability to reconstruct the mutation landscape in distinct scenarios (**Figure 2**). We used SCRIM [138] to simulate 10 ancestral recombination graphs (ARG) describing the ancestry of 2 haploid chromosomes. We obtained binary sequences by first simulating  $\mu$  landscapes (see Methods), then placing mutation

events along the branches of the genealogies proportionally to the local mutation rate. The first three sets of simulations share a flat recombination landscape. We could therefore dismiss the distribution of  $\rho$  values and analyse these data by constructing hidden states as 2-tuples with five  $\theta$  categories along with 40 time intervals ( $\theta$ -iSMC). The last simulated dataset concerns the potential confounding impact of a heterogeneous recombination landscape. Therein we configure hidden states as triplets ( $\theta$ - $\rho$ -iSMC). In all cases, the proportion of variance in simulated maps that is explained by inferred maps ( $R^2$ ) was estimated after binning the landscapes in windows of 50 kb, 200 kb, 500 kb and 1 Mb.

*Mutation landscapes.* We first simulated five scenarios of spatial variation in the mutation rate corresponding to four different patterns of magnitude and frequency of change in  $\theta$  as well as a “null” scenario with constant mutation rate along the genome (see Methods). We evaluated iSMC's ability to distinguish between scenarios by fitting it twice to each dataset (**Figure 2**, top): first, assuming that  $\theta$  is constant along the genome (standard SMC), and second, allowing it to vary ( $\theta$ -iSMC). Model selection based on Akaike's Information Criterion (AIC) shows that iSMC correctly favours the heterogeneous model in all replicates of all scenarios where there is indeed spatial heterogeneity in the mutation rate (**Table S1**). On the other hand, it selects a heterogeneous model in five out of 10 replicates when the mutation rate is actually constant along the genome. The inferred mutation landscapes, however, are flat in all these five cases, meaning that false positives in model selection will not lead to spurious identification of spatial heterogeneity in  $\theta$ . This indicates that iSMC can distinguish between local variation in polymorphism that results from genealogical variance and that which results from variation in the mutation rate.

Next, we assessed iSMC's accuracy in reconstructing each of the four heterogeneous landscapes. We report high  $R^2$  for all scenarios (**Figure 3A**, **Table S2**), ranging from 47.1% to 99.6%. The good performance in the scenario with frequent changes of low magnitude in  $\theta$  may seem counter-intuitive because subtle differences in  $\theta$  that do not span long segments could have their effect confounded by the distribution of genealogies itself. However, iSMC is able to distinguish between the effects that mildly heterogeneous mutation rates and the distribution of genealogies leave on sequence data.



**Figure 1. Schematic representation of  $\theta$ -iSMC for one pair of genomes, with five time intervals and three mutation rate categories.** **A**, In the SMC process, the spatial distribution of TMRCA can be described by a matrix of transition probabilities that depend on the population recombination rate  $\rho$  and the ancestral coalescence rates. **B**, variation in the mutation rate  $\theta$  along the genome, modelled as a Markovian process and described by a matrix of transition probabilities. **C**, the combination of both Markovian processes leads to a Markov-modulated Markovian process. The hidden states of the resulting hidden Markov model are all pairwise combinations of discretized classes in **A** and **B**. **D**. The emission probabilities of  $\theta$ -iSMC

*Demographic history.* A history of fluctuating population sizes enlarges variance in the distribution of genealogies; it results in higher density of coalescence events in epochs of small  $N_e$  and lower density of coalescence events in epochs of large  $N_e$  [35]. In the simple case of a single population bottleneck, a bimodal distribution of TMRCA is expected, with

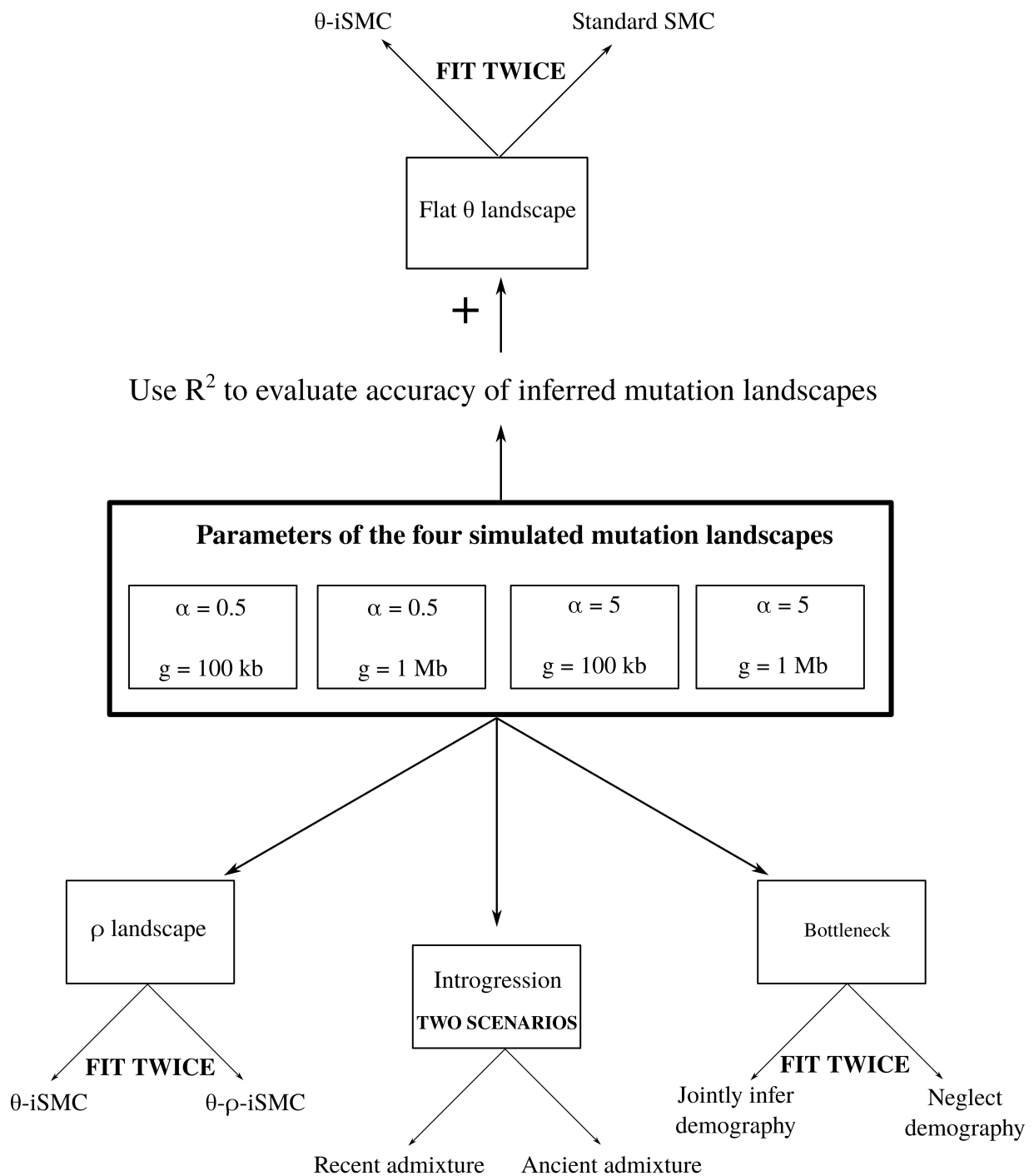
many recent coalescence events when sequences are segregating in a small population, and taking on average much more time to coalesce from the backwards-in-time point when  $N_e$  increases. Since recent coalescence events tend to involve long blocks of contiguous ancestral material while ancient coalescent events tend to involve much shorter segments, a bottleneck influences the distribution of SNPs along a diploid genome: an increased density of both long runs of homosigosity and short blocks of enriched polymorphism is expected relative to the expectation under mutation-drift balance. This suggests that using coalescent-based methods to identify regions of the genome with different mutation rates while neglecting the demographic history should lead to biased inference. We tested this hypothesis by coupling our four simulated mutation landscapes with a strong population bottleneck (see Methods), and fitting  $\theta$ -iSMC twice to each dataset (**Figure 2**, bottom right): first, erroneously assuming a flat demographic history, and second, allowing iSMC to infer piecewise constant coalescence rates in order to accommodate population size changes. As expected, the  $R^2$  is substantially higher when demography is jointly-inferred (ranging from 37.8% to 83.5%, **Figure 3C**) than when it is assumed to be constant (ranging from 2.9% to 66.7%, **Figure 3B**, **Table S3**). Notably, the difference in the distribution of  $R^2$  between demography-aware and demography-oblivious models is much sharper than the one observed in the case of recombination [180], demonstrating that the distribution of genealogies interferes more with the inference of the  $\theta$  landscape than with the inference of the  $\rho$  landscape.

*Introgression.* Next, we investigated the robustness of iSMC to the impact of introgression, which can affect the distribution of SNPs but is not accounted for by the model. After a period of isolation, gene-flow from the “source” population introduces linked polymorphism into the gene pool of the “target” population. As recombination events break introgressed chromosomes apart, repeated back-crossing maintains introgressed blocks segregating. Hence, when diploids from the target population are sampled, regions where an introgressed segment is paired with a “native” segment in the homologous chromosome should display an excess of SNPs. Thus, introgression distorts the distribution of genealogies, mimicking the effect of local changes in  $\theta$ , the magnitude of which is a function of (1) the split time between populations (the longer the time, the higher the divergence between sequences from source and target); (2) the effective number of migrants from the source that contributes to the gene pool of the target (the higher the proportion of introgressed chromosomes, the more often

they are to be found in a randomly sampled diploid); and (3) the time since the admixture event (the longer the time, the more sparsely distributed the introgressed blocks will be in the present population, and the stronger the fluctuations in their frequency imposed by drift). We sought to test this confounding effect on iSMC's estimates of the mutation maps by adding to our four simulated landscapes two scenarios of introgression that differ in their time since admixture (**Figure 2**, bottom center, see Methods). Although lower than in the simulations with no model violation,  $R^2$  values in the presence of introgression remain high (ranging from 56.0% to 94.2% for recent introgression, **Figure 3D**, and from 20.8% to 85.9% for ancient introgression, **Figure 3E**, **Table S4**).

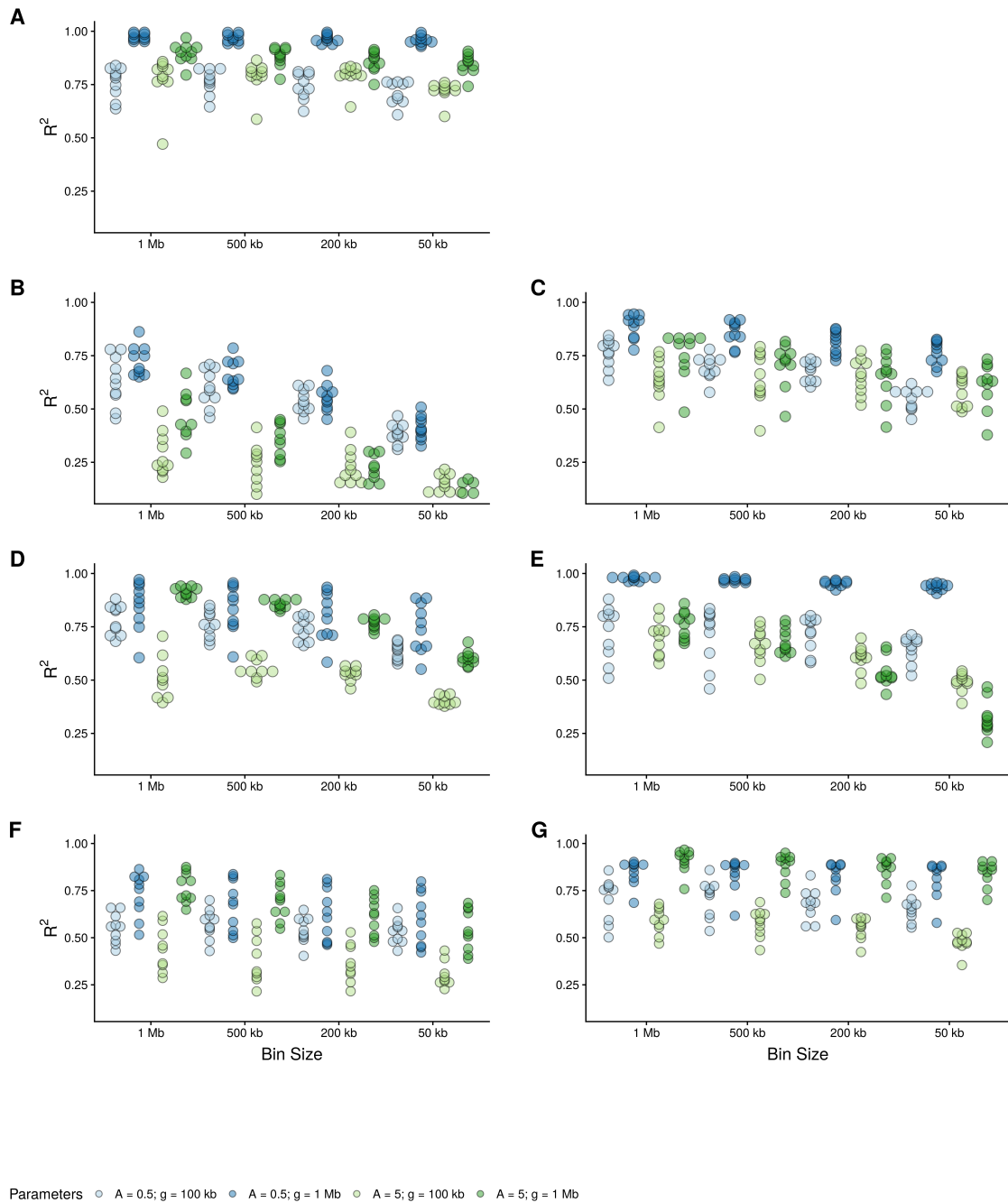
*Spatial variation in recombination rates.* Finally, we investigated the impact of spatially heterogeneous recombination rates on inference of the mutation landscape. For example, in regions of low  $\rho$ , TRMCAs will span an unusually long number of sites. They will thus mimic the effect of local differences in mutation rates by emitting more or fewer SNPs than expected under constant recombination. Therefore, if left as a ghost factor, a heterogeneous  $\rho$  landscape should affect iSMC's ability to distinguish whether genealogies or variable mutation rates are affecting the lengths of runs of homosigosity. To quantify the magnitude of this confounding factor, we imposed a landscape of frequent changes of large magnitude in recombination rates along the genome, and fitted our model twice to this dataset (**Figure 2**, bottom left): first, defining hidden states as pairs of discretised genealogies and mutation rates ( $\theta$ -iSMC, which does not model variation in  $\rho$ ); second, defining hidden states as triplets of discretised genealogies, mutation and recombination rates ( $\theta$ - $\rho$ -iSMC, which models variation in both rates). As expected, model selection based on AIC favours  $\theta$ - $\rho$ -iSMC over  $\theta$ -iSMC for all ten replicate ARGs from this scenario (**Table S5**). The proportion of variance explained when the model accounts for the recombination landscape ( $\theta$ - $\rho$ -iSMC) is systematically higher ( $R^2$  ranging from 35.4% to 96.7%, **Figure 3G**) than when the model neglects it ( $\theta$ -iSMC,  $R^2$  ranging from 21.5% to 87.5%, **Figure 3F**, **Table S6**). Because accuracy globally increases once variation in  $\rho$  is jointly-modelled with variation in  $\theta$ , we conclude that iSMC can disentangle the effects that genome-wide distributions of genealogies, mutation and recombination rates leave on the distribution of diversity.

Use AIC to compare accuracy of model selection



Use  $R^2$  to compare accuracy of inferred mutation landscapes

**Figure 2. Overview of the simulation study.** Four basic landscapes of mutation rate variation (bolded rectangle) are used to evaluate iSMC's baseline discriminatory power (top) and then appended with more complex evolutionary scenarios (bottom).



**Figure 3. Mutation map recovery under various simulated scenarios according to bin size.** Dot plots show the distribution of squared Pearson correlation coefficients ( $R^2$ ) between the simulated and inferred mutation maps. **A**, four scenarios of spatial variation in the mutation rate, corresponding to different combinations of parameters (colour). **B-C**, comparison between a model where demography is mis-specified (**B**) and another where it is jointly inferred (**C**), in the presence of an ancient bottleneck. **D-E**, comparison between a scenario of recent introgression (**D**) and a scenario of ancient introgression (**E**). **F-G**, comparison between a model where recombination rate is assumed to be flat (**F**) and another where it is allowed to vary (**G**), in the presence of a heterogeneous recombination landscape. **Legend:**  $A$  is the shape of the Gama distribution;  $g$  is the average length of blocks.



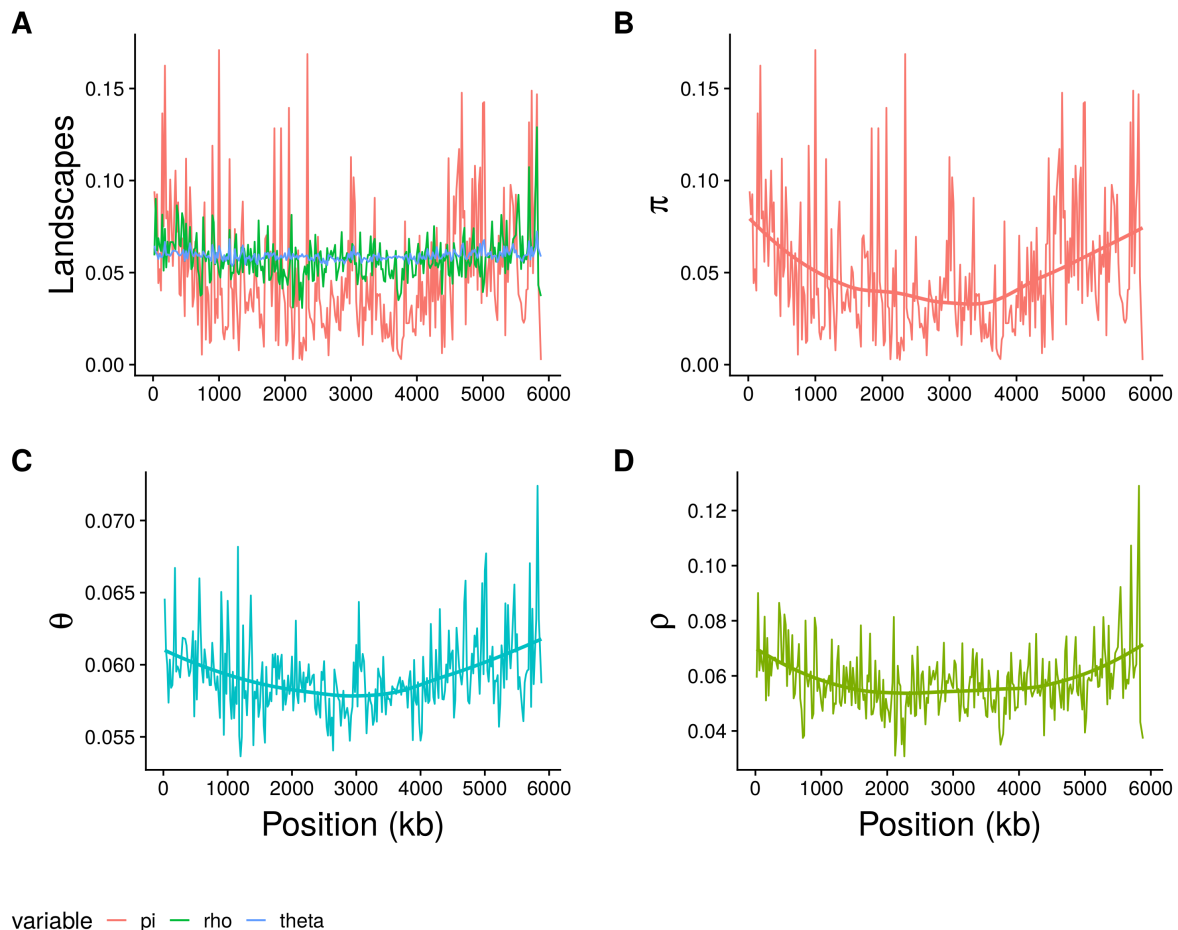
### Case study in *Zymoseptoria tritici*

We performed a case study on the mutation landscape of *Zymoseptoria tritici* using six previously published haploid genomes of this species (data from chromosome 1, see Methods). We aimed at answering two questions: (1) How much variation in diversity is explained by the landscape of *de novo* mutations in *Z. tritici*? and (2) Is there evidence of a mutagenic effect of recombination in *Z. tritici*?

In order to increase power in jointly-inferring the distributions of genealogies, recombination and mutation rates, we further extended iSMC to accommodate multiple genomes. In this augmented model, input genomes are combined in pairs such that the genealogies underlying each of them are still binary, i.e., their topology is trivial and they can be summarised by their TMRCA (**Figure 1A**). Although under Kingman's Coalescent [36] the genealogies of multiple pairs of genomes are not independent, we approximate and compute the composite log-likelihood of the entire dataset by summing the log-likelihoods of each pair. At any generation, variation in the genomic background of individuals should lead to variation in their spatial distribution of mutation and recombination events. For example, population-level diversity in trans-acting modifiers of recombination will cause individuals carrying different sets of alleles to have different expectations for the position of their cross-over events. However, these differences are averaged out within the time-frame of the Coalescent, since chromosomes are expected to spend the same number of generations in each type during the ancestral process. Hence, under panmixia, all pairs of genomes should carry the signatures of the same molecular landscapes. In accordance, iSMC enforces all pairs of genomes to share parameters values, but it does not explicitly enforce a common landscape. Rather, after optimisation of model parameters, it uses the so-called posterior decoding method [134] to infer mutation and recombination maps of single-nucleotide resolution separately for each pair of genomes. Variance of the ARG results in variation among these individual maps. To obtain a consensus of the whole sample, iSMC computes the average  $\theta$  and  $\rho$  (over all pairs of genomes) for all sites in the genome.

Since the genome sequences in the *Z. tritici* dataset were obtained from the haploid phase of each individual, we were able to combine the six haploid sequences in 15 distinct (overlapping) pairs of genomes. We first fitted  $\theta$ - $\rho$ -iSMC with a hotspot model of variation in

$\rho$  to these data. There were, however, estimation issues, suggesting that the hotspot model of recombination interferes with the distribution of mutation rates. When fitting iSMC using Gamma distributions to describe variation in both  $\theta$  and  $\rho$ , we were able to obtain their consensus maps (**Figure 4**). To check whether mutation rate heterogeneity influences the inference of recombination maps in this species, we first estimated the proportion of variance in an experimental cross-over map [147] that is explained by the consensus recombination map inferred with  $\theta$ - $\rho$ -iSMC ( $R^2 = 65.98\%$ ). We then fitted  $\rho$ -iSMC (40x5 configuration) to the same dataset of 15 pairs of genomes and obtained a  $R^2$  of 62.52%. The small improvement confirms that the mutation landscape introduces only a small bias when inferring the recombination landscape [180], and that our model can disentangle the effects that they both leave on the distribution of diversity along the genome.



**Figure 4. The genomic landscapes along chromosome 1 of *Z. tritici*.** **A**, Genetic diversity, mutation and recombination maps, individually normalised to highlight the contrast in their degree of variation. **B**, genetic diversity, as measured by Tajima's  $\pi$ . **C**, mutation rate landscape inferred with iSMC. **D**, recombination rate landscape inferred with iSMC.

To further investigate the joint influence of mutation and recombination in shaping genome-wide diversity in *Z. tritici*, we built a linear model with the average pair-wise diversity among all pairs of genomes ( $\pi$ ) as the dependent variable and the consensus  $\theta$  and  $\rho$  maps as independent variables, allowing for an interaction between them. We found that spatial variation in both mutation and recombination rates have a significant positive effect on the distribution of diversity (**Table S7**). The positive influence of the mutation rate is expected since regions of the genome with higher  $\mu$  should – other things being equal – display higher polymorphism. On the other hand, the positive influence of recombination can be interpreted as a result of a reduced effect of linked selection in regions of higher recombination rate. We also found a significant negative interaction between mutation and recombination, which goes in the reverse direction that would be expected if the mutagenic effect of recombination was an important determinant of genome-wide diversity [181]. This can be explained by (1) continuous evolution of the recombination landscape diluting the signature that imperfect repair of double-strand breaks leaves on polymorphism levels or (2) negligibility of the mutagenic effect of recombination compared to other factors shaping diversity in this species, or a combination of both. The negative interaction suggests that in genomic regions where both  $\theta$  and  $\rho$  are high, the two rates “compete” for leaving their footprints on diversity. Overall, our linear model explains 50.64% of variance in the distribution of diversity (adjusted  $R^2$ ). To assess the contribution of each variable, we performed relative importance and ANOVA tests (**Table S8**). The landscape of *de novo* mutations is the strongest factor (0.60 relative importance), contributing to 38.08% of variance in diversity, whereas the landscape of recombination (0.31 relative importance, 8.6% variance explained) and the interaction between mutation and recombination (0.09 relative importance, 4.48% variance explained) are also significant. Taken together, these results suggests that spatial variation in  $\theta$  is an important factor shaping diversity in *Z. tritici*.

### 3. DISCUSSION:

Our new implementation of iSMC can reconstruct the genome-wide landscape of the population mutation rate  $\theta$ . Its joint-inference approach disentangles the effects of three evolutionary forces on polymorphism data, namely, genetic drift (as modulated by demography), mutation and recombination. To illustrate this point, we re-consider the case of

locally increased diversity in a region of the genome. If the population is panmictic and evolves neutrally, this observation could result from either a higher mutation rate or deeper TMRCA. iSMC will assign posterior probabilities to these competing explanations by taking into account both their global pattern of variation and local information from surrounding positions. If  $\theta$  tends to vary at a broad scale, it is less likely to explain a narrow hotspot of SNPs. Conversely, if the region of increased diversity is long, a large block with ancient TMRCA would be unlikely, unless the local recombination rate is unusually low or the demographic history favours a high density of coalescence events in the deep past. In accordance with the rationale of weighting competing explanations, we have shown that simultaneously modelling multiple evolutionary forces leads to more accurate inference of both mutation and recombination maps.

So far, inference of the mutation landscape has been restricted to either direct observation of *de novo* mutations in large-scale pedigree studies or indirect estimates from diversity or divergence data. iSMC now allows for parametric inference of the mutation landscape from polymorphism data, making estimates of spatial variation in  $\theta$  more accessible. As is the case with the recombination landscape [180], it should be noted that the mutation landscape obtained from pedigree studies and population genomic methods are fundamentally different. The former estimates the distribution of mutation events in the present population, while the latter estimates a mutation landscape of historical influence dating back to the various TMRCA's spread along the genome. Importantly, for the  $\theta$  landscape inferred by iSMC to be interpreted strictly as variation in the rate of *de novo* mutations, two assumptions must be met. First, that the generation time remains constant within the time-scale of interest ( $\sim 4.0$  coalescent units, a measure of time scaled by  $N_e$ ). For example, if there has been a considerable decrease in generation time around 2.0 coalescent units ago, regions of the genome with a TMRCA  $< 2.0$  will tend to harbour more SNPs than expected because these segments will have experienced more reproductive events (hence more frequent DNA replication). Second, that diversity evolves neutrally. In regions of the genome that are under selection, levels of polymorphism at linked sites will be distorted resulting in biased estimates of the local mutation rate. This demonstrates that more complex modelling incorporating an increasing number of biological processes has the potential to unravel more insights into the evolutionary factors shaping the distribution of diversity along genomes.

### 3. METHODS:

#### Modelling spatial variation in $\theta$

Because iSMC models pairs of genomes, the genealogies underlying each orthologous site can be summarized by  $\tau$ , the time to their most recent common ancestor (TMRCA). The pair of DNA sequences can be described as a binary sequence where 0 represents a homozygous position and 1 represents a heterozygous position (thus phasing information is discarded). Under the infinite-sites mutation model, the probability of observing a 0 or 1 at any given position of the genome depends only on  $\tau$  and the population mutation rate  $\theta$ . If the hidden state configuration precludes variation in the mutation rate, then  $\theta$  is assumed to be a global parameter such that the emission probabilities of homozygous and heterozygous states can be compute for every site as  $P(0 | \tau) = \exp(-\theta\tau)$ , and  $P(1 | \tau) = 1 - \exp(-\theta\tau)$ , respectively.

To incorporate spatial heterogeneity in the mutation rate, we set Watterson's estimator of  $\theta$  as the genome-wide average mutation rate, and modulate it drawing scaling factors from a discretised prior distribution with mean equal to 1.0. The parameters shaping this distribution can be viewed as hyper-parameters of our HMM. We model the changes in mutation rate along the genome as a Markov process where the transition probability between each pair of categories is the same (parameter  $\omega$ ). The justification for the Markov model is that sites in close proximity are expected to have similar mutation rates, for example, as is the case when the efficiency of the replication machinery decreases with increasing distance from the start of the replication fork. Since the emission probabilities depend on  $\theta$ , the resulting process is Markov-modulated by the mutation rate. Let  $t$  be the number of discretised TMRCA intervals, and  $k$  be the number of discretised categories of the prior distribution of scaling factors of  $\theta$ . The ensuing Markov-modulated HMM is an HMM with  $n = t \times k$  hidden states, whose emission probabilities are captured by the  $n \times 2$  matrix depicted in **Figure 1d**. Moreover, even though  $\theta$  itself does not affect the transition probabilities of the HMM, the transition between its categories weight in the transition probabilities between hidden states [180]. The transition matrix for spatial variation in  $\theta$  is:

$$Q_{\theta} = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & & & P_{2k} \\ \vdots & & & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix} = \begin{bmatrix} 1-\omega & \frac{\omega}{k-1} & \cdots & \frac{\omega}{k-1} \\ \frac{\omega}{k-1} & 1-\omega & & \frac{\omega}{k-1} \\ \vdots & & & \vdots \\ \frac{\omega}{k-1} & \frac{\omega}{k-1} & \cdots & 1-\omega \end{bmatrix}$$

And the forward recursion for this model at genomic position  $i$  can be written as:

$$F_{i,G_v}(\theta_m) = \left( \sum_{l=1}^k \left( \sum_{u=1}^t F_{i-1,G_u}(\theta_l) \cdot Pr(G_u \rightarrow G_v) \cdot Pr(\theta_l \rightarrow \theta_m) \right) \right) \cdot Pr(G_v \rightarrow S_i | \theta_m) \quad (1)$$

where  $\theta_m$  is the product of the genome-wide average mutation rate and the value of the  $m$ th discretised category drawn from its prior distribution. The emission probability of binary state  $S_i$  depends on the TMRCA of genealogy  $G_v$  and the focal mutation rate  $\theta_m$ . The forward recursion integrates over all  $k$  discretised values of  $\theta$  and over all  $t$  TMRCA intervals. In the double-modulated model, this integration is performed over discretised values of  $\theta$ , TMRCA intervals as well as  $a$  discretised values of  $\rho$ , which weight in the transition probability between genealogies:

$$F_{i,G_v}(\theta_m, \rho_c) = \left( \sum_{l=1}^k \left( \sum_{b=1}^a \left( \sum_{u=1}^t F_{i-1,G_u}(\theta_l, \rho_b) \cdot Pr(G_u \rightarrow G_v | \rho_b) \cdot Pr(\theta_l \rightarrow \theta_m) \cdot Pr(\rho_b \rightarrow \rho_c) \right) \right) \right) \cdot Pr(G_v \rightarrow S_i | \theta_m) \quad (2)$$

And the single-nucleotide mutation landscapes is obtained as in [180] (chapter 1).

### Simulation study

*Four scenarios of spatial variation in  $\theta$ .* We simulated a piecewise constant recombination rate along the genome by drawing values from a Gamma distribution with parameters  $\alpha$  and  $\beta$ , and segment lengths from a geometric distribution with mean length  $g$ . We considered four possible scenarios where  $\alpha = \beta = 0.5$  or  $5.0$ , and  $g = 100$  kb or  $1$  Mb. For each of the four combinations, we simulated 10 independent pairs of two 30 Mb haploid chromosomes under a constant population size model, assuming  $\theta = 0.003$  and  $\rho = 0.0012$ . All of the following

simulated scenarios share base parameters values and have an extra layer of complexity. A schematic representation of the simulation study can be found in **Figure 2**. The  $R^2$  values reported correspond to the square of the Pearson correlation coefficient between simulated and inferred maps.

*Demographic history.* We simulated the scenario of population bottleneck 0.5 coalescent time units ago. Translating it to generations based on effective population sizes, the bottleneck happened 30,000 generations ago.

*Introgression.* We simulated two introgression scenarios where a source population introduces a pulse of genetic material into a target population. In both scenarios, the split between source and target populations happened 2.0 coalescent time units ago, and the source replaces 10% of the genetic pool of the target. In the first scenario, secondary contact happened 0.125 coalescent time units ago; in the second, it happened 0.25 coalescent time units ago. Translating these coalescent times based on effective population sizes, the split between population happened 120,000 generations ago, and the introgression events happened 7,500 and 15,000 generations ago, respectively.

*Variation in the recombination rate.* We simulated a piecewise constant recombination rate along the genome by drawing rate values from a gamma distribution with parameters  $\alpha = \beta = 0.5$  and segment lengths from a geometric distribution with mean length 100 kb.

### **Analysis of the mutation landscape in *Z. tritici***

To fit the double-modulated iSMC model to data from *Z. tritici*, we used as input all the 15 pair-wise combinations of the following six haploid sequences from chromosome 1 [146]: Zt05, Zt07, Zt09, Zt150, Zt154, Zt155. After obtaining consensus mutation and recombination maps of this sample at the 20 kb scale, we first excluded nine of the 294 windows that displayed an excess of diversity and are likely the product of introgression (personal communication from Alice Feurtey, manuscript in preparation). We then built an ordinary least squares regression model with the average pair-wise diversity ( $\pi$ ) as response variable and the inferred maps as explanatory variables, allowing for an interaction between them. The model was Box-Cox transformed using the MASS package [182] to bring the

distribution of residuals closer to normality (nevertheless, the Shapiro-Wilk test of normality was significant with p-value = 0.0052). Both the Harrison-McCabe test for homoscedasticity and the Durbin-Watson test for auto-correlation yielded non-significant results. We performed an ANOVA test and computed the proportion of variance explained by each variable as their relative sum of squares; we computed their relative importance using the relaimpo package [183].

## SUPPLEMENTAL MATERIAL:

The following supplemental material is available in the digital archive

**Table S1:** AIC values according to replicate, combination of parameters in the mutation landscape and iSMC model (homogeneous or  $\theta$ -modulated).

**Table S2:**  $R^2$  values according to replicate, combination of parameters in the mutation landscape and bin size used to average the maps.

**Table S3:**  $R^2$  values according to replicate, combination of parameters in the mutation landscape, bin size used to average the maps and whether the demography was jointly-inferred.

**Table S4:**  $R^2$  values according to replicate, combination of parameters in the mutation landscape, bin size used to average the maps and time since introgression event.

**Table S5:** AIC values according to replicate, combination of parameters in the mutation landscape and iSMC model ( $\theta$ -modulated or  $\theta$ - $\rho$ -modulated).

**Table S6:**  $R^2$  values according to replicate, combination of parameters in the mutation landscape, bin size used to average the maps and whether the recombination landscape is jointly-modelled.

**Table S7:** Summary of the linear model fitted to *Z. tritici* data:  $\pi = \beta_0 + \beta_1\theta + \beta_2\rho + \beta_3\theta\rho + e_i$ .

**Table S8:** ANOVA table based on the linear model fitted to *Z. tritici* data.



## CHAPTER 3

# Selection at the pathway level drives the evolution of gene-specific transcriptional noise

**Authors:** Gustavo Valadares Barroso<sup>1</sup>; Natasa Puzovic<sup>1</sup> and Julien Y Dutheil<sup>1</sup>,

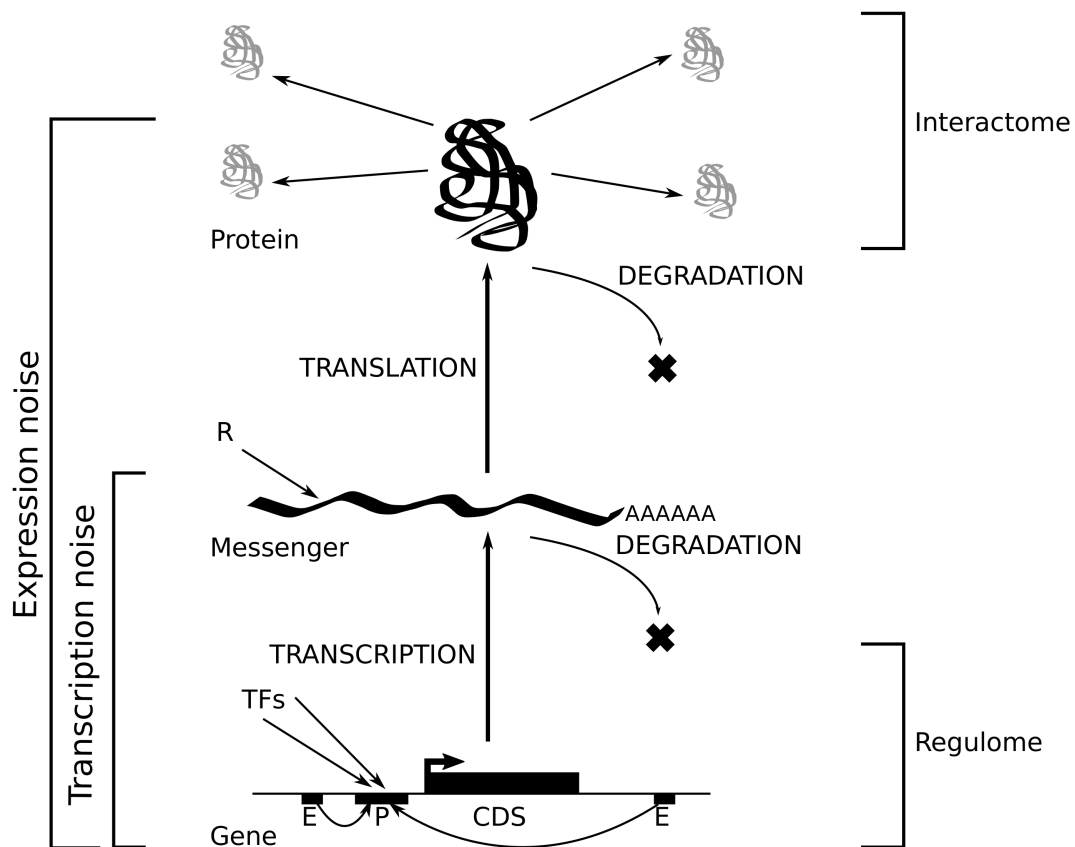
**Affiliations:** 1) Max Planck Institute for Evolutionary Biology, Department of Evolutionary Genetics, August-Thienemann-Straße 2 24306 Plön – GERMANY

### ABSTRACT:

Biochemical reactions within individual cells result from the interactions of molecules, typically in small numbers. Consequently, the inherent stochasticity of binding and diffusion processes generate noise along the cascade that leads to the synthesis of a protein from its encoding gene. As a result, isogenic cell populations display phenotypic variability even in homogeneous environments. The extent and consequences of this stochastic gene expression have only recently been assessed on a genome-wide scale, in particular owing to the advent of single cell transcriptomics. However, the evolutionary forces shaping this stochasticity have yet to be unraveled. We take advantage of two recently published data sets of the single-cell transcriptome of the domestic mouse *Mus musculus* in order to characterize the effect of natural selection on gene-specific transcriptional stochasticity. We show that noise levels in the mRNA distributions (*a.k.a.* transcriptional noise) significantly correlate with three-dimensional nuclear domain organization, evolutionary constraint on the encoded protein and gene age. The position of the encoded protein in biological pathways, however, is the main factor that explains observed levels of transcriptional noise, in agreement with models of noise propagation within gene networks. Because transcriptional noise is under widespread selection, we argue that it constitutes an important component of the phenotype and that variance of expression is a potential target of adaptation. Stochastic gene expression should therefore be considered together with mean expression level in functional and evolutionary studies of gene expression.

# 1. INTRODUCTION:

Isogenic cell populations display phenotypic variability even in homogeneous environments [72]. This observation challenged the clockwork view of the intra-cellular molecular machinery and led to the recognition of the stochastic nature of gene expression. Since biochemical reactions result from the interactions of individual molecules in small numbers [184], the inherent stochasticity of binding and diffusion processes generates noise along the biochemical cascade leading to the synthesis of a protein from its encoding gene (**Figure 1**). The study of stochastic gene expression (SGE) classically recognizes two sources of expression noise. Following the definition introduced by Elowitz et al [75], extrinsic noise results from variation in concentration, state and location of shared key molecules involved in the reaction cascade from transcription initiation to protein folding. This is because molecules that are shared among genes, such as ribosomes and RNA polymerases, are typically present in low copy numbers relative to the number of genes actively transcribed [185]. Extrinsic factors also include physical properties of the cell such as size and growth rate, likely to impact the diffusion process of all molecular players. Extrinsic factors therefore affect every gene in a cell equally. Conversely, intrinsic factors generate noise in a gene-specific manner. They involve, for example, the strength of cis-regulatory elements [186] as well as the stability of the mRNA molecules that are transcribed [77,187]. Every gene is affected by both sources of stochasticity and the relative importance of each has been discussed in the literature [188,189]. Shahrezaei and Swain [185] proposed a more general, systemic and explicit definition for any organization level, where intrinsic stochasticity is “generated by the dynamics of the system from the random timing of individual reactions” and extrinsic stochasticity is “generated by the system interacting with other stochastic systems in the cell or its environment”. This generic definition therefore includes Raser and O’Shea’s [190] suggestion to further distinguish extrinsic noise occurring “within pathways” and “between pathways”. Other organization levels of gene expression are also likely to affect expression noise, such as chromatin structure [82,191], and three-dimensional genome organization [192].



**Figure 1: A systemic view of gene expression.**

Pioneering work by Fraser et al [88] has shown that SGE is an evolvable trait which is subject to natural selection. First, genes involved in core functions of the cell are expected to behave more deterministically [76] because temporal oscillations in the concentration of their encoded proteins are likely to have a deleterious effect. Second, genes involved in immune response [193,194] and response to environmental conditions can benefit from being unpredictably expressed in the context of selection for bet-hedging [195]. As the relation between fitness and stochasticity depends on the function of the underlying gene, selection on SGE is expected to act mostly at the intrinsic level [196–198]. The molecular mechanisms by which natural selection operates to regulate expression noise, however, remain to be elucidated.

Due to methodological limitations, seminal studies on SGE (both at the mRNA and protein

levels) have focused on only a handful of genes [75,199,200]. The canonical approach consists in selecting genes of interest and recording the change of their noise levels in a population of clonal cells as a function of either (1) the concentration of the molecule that allosterically controls affinity of the transcription factor to the promoter region of the gene [82,201] or (2) mutations artificially imposed in regulatory sequences [199]. In parallel with theoretical work [202–205], these pioneering studies have provided the basis of our current understanding of the proximate molecular mechanisms behind SGE, namely complex regulation by transcription factors, architecture of the upstream region (including the presence of TATA box) and gene orientation [206], translation efficiency and mRNA / protein stability [207], properties of the protein-protein interaction network [208]. Measurements at the genome scale coupled with rigorous statistical analyses are however needed in order to go beyond gene idiosyncrasies and particular histories, and test hypotheses about the evolutionary forces shaping SGE [209].

The recent advent of single-cell RNA sequencing makes it possible to sequence the transcriptome of each individual cell in a collection of clones, and to observe the variation of gene-specific mRNA quantities across cells. This provides a genome-wide assessment of transcriptional noise. While not accounting for putative noise resulting from the process of translation of mRNAs into proteins, transcriptional noise accounts for noise generated by both synthesis and degradation of mRNA molecules (**Figure 1**). Previous studies, however, have shown that transcription is a limiting step in gene expression, and that transcriptional noise is therefore a good proxy for expression noise [196,210]. Here, we used publicly available single-cell transcriptomics data sets to quantify gene-specific transcriptional noise and relate it to other genomic factors, including protein conservation and position in the interaction network, in order to uncover the molecular basis of selection on stochastic gene expression.

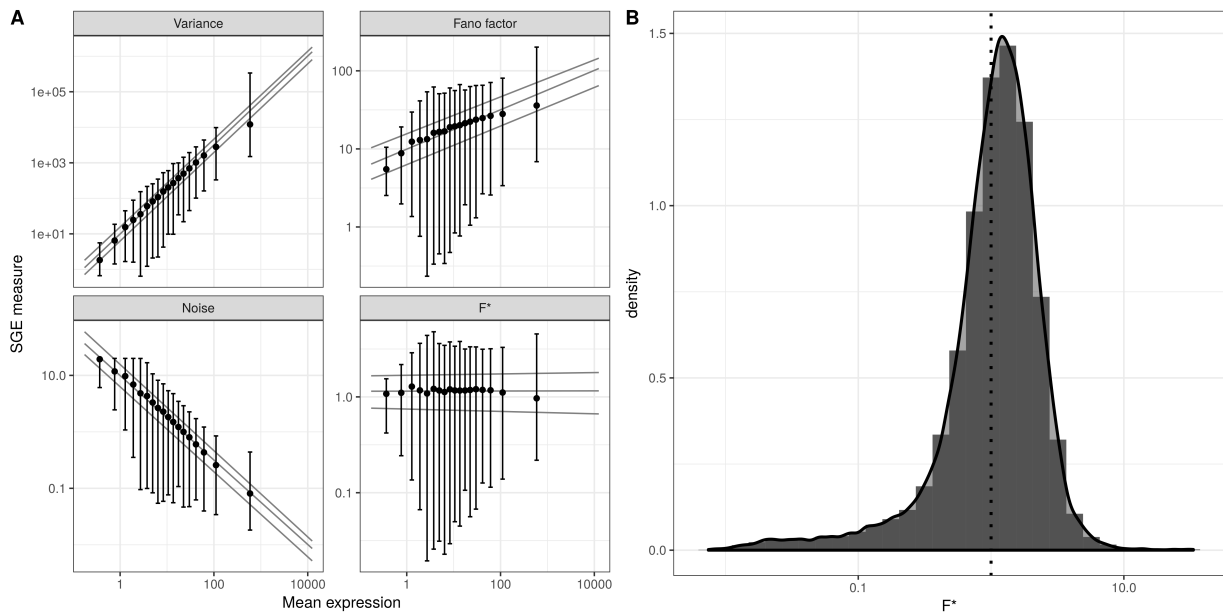
## **2. RESULTS:**

### **A new measure of noise to study genome-wide patterns of stochastic gene expression**

We used the dataset generated by Sasagawa et al (2013), which quantifies gene-specific

amounts of mRNA as fragments per kilobase of transcripts per million mapped fragments (FPKM) values for each gene and each individual cell. Among these, we selected all genes in a subset containing 20 embryonic stem cells in G1 phase in order to avoid recording variance that is due to different cell types or cell-cycle phases. The Quartz-Seq sequencing protocol captures every poly-A RNA present in the cell at one specific moment, allowing to assess transcriptional noise. Following Shalek et al (2014) we first filtered out genes that were not appreciably expressed in order to reduce the contribution of technical noise to the total noise. For each gene we further calculated the mean  $\mu$  in FPKM units and variance  $\sigma^2$  in FPKM<sup>2</sup> units, as well as two previously published measures of stochasticity: the *Fano factor*, usually referred to as the bursty parameter, defined as  $\sigma^2/\mu$  and *Noise*, defined as the coefficient of variation squared ( $\sigma^2/\mu^2$ ). Both the variance and *Fano factor* are monotonically increasing functions of the mean (**Figure 2A**). *Noise* is inversely proportional to mean expression (**Figure 2A**), in agreement with previous observations at the protein level [201,210]. While this negative correlation was theoretically predicted [211] it may confound the analyses of transcriptional noise at the genome level, because mean gene expression is under specific selective pressure [212] In order to disentangle these effects, we developed a new quantitative measure of noise, independent of the mean expression level of each gene.

To achieve this we performed polynomial regressions in the log-space plot of variance versus mean. We defined  $F^*$  as  $\sigma_{obs}^2/\sigma_{pred}^2$  (see Material and Methods) that is, the ratio of the observed variance over the variance component predicted by the mean expression level. We selected the simplest model for which no correlation between  $F^*$  and mean expression was observed, and found that a degree 3 polynomial model was sufficient to remove further correlation (Kendall's tau = -0.0037, p-value = 0.5217, **Figure 2A**). Genes with  $F^* < 1$  have a variance lower than expected according to their mean expression whereas genes with  $F^* > 1$  behave the opposite way (**Figure 2B**). This approach fulfills the same goal as the running median approach of Newman et. al [196], whilst it includes the effect of mean expression directly into the measure of stochasticity instead of correcting a posteriori a dependent measure (in that case, the Fano factor). We therefore use  $F^*$  as a measure of SGE throughout this study.



**Figure 2: Transcriptional noise and mean gene expression.** **A**, Measures of noise plotted against the mean gene expression for each gene, in logarithmic scales: Variance, Fano factor (variance / mean), noise (square of the coefficient of variation, variance / mean<sup>2</sup>) and F\* (this study). Lines represent quantile regression fits (median, first and third quartiles). Point and bars represent median, first and third quartiles for each category of mean expression obtained by discretization of the x axis. **B**, Distribution of F\* over all genes in this study. Vertical line corresponds to F\* = 1.

### Stochastic gene expression correlates with the three-dimensional structure of the genome

We first sought to investigate whether genome organization significantly impacts the patterns of stochastic gene expression. We assessed whether genes in proximity along chromosomes display more similar amount of transcriptional noise than distant genes. We tested this hypothesis by computing the primary distance on the genome between each pair of genes, that is, the number of base pairs separating them on the chromosome, as well as the relative difference in their transcriptional noise (see Methods). We found no significant association between the two distances (Mantel tests, each chromosome tested independently). Contiguous genes in one dimension, however, have significantly more similar transcriptional noise than non-contiguous genes (permutation test, p-value < 1E-4, **Figure S1**). Using Hi-C data from mouse embryonic cells [213], we report that genes in contact in three-dimensions have significantly more similar transcriptional noise than genes not in contact (permutation test, p-value < 1E-3, **Figure S1**). Most contiguous genes in one-dimension also appear to be close in three-dimensions and the effect of 3D contact is stronger than that of 1D contact. These

results therefore suggest that the three-dimensional structure of the genome has a stronger impact on stochastic gene expression than the position of the genes along the chromosomes. We further note that while highly significant, the size of this effect is small, with a difference in relative expression of -1.10% (**Figure S1**).

### **Transcription factors binding and histone methylation impact stochastic gene expression**

The binding of transcription factors (TF) to promoter constitutes one notable source of transcriptional noise (**Figure 1**) [82,196]. In eukaryotes, the accessibility of promoters is determined by the chromatin state, which is itself controlled by histone methylation. We assessed the extent to which transcriptional noise is linked to particular TFs and histone marks by using data from the Ensembl regulatory build [214], which provides data from experimental evidence of TF binding and methylation sites along the genome. First we contrasted the  $F^*$  values of genes with binding evidence for each annotated TF independently. Among 13 TF represented by at least 5 genes in our data set, we found that 4 of them significantly influence  $F^*$  after adjusting for a global false discovery rate of 5%: the transcription repressor CTCF (adjusted p-value = 0.0321), the transcription factor CP2-like 1 (Tcfcp2l1, adjusted p-value = 0.0087), the X-Linked Zinc Finger Protein (Zfx, adjusted p-value = 0.0284) and the Myc transcription factor (MYC, adjusted p-value = 0.0104). Interestingly, association with each of these four TFs led to an increase in transcriptional noise. We also report a weak but significant positive correlation between the number of transcription factors associated with each gene and the amount of transcriptional noise (Kendall's tau = 0.0238, p-value = 0.0007). This observation is consistent with the idea that noise generated by each TF is cumulative [215]. We then tested if particular histone marks are associated with transcriptional noise. Among five histone marks represented in our data set, three were found to be highly significantly associated to a higher transcriptional noise: H3K4me3 (adjusted p-value = 1.9981E-146), H3K4me2 (adjusted p-value = 5.4524E-121) and H3K27me3 (adjusted p-value = 5.2985E-34). Methylation on the fourth Lysine of histone H3 is associated with gene activation in humans, while tri-methylation on lysine 27 is usually associated with gene repression [216]. These results suggest that both gene activation and silencing contribute to the stochasticity of gene expression, in agreement with the view that

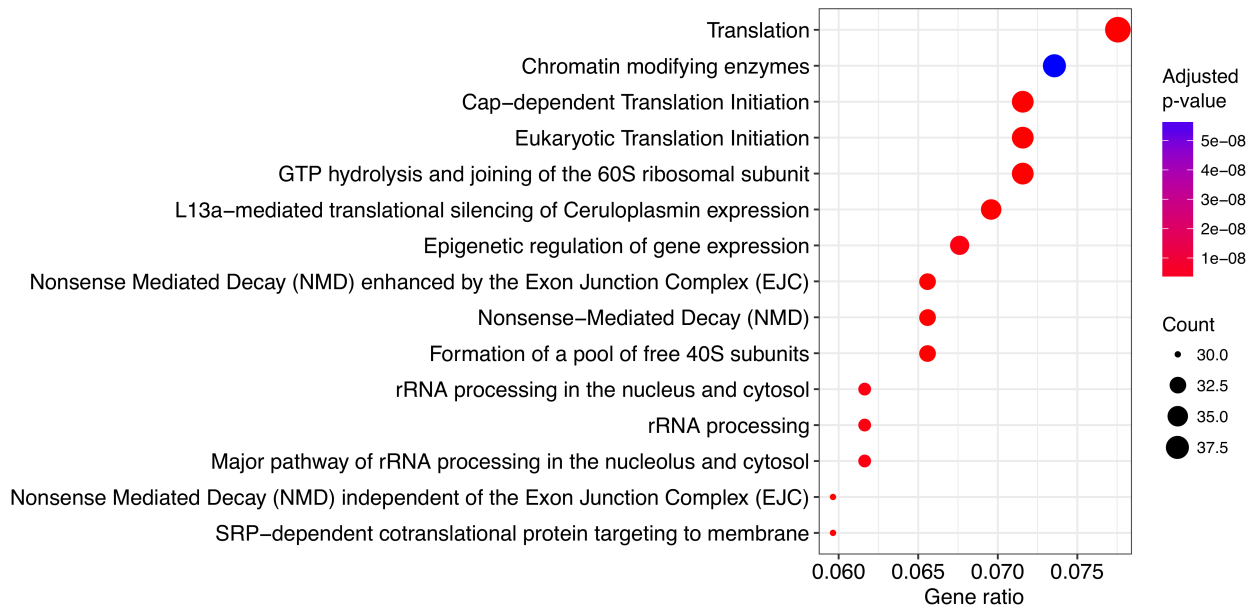
bursty transcription leads to increased noise [82,196,203].

### **Low noise genes are enriched for housekeeping functions**

We investigated the function of genes at both ends of the  $F^*$  spectrum. We defined as candidate gene sets the top 10% least noisy or the top 10% most noisy genes in our data set, and tested for enrichment of GO terms and Reactome pathways (see Methods). It is expected that genes encoding proteins participating in housekeeping pathways are less noisy because fluctuations in concentration of their products might have stronger deleterious effects [86]. On the other hand, stochastic gene expression could be selectively advantageous for genes involved in immune and stress response, as part of a bet-hedging strategy [193,217]. GO terms enrichment test revealed significant categories enriched in the low noise gene set only: molecular functions “nucleic acid binding” and “structural constituent of ribosome”, the biological processes “nucleosome assembly”, “innate immune response in mucosa” and “translation”, as well as the cellular component “cytosolic large ribosomal subunit” (**Table 1**). All these terms but one relate to gene expression, in agreement with previously reported findings in yeast [196]. We further find a total of 41 Reactome pathways significantly over-represented in the low-noise gene set (false discovery rate set to 1%). Interestingly, the top most significant pathways belong to modules related to translation (RNA processing, initiation of translation and ribosomal assembly), as well as several modules relating to gene expression, including chromatin regulation and mRNA splicing (**Figure 3**). Only one pathway was found to be enriched in the high noise set: TP53 regulation of transcription of cell cycle genes (p-value = 0.0079). This finding is interesting because TP53 is a central regulator of stress response in the cell [218]. These results therefore corroborate previous findings that genes involved in stress response might be evolving under selection for high noise as part of a bet hedging strategy [217,219].

The small amount of significantly enriched Reactome pathways by high noise genes can potentially be explained by the nature of the data set: as the original experiment was based on unstimulated cells, genes that directly benefit from high SGE might not be expressed in these conditions.





**Figure 3: Enriched pathways in the low-noise gene set.** Depicted pathways are the fifteen most significant in the 10% genes with lowest transcriptional noise.

### Highly connected proteins are synthesized by low-noise genes

The structure of the interaction network of proteins inside the cell can greatly impact the evolutionary dynamics of genes [80,220]. Furthermore, the contribution of each constitutive node within a given network varies. This asymmetry is largely reflected in the power-law-like degree distribution that is observed in virtually all biological networks [221] with a few genes displaying many connections and a majority of genes displaying only a few. The individual characteristics of each node in a network can be characterized by various measures of centrality [222]. Following previous studies on protein evolutionary rate [223–225] and protein-protein interaction (PPI) networks [208] we asked whether, at the gene level, there is a link between centrality of a protein and the amount of transcriptional noise. We study six centrality metrics measured on two types of network data: (i) pathway annotations from the Reactome database [226] and (ii) PPI data from the iRefIndex database. PPI data are typically more complete (5,553 genes with gene expression data) but do not provide functional evidence. The Reactome database is based on published functional evidence, but encompasses less genes (4,454 genes for which expression data is available). In addition, graph representing PPI network are not oriented while graph representing Pathway annotations are, implying that distinct statistics can be computed on both types of networks.

We first estimated the pleiotropy index of each gene by counting how many different pathways the corresponding proteins are involved in. We then computed centrality measures as averages over all pathways in which each gene is involved. These measures include (1) *node degree*, which corresponds to the number of other nodes a given node is directly connected with, (2) *hub score*, which estimates the extent to which a node links to other central nodes, (3) *authority score*, which estimates the importance of a node by assessing how many hubs link to it, (4) *transitivity*, or *clustering coefficient*, defined as the proportion of neighbors that also connect to each other, (5) *closeness*, a measure of the topological distance between a node and every other reachable node (the fewer edge hops it takes for a protein to reach every other protein in a network, the higher its closeness), and (6) *betweenness*, a measure of the frequency with which a protein belongs to the shortest path between every pairs of nodes.

We find that node degree, hub score, authority score and transitivity are all significantly negatively correlated with transcriptional noise on pathway-based networks: the more central a protein is, the less transcriptional noise it displays (**Figure 4A-D** and **Table 2**). We also observed that pleiotropy is negatively correlated with  $F^*$  (Kendall's tau = -0.0514, p-value = 8.31E-07, **Figure 4E**, **Table 2**), suggesting that a protein that potentially performs multiple functions at the same time needs to be less noisy. This effect is not an artifact of the fact that pleiotropic genes are themselves more central (e.g. correlation of pleiotropy and node degree: Kendall's tau = 0.2215, p-value < 2.2E-16) or evolve more slowly (correlation of pleiotropy and  $K_a / K_s$  ratio: Kendall's tau = -0.1060, p-value < 2.2E-16) since it is still significant after controlling for these variables (partial correlation of pleiotropy and  $F^*$ , accounting for centrality measures and  $K_a / K_s$ : Kendall's tau = -0.0254, p-value = 7.45E-06). Closeness and betweenness, on the other hand, show a negative correlation with  $F^*$ , yet much less significant (Kendall's tau = -0.0254, p-value = 0.0109 for closeness and tau = -0.0175, p-value = 0.0865 for betweenness, see **Figure 4F-G** and **Table 2**). In modular networks [81] nodes that connect different modules are extremely important to the cell [227] and show high betweenness scores. In yeast, high betweenness proteins tend to be older and more essential [228], an observation also supported by our data set (betweenness vs gene age, Kendall's tau = 0.0619, p-value = 1.09E-07; betweenness vs  $K_a/K_s$ , Kendall's tau = -0.0857, p-value = 3.83E-16). It has been argued, however, that in protein-protein interaction networks high

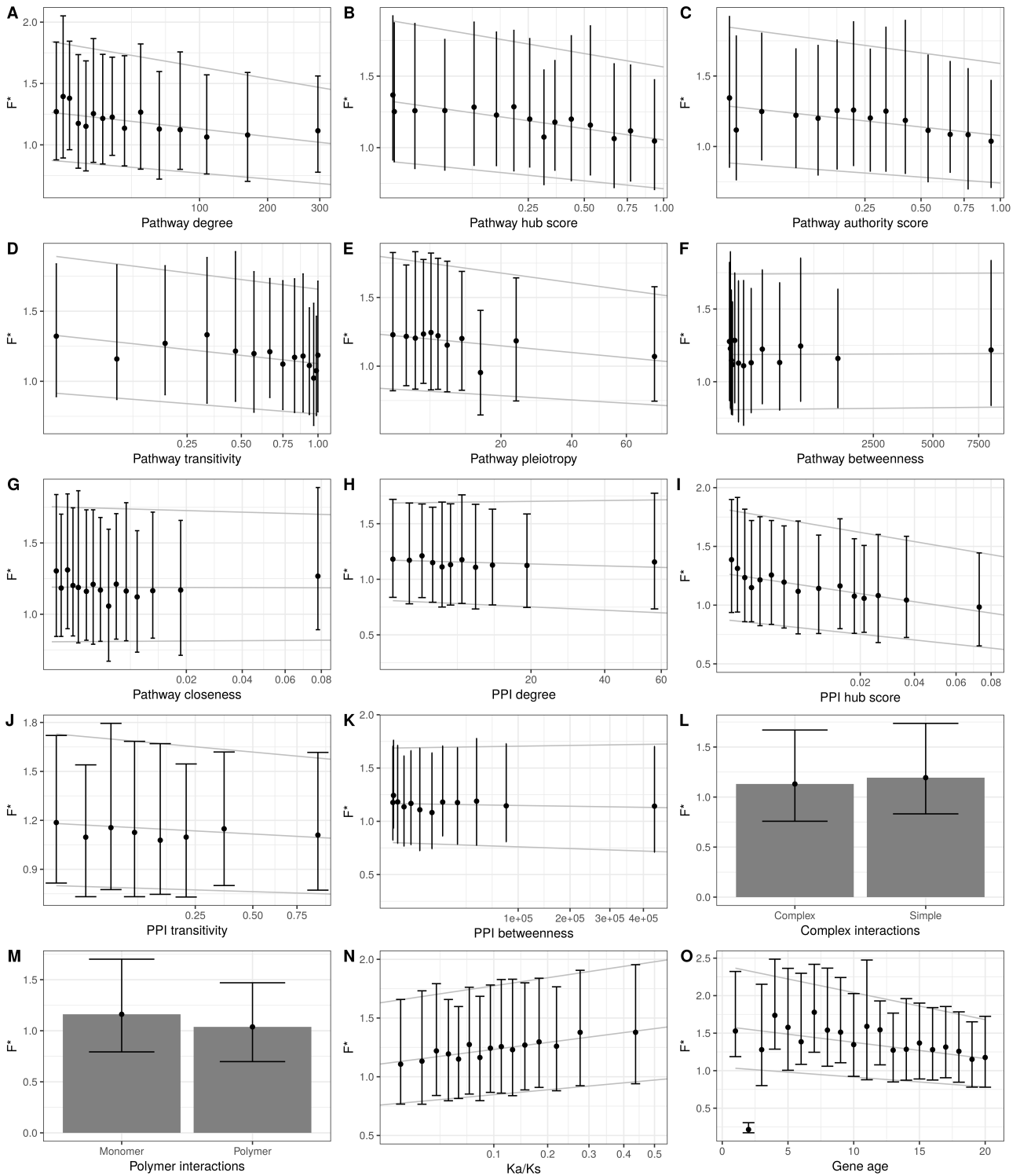
betweenness proteins are less essential due to the lack of directed information flow, compared to, for instance, regulatory networks [229], a hypothesis which could explain the observed lack of correlation.

By applying similar measures on the PPI network, we report significant negative correlation between  $F^*$  and PPI centrality measures (**Figure 4H-K**, **Table 2**). Because the PPI network is not directed, authority scores and hub scores cannot be distinguished. The results obtained with the mouse PPI interaction network are qualitatively similar to the ones obtained by Li et al (2010) on Yeast expression data [208]. In addition, we further report that genes involved in complex interactions (that is, genes which interact with more than one other protein simultaneously) have reduced noise in gene expression (Wilcoxon rank test, p-value =  $8.053E-05$ , **Figure 4L**), corroborating previous findings in Yeast [88]. Conversely, genes involved in polymeric interactions, that is, where multiple copies of the encoded protein interact with each other, did not show significantly different noise than other genes (Wilcoxon rank test, p-value = 0.0821, **Figure 4M**).

It was previously shown that centrality measures negatively correlate with evolutionary rate [230]. Our results suggest that central genes are selectively constrained for their transcriptional noise, and that centrality therefore also influences the regulation of gene expression. Interestingly, it has been reported that central genes tend to be more duplicated [231]. The authors proposed that such duplication events would have been favored as they would confer greater robustness to deleterious mutations in proteins. Our results are compatible with another, non exclusive, possible advantage: having more gene copies could reduce transcriptional noise by averaging the amount of transcripts produced by each gene copy [190].

### **Network structure impacts transcriptional noise of constitutive genes**

Whereas estimators of node centrality highlight gene-specific properties inside a given network, measures at the whole-network level enable the comparison of networks with distinct properties.



**Figure 4: Factors driving stochastic gene expression.** Correlation of  $F^*$  and all tested network centrality measures, as well as protein conservation ( $K_a / K_s$  ratio) and gene age. Point and bars represent median, first and third quartiles for each category of mean expression obtained by discretization of the x axis, together with the quantile regression lines estimated on the full data set.

We computed the size, diameter and global transitivity for each annotated network in our data set (1,364 networks, Supplementary Material) which we compare with the average  $F^*$  measure of all constitutive nodes. The size of a network is defined as its total number of nodes, while diameter is the length of the shortest path between the two most distant nodes. Transitivity is a measure of connectivity, defined as the average of all nodes' clustering coefficients. Interestingly, while network size is positively correlated with average degree and transitivity (Kendall's tau = 0.5880, p-value < 2.2E-16 and Kendall's tau = 0.1166, p-value = 1.08E-10, respectively), diameter displays a positive correlation with average degree (Kendall's tau = 0.2959, p-value < 2.2E-16) but a negative correlation with transitivity (Kendall's tau = -0.0840, p-value = 2.17E-05). This is because diameter increases logarithmically with size, that is, addition of new nodes to large networks do not increase the diameter as much as additions to small networks. This suggests that larger networks are relatively more compact than smaller ones, and their constitutive nodes are therefore more connected. We find that average transcriptional noise correlates negatively with network size (Kendall's tau = -0.0514, p-value = 0.0039), while being independent of the diameter (Kendall's tau = 0.0061, p-value = 0.7547 see **Table 3**). These results are in line with the node-based analyses, and show that the more connections a network has, the less stochastic the expression of the underlying genes is. This supports the view of Raser and Oshea [190] that the gene-extrinsic, pathway-intrinsic level is functionally pertinent and needs to be distinguished from the globally extrinsic level. We further asked whether genes with similar transcriptional noise tend to synthesize proteins that connect to each other (positive assortativity) in a given network, or on the contrary, tend to avoid each other (negative assortativity). We considered all Reactome pathways annotated to the mouse and estimated their respective  $F^*$  assortativity. We found the mean assortativity to be significantly negative, with a value of -0.1384 (one sample Wilcoxon rank test, p-value < 2.2E-16), meaning that proteins with different  $F^*$  values tend to connect with each other (**Figure S3**). Maslov & Sneppen [232] reported a negative assortativity between hubs in protein-protein interaction networks, which they hypothesized to be the result of selection for reduced vulnerability to deleterious perturbations. In our data set, however, we find the assortativity of hub scores to be significantly positive (average of 0.1221, one sample Wilcoxon rank test, p-value = 1.212E-12, **Figure S5**), although with a large distribution of assortativity values. As we showed that hub scores correlates negatively with  $F^*$  (**Table 2**), we asked whether the

negative assortativity of hub proteins can at least partly explain the negative assortativity of  $F^*$ . We found a significantly positive correlation between the two assortativity measures (Kendall's tau = 0.2581, p-value < 2.2E-16). The relationship between the measures, however, is not linear (**Figure S5**), suggesting a distinct relationship between hub score and  $F^*$  for negative and positive hub score assortativity. Negative assortativity of hub proteins contributes to a negative assortativity of SGE (Kendall's tau = 0.2730, p-value < 2.2E-16), while for pathways with positive hub score assortativity the effect vanishes (Kendall's tau = 0.0940, p-value = 3.135E-4). While assortativity of  $F^*$  is closer to 0 for pathways with positive assortativity of hub score, we note that it is still significantly negative (average = -0.0818, one sample Wilcoxon test with p-value < 2.2E-16). This suggests the existence of additional constraints that act on the distribution of noisy proteins in a network.

### **Transcriptional noise is positively correlated with the evolutionary rate of proteins**

In the yeast *Saccharomyces cerevisiae*, evolutionary divergence between orthologous coding sequences correlates negatively with fitness effect on knock-out strains of the corresponding genes [233], demonstrating that protein functional importance is reflected in the strength of purifying selection acting on it. Fraser et al [88] studied transcription and translation rates of yeast genes and classified genes in distinct noise categories according to their expression strategies. They reported that genes with high fitness effect display lower expression noise than the rest. Following these pioneering observations, we hypothesized that genes under strong purifying selection at the protein sequence level should also be highly constrained for their expression and therefore display a lower transcriptional noise. To test this hypothesis, we correlated  $F^*$  with the ratio of non-synonymous ( $K_a$ ) to synonymous substitutions ( $K_s$ ), as measured by sequence comparison between mouse genes and their human orthologs, after discarding genes with evidence for positive selection ( $n = 5$ ). In agreement with our prediction, we report a significantly positive correlation between the  $K_a / K_s$  ratio and  $F^*$  (**Figure 4N**, Kendall's tau = 0.0557, p-value < 1.143E-05), that is, highly constrained genes (low  $K_a / K_s$  ratio) display less transcriptional noise (low  $F^*$ ) than fast evolving ones. This result demonstrates that genes encoding proteins under strong purifying selection are also more constrained on their transcriptional noise.

### **Older genes are less noisy**

Evolution of new genes was long thought to occur via duplication and modification of existing genetic material (“evolutionary tinkering”, [234]). Evidence for *de novo* gene emergence is however becoming more and more common [235,236]. *De novo* created genes undergo several optimization steps, including their integration into a regulatory network [237]. We tested whether the historical process of incorporation of new genes into pathways impacts the evolution of transcriptional noise. We used the phylostratigraphic approach of Neme & Tautz [237], which categorizes genes into 20 strata, to compute gene age and tested for a correlation with  $F^*$ . As older genes tend to be more conserved [238], more central (according to the preferential attachment model of network growth [220,239]) and more pleiotropic, we controlled for these confounding factors (Kendall's tau = -0.0663, p-value = 1.58E-37 ; partial correlation controlling for Ka / Ks ratio, centrality measures and pleiotropy level, **Figure 4O**). These results suggest that older genes are more deterministically expressed while younger genes are more noisy. While we cannot rule out that functional constraints not fully accounted for by the Ka / Ks ratio or unavailable functional annotations could explain at least partially the correlation of gene age and transcriptional noise, we hypothesise that the observed correlation results from ancient genes having acquired more complex regulation schemes through time. Such schemes include for instance negative feedback loops, which have been shown to stabilize gene expression and reduce expression noise [77,240].

### **Position in the protein network is the main driver of transcriptional noise**

In order to jointly assess the effect of network topology, epigenomic factors, Ka / Ks ratio and gene age, we modeled the patterns of transcriptional noise as a function of multiple predictive factors within the linear model framework. This analysis could be performed on a set of 2,794 genes for which values were available jointly for all variables. In order to avoid colinearity issues because some of these variables are intrinsically correlated, we performed data reduction procedures prior to modeling. For continuous variables, including Pathway and PPI network variables, Ka / Ks ratio and gene age, we conducted a principal component analysis (PCA) and used as synthetic measures the first eight principal components (PC), explaining together more than 80% of the total inertia (**Figure S2A**). The first principal component (PC1) of the PCA analysis is associated with pathway centrality measures (degree, hub score,

authority score and transitivity, **Figure S2B**). The second principal component (PC2) corresponds to PPI centrality measures (degree, hub score and betweenness), while the third component (PC3) relates to gene age and Ka / Ks ratio. The fourth component (PC4) is associated with PPI complex interactions and transitivity. PC5 and PC6 are essentially associated to betweenness and closeness of the pathway network, PC7 with PPI polymeric interactions and PC8 with pathway pleiotropy. As transcription factors and histone marks data are binary (presence / absence for each gene), we performed a logistic PCA for both type of variables [241]. For transcription factors, we selected the three first components (hereby noted TFPC), which explained 78% of deviance (**Figure S3A**). The loads on the first component (TFPC1) are all negative, meaning that TFPC1 captures a global correlation trend and does not discriminate between TFs. Tcfcp2l1 appears to be the TF with the highest correlation to TFPC1. The second component TFPC2 is dominated by TCFC (positive loading) and Oct4 (negative loading), while the third component TFPC3 is dominated by Esrrb (positive loading) and MYC, nMyc and E2F1 (negative loadings, **Figure S3B**). For histone marks, the two first components (hereby noted HistPC) explained 95% of variance and were therefore retained (**Figure S4A**). HistPC1 is dominated by marks H3K27me3 linked to gene repression (negative loadings) and HistPC2 by marks H3K4me1 and H3K4me3 linked to gene activation (positive loadings, **Figure S4A**).

We fitted a linear model with  $F^*$  as a response variable and all 13 synthetic variables as explanatory variables. We find that PC1 has a significant positive effect on  $F^*$  (**Table 3**). As the loadings of the centrality measures on PC1 are negative (**Figure S2C**), this result is consistent with our finding of a negative correlation of pathway-based centrality measure with  $F^*$ . PC3 has a highly significant negative effect on  $F^*$ , which is consistent with a negative correlation with gene age (positive loading on PC3) and a positive correlation with the Ka / Ks ratio (negative loading on PC3, **Figure S2D**). The last highly significant variable is the first principal component of the logistic PCA on histone methylation patterns, HistPC1, which has a negative effect on  $F^*$ . Because the loadings are essentially negative on HistPC1, this suggests a positive effect of methylation, in particular the repressive H3K27me3. Altogether, the linear model with all variables explained 4.01% of the total variance (adjusted  $R^2$ ). This small value indicates either that gene idiosyncrasies largely predominate over general effects, or that our estimates of transcriptional noise have a large measurement error,



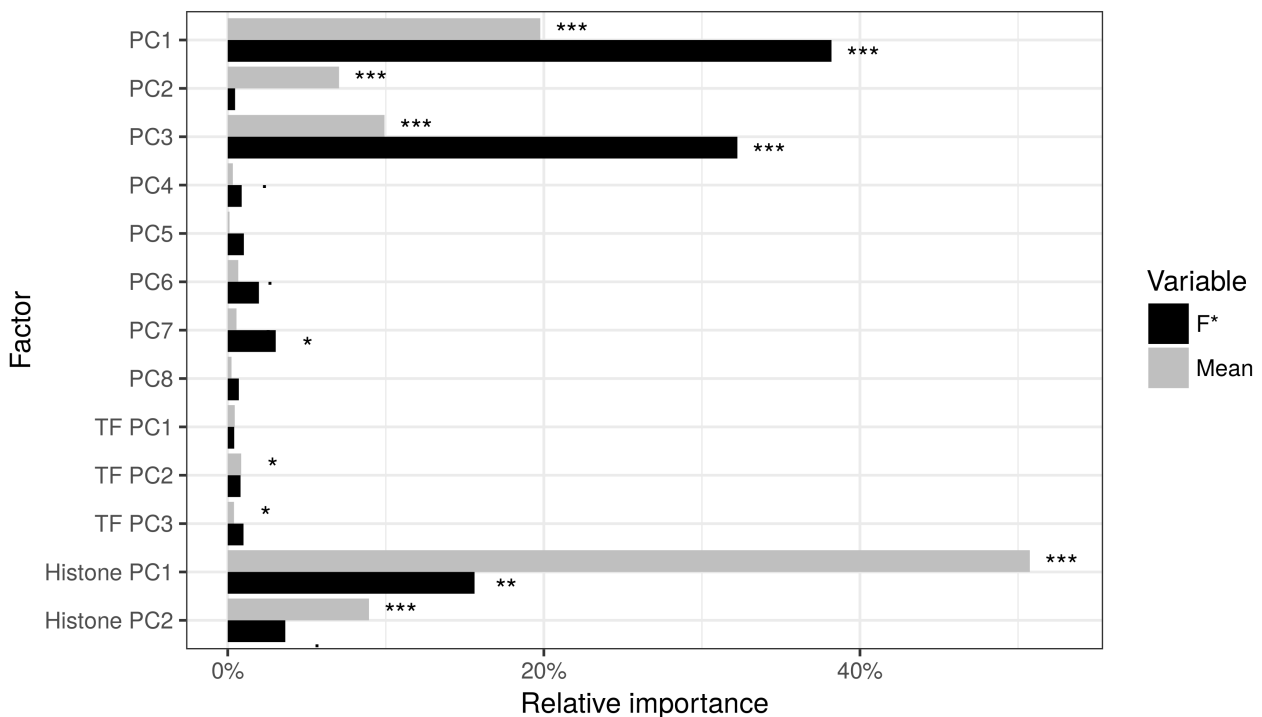
or both. To compare the individual effects of each explanatory variable, we conducted a relative importance analysis. As a mean of comparison, we fitted a similar model with mean expression as a response variable. We find that pathway centrality measures (PC1 variables) account for 38% of the explained variance, while protein constraints and gene age (PC3) account for 32%. Chromatin state (HistPC1) account for another 15% of the variance (**Figure 5**). These results contrast with the model of mean expression, where HistPC1 and HistPC2 respectively account for 51% and 9% of the explained variance, and PC1 and PC3 20% and 10% only (**Figure 5**). This suggests (1) that among all factors tested, position in protein network is the main driver of the evolution of gene-specific stochastic expression, followed by protein constraints and gene age and (2) that different selective pressures act on the mean and cell-to-cell variability of gene expression.

We further included the effect of three-dimensional organization of the genome in order to assess whether it could act as a confounding factor. We developed a correlation model allowing for genes in contact to have correlated values of transcriptional noise. The correlation model was fitted together with the previous linear model in the generalized least square (GLS) framework. This model allows for one additional parameter,  $\lambda$ , which captures the strength of correlation due to three-dimensional organization of the genome (see Methods). The estimate of  $\lambda$  was found to be 0.0016, which means that the spatial autocorrelation of transcriptional noise is low on average. This estimate is significantly higher than zero, and model comparison using Akaike's information criterion favors the linear model with three-dimensional correlation (AIC = 4880.858 vs. AIC = 4890.396 for a linear model without three-dimensional correlation). Despite the significant effect of 3D genome correlation, our results were qualitatively and quantitatively very similar to the model ignoring 3D correlation (**Table 3**).

### **Analysis of bone marrow-derived dendritic cells supports the generality of the results**

We assessed the reproducibility of our results by analyzing an additional single-cell transcriptomics data set of 95 unstimulated bone marrow-derived dendritic cells (BMDC) [242]. After filtering (see Methods), the data set consisted of 11,640 genes. Using the same normalization procedure as for the ESC data set, we nonetheless report a weak but significant

negative correlation between  $F^*$  and the mean expression, even with a degree-5 polynomial regression ( $-0.0459$ ,  $p\text{-value} < 1.13E-13$ ). This effect is due to the distribution of per-gene, between cell RPKM values being extremely skewed in this data set. In order to assess the impact of the residual correlation with the mean, we computed a value of  $F^*$  (noted  $F_R^*$ ) on a restricted dataset where the variance was between 1/8 and 8 times the mean (75% of all genes) using a quantile regression on the median instead of a linear regression. A second degree polynomial quantile regression proved to be sufficient to remove the effect of mean expression (Kendall's tau = 0.0114,  $p\text{-value} = 0.1125$ ) on this restricted data set. As all results were consistent when using the  $F_R^*$  and  $F^*$  measures, we only discuss here results obtained with  $F^*$  and refer to **Supplementary Data 1** for detailed results obtained with the  $F_R^*$  measure.



**Figure 5: Relative importance of explanatory factors on mean gene expression and  $F^*$ .** Significance codes refer to ANOVA test of variance, \*\*\* < 0.001 < \*\* < 0.01 < \* < 0.05 < . < 0.1.

We report a highly significant positive correlation between  $F^*$  values measured on the 8,792 genes with expression in both data sets, suggesting that cell-to-cell variance in gene expression is to a large extent conserved among the two cell types (Kendall's tau = 0.1289,  $p\text{-value} < 2.2E-16$ , **Figure S6A**). GO terms or reactome pathways enrichment analyses reveal

less significant but consistent terms with the ESM analysis: the high  $F^*$  gene set did not show any significantly enriched GO term or reactome pathway (FDR set to 1%) and the low  $F^*$  gene set revealed RNA-binding as a significantly enriched molecular function, as well as 21 enriched pathways (**Figure S7**). In agreement with results from the ESM analysis, many of the most significant enriched pathways relate to gene expression, including translation and splicing. Interestingly, the two most significant pathways, however, are “Vesicle-mediated transport” and “Membrane trafficking”, two essential pathways for the functioning of dendritic cells. Analyses of network centrality measures also generally show consistent results with the ESC data set, more central genes displaying reduced gene expression noise (**Figure S6B-N, Table S1**). Quantitative differences consists of PPI betweenness, as well as pathway closeness and betweenness are highly significantly negatively correlated with  $F^*$  while they were only weakly or non-significant with the ESC data set. The only discrepancies that we report between the two data sets relate to pathway level statistics. Pathway size appears to be significantly positively correlated with mean  $F^*$ , while it was negatively correlated on the ESC data set, yet with a comparatively higher p-value. Similarly pathway diameter is significantly positively correlated with mean  $F^*$  in the BMDC data set, while it was not significant with the ESC data. We currently have no hypothesis to explain this particular discrepancy. While these results support the generality of our observations, they also illustrate that in details, the fine structure of translational noise may vary in a cell type-specific manner.

We fitted linear models as for the embryonic stem cell (ESC) data set, with the exception that no epigenomic data was available for this cell type. Data reduction was performed using a principal component analysis, with the eight first principal components explaining 81% of the total deviance (**Figure S8A**). We report consistent results with the ESC analysis, with all major effects similar in direction and intensity, highlighting the impact of network centrality measures on expression noise (**Table S2**). With the BMDC data, however, the second principal component PC2 which is associated with PPI centrality measures (**Figure S8B**) appears to have a significant negative impact on  $F^*$ , while it was not significant with the ESC dataset. As the loading of the PPI centrality measures are positive on PC2, this is consistent with central genes having a lower transcriptional noise as for the pathway network metrics (**Figure S8C**). When taking 3D genome correlations into account, we estimated a low

correlation coefficient as for the ESC dataset ( $\lambda = 0.0004$ ), and the AIC favored the model without correlation in this case. Relative importance analysis revealed that network centrality measures contributed most to the explained variance (48% and 21% for PC1 and PC2 respectively), while the contribution of protein constraints and gene age (PC3) was 24%.

### **Biological, not technical noise is responsible for the observed patterns**

The variance in gene expression measured from single-cell transcriptomics is a combination of biological and technical variance. While the two sources of variance are a priori independent, gene-specific technical variance has been observed in micro-array experiments [243] making a correlation of the two types of variance plausible. If similar effects also affect RNA-Seq experiments, technical variance could be correlated to gene function and therefore act as a covariate in our analyses. In order to assess whether this is the case, we used the dataset of Shalek et al [217], which contains both single-cell transcriptomics and 3 replicates of 10,000 pooled-cell RNA sequencing. In traditional RNA sequencing, which is typically performed on pooled populations of several thousands of cells, biological variance is averaged out so that the resulting measured variance between replicates is essentially the result of technical noise. We computed the mean and variance in expression of each gene across the three populations of cells. By plotting the variance versus the mean in log-space, we were able to compute a “technical”  $F_t^*$  value for each gene (see Methods). We fitted linear models as for the single cell data, using  $F_t^*$  instead of  $F^*$ . We report that no variable had a significant effect on  $F_t^*$  (**Table S3**). In addition, there was no enrichment of the lower 10<sup>th</sup>  $F_t^*$  percentile for any particular pathway or GO term. The upper 90<sup>th</sup> percentile showed no GO term enrichment, but four pathways appeared to be significant: “Chromosome maintenance” (adjusted p-value = 0.0043), “Polymerase switching on the C-strand of the telomere” (adjusted p-value = 0.0062), “Polymerase switching” (adjusted p-value = 0.0062) and “Leading strand synthesis” (adjusted p-value = 0.0062), which relate to DNA replication. While it is unclear why genes involved in these pathways would display higher technical variance in RNA sequencing, these results strikingly differ from our analyses of single cell RNA sequencing and therefore suggest that technical variance does not act as a confounding factor in our analyses.

Because only three replicates were available in the pooled RNA-Seq data set, we asked whether the resulting estimate of mean and variance in expression is accurate enough to allow proper inference of noise and its correlation with other variables. We conducted a jackknife procedure where we sampled the original cells from the ESC data set and re-estimated  $F^*$  for each sample. We tested combinations of 3, 5, 10 and 15 cells, with 1,000 samples in each case. In each sample, we computed  $F^*$  with the same procedure as for the complete data set, and fitted a linear model with all 13 synthetic variables. For computational efficiency, we did not include 3D correlation in this analysis. We compute for each variable the number of samples where the effect is significant at the 5% level and has the same sign as in the model fitted on the full data set. We find that the model coefficients are very robust to the number of cells used (**Figure S9A**) and that 3 cells are enough to infer the effect of the PC1 and PC3 variables, the most significant in our analyses. Two main conclusions can be drawn from this jackknife analysis: (1) that the lack of significant effect of our explanatory variables on technical noise is not due to the low number of replicates used to compute the mean and variance in expression and (2) that our conclusions are very robust to the actual cells used in the analysis, ruling out drop-out and amplification biases as possible source of errors [244].

### **3. DISCUSSION:**

Throughout this work, we provided the first genome-wide evolutionary and systemic study of transcriptional noise, using mouse cells as a model. We have shown that transcriptional noise correlates with functional constraints both at the level of the gene itself via the protein it encodes, but also at the level of the pathway(s) the gene belongs to. We further discuss here potential confounding factors in our analyses and argue that our results are compatible with selection acting to reduce noise-propagation at the network level.

In this study, we exhibited several factors explaining the variation in transcriptional noise between genes. While highly significant, the effects we report are of small size, and a complex model accounting for all tested sources of variation only explains a few percent of the total observed variance. There are several possible explanations for this reduced explanatory power: (1) transcriptional noise is a proxy for noise in gene expression, at which

selection occurs (**Figure 1**). As transcriptional noise is not randomly distributed across the genome, it must constitute a significant component of expression noise, in agreement with previous observations [82,196]. Translational noise, however, might constitute an important part of the expression noise and was not assessed in this study. (2) Gene expression levels were assessed on embryonic stem cells in culture. Such an experimental system may result in gene expression that differs from that in natural conditions under which natural selection acted. (3) Functional annotations, in particular pathways and gene interactions are incomplete, and network-based measures have most likely large error rates. (4) While the newly introduced  $F^*$  measure allowed us to assess the distribution of transcriptional noise independently of the average mean expression, it does not capture the full complexity of SGE. Explicit modeling, for instance based in the Beta-Poisson model [245] is a promising avenue for the development of more sophisticated quantitative measures.

In a pioneering study, Fraser et al [88], followed by Shalek et al [217], demonstrated that essential genes whose deletion is deleterious, and genes encoding subunits of molecular complexes as well as housekeeping genes display reduced gene expression noise. Our findings go beyond these early observations by providing a statistical assessment of the joint effect of multiple explanatory factors. Our analyses reveal that network centrality measures are the explanatory factors that explained the most significant part of the distribution of transcriptional noise in the genome. Network-based statistics were first tested by Li et al [208] using PPI data in Yeast. While we are able to extend these results to mouse cells, we show that more detailed annotation as provided by the Reactome database lead to new insights into the selective forces acting on expression noise. Our results suggest that “pathways” constitute a relevant systemic level of organisation, at which selection can act and drive the evolution of SGE at the gene level. This multi-level selection mechanism, we propose, can be explained by selection against noise propagation within networks. It has been experimentally demonstrated that expression noise can be transmitted from one gene to another gene with which it is interacting [86]. Large noise at the network level is deleterious [76] but each gene does not contribute equally to it, thus the strength of selective pressure against noise varies among genes in a given network. We have shown that highly connected, “central” proteins typically display reduced transcriptional noise. Such nodes are likely to constitute key players in the flow of noise in intra-cellular networks as they are more likely to

transmit noise to other components. In accordance with this hypothesis, we find genes with the lowest amount of transcriptional noise to be enriched for top-level functions, in particular involved in the regulation of other genes.

These results have several implications for the evolution of gene networks. First, this means that new connections in a network can potentially be deleterious if they link genes with highly stochastic expression. Second, distinct selective pressures at the “regulome” and “interactome” levels (**Figure 1**) might act in opposite direction. We expect genes encoding highly connected proteins to have more complex regulation schemes, in particular if their proteins are involved in several biological pathways. In accordance, several studies demonstrated that expression noise of a gene positively correlates with the number of transcription factors controlling its regulation [215], a correlation that we also find significant in the data set analyzed in this work. Central genes, while being under negative selection against stochastic behavior, are then more likely to be controlled by numerous transcription factors which increase transcriptional noise. As a consequence, if the number of connections at the interactome level is correlated with the number of connections at the regulome level, we predict the existence of a trade-off in the number of connections a gene can make in a network. Alternatively, highly connected genes might evolve regulatory mechanisms allowing them to uncouple these two levels: negative feedback loops, for instance, where the product of a gene down-regulates its own production have been shown to stabilize expression and significantly reduce stochasticity [211,240,246]. We therefore predict that negative feedback loops are more likely to occur at genes that are more central in protein networks, as they will confer greater resilience against high SGE, which is advantageous for this class of genes.

Our results enabled the identification of possible selective pressures acting on the level of stochasticity in gene expression. The mechanisms by which the amount of stochasticity can be controlled remain however to be elucidated. We evoked the existence of negative feedback loops which reduce stochasticity and the multiplicity of upstream regulator which increase it. Recent work by Wolf et al [247] and Metzger et al [248] add further perspective to this scheme. Wolf and colleagues found that in *Escherichia coli* noise is higher for natural than experimentally evolved promoters selected for their mean expression level. They

hypothesized that higher noise is selectively advantageous in case of changing environments. On the other hand, Metzger and colleagues performed mutagenesis experiments and found signature of selection for reduced noise in natural populations of *Saccharomyces cerevisiae*. These seemingly opposing results combined with our observations provide additional evidence that the amount of stochasticity in the expression of single genes has an optimum, as high values are deleterious because of noise propagation in the network, whilst lower values, which result in reduced phenotypic plasticity, are suboptimal in case of dynamic environment.

## **4. CONCLUSION:**

Using a new measure of transcriptional noise, our results demonstrate that the position of the protein in the interactome is a major driver of selection against stochastic gene expression. As such, transcriptional noise is an essential component of the phenotype, in addition to the mean expression level and the actual sequence and structure of the encoded proteins. This is currently an under-appreciated phenomenon, and gene expression studies that focus only on the mean expression of genes may be missing key information about expression diversity. The study of gene expression must consider changes in noise in addition to change in mean expression level as a putative explanation for adaptation. Further work aiming to unravel the exact structure of the regulome is however needed in order to fully understand how transcriptional noise is generated or inhibited.

## **5. METHODS:**

### **Single-cell gene expression data set**

We used the dataset generated by Sasagawa et al. [249] retrieved from the Gene Expression Omnibus repository (accession number GSE42268). We analyzed expression data corresponding to embryonic stem cells in G1 phase, for which more individual cells were sequenced. A total of 17,063 genes had non-zero expression in at least one of the 20 single cells. Similar to Shalek et al [242], a filtering procedure was performed where only genes whose expression level satisfied  $\log(\text{FPKM}+1) > 1.5$  in at least one single cell were kept for



further analyses. This filtering step resulted in a total of 13,660 appreciably expressed genes for which transcriptional noise was evaluated.

### Measure of transcriptional noise

The expression mean ( $\mu$ ) and variance ( $\sigma^2$ ) of each gene over all single cells were

computed. We measured stochastic gene expression as the ratio  $F^* = \frac{\sigma^2}{\hat{\sigma}^2(\mu)}$ , where

$\hat{\sigma}^2(\mu)$  is the expected variance given the mean expression. In order to compute  $\hat{\sigma}^2(\mu)$ , we performed several polynomial regressions with  $\log(\sigma^2)$  as a function of  $\log(\mu)$ , with degrees between 1 and 5. We then tested the resulting  $F^*$  measures for residual correlation with mean expression using Kendall's rank correlation test. We find that a degree-3 polynomial regression was sufficient to remove any residual correlation with  $F^*$  (Kendall's tau = 0.0037, p-value = 0.5217).  $F^*$  can be seen as a general expression for the Fano factor and noise measure: when using a polynome of degree 1, the expression of  $F^*$  becomes

$F^* = \frac{\sigma^2}{\exp(a+b \cdot \log(\mu))} = \frac{\sigma^2}{\exp(a) \cdot \mu^b}$ , and is therefore equivalent to the Fano factor when  $a = 0$  and  $b = 1$ , and equivalent to noise when  $a = 0$  and  $b = 2$ .

### Genome architecture

The mouse proteome from Ensembl (genome version: mm9) was used in order to get coordinates of all genes. The Hi-C dataset for embryonic stem cells (ES) from Dixon et al [213] was used to get three-dimensional domain information. Two genes were considered in proximity in one dimension (1D) if they are on the same chromosome and no protein-coding gene was found between them. The primary distance (in number of nucleotides) between their midpoint coordinates was also recorded as 1D a distance measure between the genes. Two genes were considered in proximity in three dimensions (3D) if the normalized contact number between the two windows the genes belong was non-null. Two genes belonging to the same window were considered in proximity. We further computed the relative difference of stochastic gene expression between two genes by computing the ratio

$(F_2^* - F_1^*) / (F_2^* + F_1^*)$ . For each chromosome, we independently tested if there was a correlation between the primary distance and the relative difference in stochastic gene expression with a Mantel test, as implemented in the `ade4` package [250]. In order to test whether genes in proximity (1D and 3D) had more similar transcriptional noise than distant genes, we contrasted the relative differences in transcription noise between pairs of genes in proximity and pairs of distant genes. As we test all pairs of genes, we performed a randomization procedure in order to assess the significance of the observed differences by permuting the rows and columns in the proximity matrices 10,000 times. Linear models accounting for spatial interactions with genes were fitted using the generalized least squares (GLS) procedure as implemented in the “nlme” package for R. A correlation matrix between all tested genes was defined as  $G = \{g_{i,j}\}$ , where  $g_{i,j}$  is the correlation between genes  $i$  and  $j$ . We defined  $g_{i,j} = 1 - \exp(-\lambda \delta_{i,j})$ , where  $\delta_{i,j}$  takes 1 if genes  $i$  and  $j$  are in proximity, 0 otherwise (binary model). Alternatively,  $\delta_{i,j}$  can be defined as the actual number of contacts between the two 20 kb regions (as defined by Dixon et al) the genes belong to (proportional model). Parameter  $\lambda$  was estimated jointly with other model parameters, it measures the strength of the genome “spatial” correlation. Models were compared using Akaike’s information criterion (AIC). We find that the proportional correlation model fitted the data better and therefore selected it for further analyses.

### **Transcription factors and histone marks**

Transcription factor (TF) mapping data from the Ensembl regulatory build [214] were obtained via the `biomaRt` package for R. We used the Grch37 build as it contained data for stem cells epigenomes. Genes were considered to be associated with a given TF when at least one binding evidence was present in the 3 kb upstream flanking region. Transcription factors associated with less than 5 genes for which transcriptional noise could be computed were not considered further. A similar mapping was performed for histone marks by counting the evidence of histone modification in the 3 kb upstream and downstream regions of each gene. A logistic principal component analysis was conducted on the resulting binary contingency tables using the `logisticPCA` package for R [241], for TF and histone marks separately. Principal components were used to define synthetic variables for further analyses.

## **Biological pathways, protein-protein interactions and network topology**

We defined genes either in the top 10% least noisy or in the top 10% most noisy as candidate sets and used the Reactome PA package [251] to search the mouse Reactome database for overrepresented pathways with a 1% false discovery rate.

Centrality measures were computed using a combination of the “igraph” [252] and “graphite” [253] packages for R. As the calculation of assortativity does not handle missing data (that is, nodes of the pathway for which no value could be computed), we computed assortativity on the sub-network with nodes for which data were available. Reactome centrality measures could be computed for a total of 4,454 genes with expression data.

Protein-protein interactions (PPI) were retrieved from the iRefIndex database [254] using the iRefR package for R [255]. Interactions were converted to a graph using the dedicated R functions in the package, and the same methods were used to compute centrality measures as for the pathway analysis. Because the PPI-based graph was not oriented, authority scores were not computed for this data (as this gave identical results to hub scores). Furthermore, as most genes are part of a single graph structure in the case of PPI interactions, closeness values were not further analysed as they were virtually identical for all genes.

## **Gene Ontology Enrichment**

Eight thousand three hundreds and twenty five out of the 13,660 genes were associated with Gene Ontology (GO) terms. We tested genes for GO terms enrichment at both ends of the F\* spectrum using the same threshold percentile of 10% low / high noise genes as we did for the Reactome analysis. We carried out GO enrichment analyses using two different algorithms: “Parent-child” [256] and “Weight01”, a mixture of two algorithms developed by Alexa et al [257]. We kept only the terms that appeared simultaneously on both Parent-child and Weight01 under 1% significance level, controlling for multiple testing using the FDR method [258].

## **Sequence divergence**

The Ensembl's Biomart interface was used to retrieve the proportion of non-synonymous (Ka)

and synonymous ( $K_s$ ) divergence estimates for each mouse gene relative to the human ortholog. This information was available for 13,136 genes.

### **Gene Age**

The relative taxonomic ages of the mouse genes have been computed and is available in the form of 20 Phylostrata [237]. Each Phylostratum corresponds to a node in the phylogenetic tree of life. Phylostratum 1 corresponds to “All cellular organisms” whereas Phylostratum 20 corresponds to “*Mus musculus*”, with other levels in between. We used this published information to assign each of our genes to a specific Phylostratum and used this as a relative measure of gene age: Age = 21 - Phylostratum, so that an age of 1 corresponds to genes specific to *M. musculus* and genes with an age of 20 are found in all cellular organisms.

### **Linear modeling**

We simultaneously assessed the effect of different factors on transcriptional noise by fitting linear models to the gene-specific  $F^*$  estimates. To avoid collinearity issues of intrinsically correlated explanatory variables, we conducted a data reduction procedure using multivariate analysis. We used variants of principal component analysis (PCA) on explanatory variables in three groups: network centrality measures,  $K_a / K_s$  and gene age with standard PCA, transcription factor binding evidence and histone methylation patterns using logistic PCA, a generalization of PCA for binary variables [241]. In each case, we used the most representative components (totaling at least 75% of the total deviance) as synthetic variables. PCA analysis was conducted using the *ade4* package for R [250], logistic PCA was performed using the *logisticPCA* package [241].

We built a linear model with  $F^*$  as a response variable and thirteen synthetic variables as explanatory variables. As the synthetic variables are principal components, they are orthogonal by construction. The fitted model displayed significant departure to normality and was further transformed using the Box-Cox procedure (“*boxcox*” function from the *MASS* package for R [182]). Residues of the selected model had normal, independent residue distributions (Shapiro-Wilk test of normality,  $p$ -value = 0.121, Ljung-Box test of independence,  $p$ -value = 0.2061) but still displayed significant heteroscedasticity (Harrison-

McCabe test,  $p$ -value = 0.003). In order to ensure that this departure from the Gauss-Markov assumptions does not bias our inference, we used the “robcov” function of the “rms” package in order to get robust estimates of the effect significance [259]. Relative importance of each explanatory factor was assessed using the method of Lindeman, Merenda and Gold [260] as implemented in the R package “relaimpo”. The significance of the level of variance explained by each factor was computed using standard ANOVA procedure.

### **Additional data sets**

The aforementioned analyses were additionally conducted on the bone marrow-derived dendritic cells data set of Shalek et al [242]. Following the filtering procedure established by the authors in the original paper, genes which did not satisfy the condition of being expressed by an amount such that  $\log(\text{TPM}+1) > 1$  in at least one of the 95 single cells were further discarded, where TPM stands for transcripts per million. This cut-off threshold resulted in 11,640 genes being kept for investigation. The rest of the analyses was conducted in the same way as for the ESM data set.

### **Jackknife procedure**

A jackknife procedure was conducted in order to assess (1) the robustness of our results to the choice of actual cells used to estimate mean and variance in gene expression and (2) the power of the pooled RNA-seq analysis for which only three replicates were available. This analysis was conducted by sampling 3, 5, 10 and 15 of the original 20 single cells of the ESM data set [249], 1,000 times in each case. The exact same analysis was conducted on each random sample as for the complete data set, and model coefficients and their associated  $p$ -values were recorded.

### **Acknowledgements**

The authors would like to thank Rafiq Neme-Garrido, Frederic Bartels and Estelle Renaud for fruitful discussions about this work, Andrew Landgraf for help with the logistic PCA analysis as well as Diethard Tautz for comments on an earlier version of this manuscript. JYD acknowledges funding from the Max Planck Society. This work was supported by the

German Research Foundation (DFG), within the priority program (SPP) 1590.

## DECLARATION:

This chapter is published in the journal Genetics with DOI:

<https://doi.org/10.1534/genetics.117.300467>

## SUPPLEMENTAL MATERIAL:

The following supplemental material is available in the digital archive

**Table S1:** Correlation of transcriptional noise with genes centrality measures and pleiotropy for the bone marrow-derived dendritic cells data set. Legends as in **Table 2**.

**Table S2:** Linear models of transcriptional noise with genomic factors for the bone marrow-derived dendritic cells data set. Legend as in Table 4.

**Table S3:** Linear model of transcriptional noise with genomic factors with pooled RNA-Seq data. Legend as in Table 4.

**Figure S1: Impact of genome organization on the distribution of transcriptional noise.**

The x-axis shows the mean relative difference in transcriptional noise. Vertical lines show observed values and histograms the distribution over 10,000 permutations (see Methods). Left panel: distribution for neighbor genes along the genome. Right panel: distribution for genes in contact in three-dimensions.

**Figure S2: Principal component analysis of pathways centrality measures.** **A**, Proportion of deviance explained by models with 1, 2, etc. principal components. **B**, Contributions, computed as proportion of deviance, of each input variable to each principal component. **C**, Loadings of each variable on the 2 first components. **D**, Loadings of each variable on the 3rd and 4th principal components.

**Figure S3: Logistic principal component analysis of transcription factor binding evidences.** **A**, Proportion of deviance explained by models with 1, 2, etc. principal components. **B**, Contributions, computed as proportion of deviance, of each input variable to each principal component. **C**, Loadings of each variable on the 2 first components. **D**, Loadings of each variable on the 2nd and 3rd principal components.

**Figure S4: Logistic principal component analysis of histone marks.** **A**, Proportion of deviance explained by models with 1, 2, etc. principal components. **B**, Contributions, computed as proportion of deviance, of each input variable to each principal component. **C**, Loadings of each variable on the 2 first components.

**Figure S5: Assortativity in networks.** **A**, Distribution of assortativity values for hub scores. **B**, Distribution of assortativity values for F\*. **C**, Assortativity for F\* and hub scores are plotted against each other. Solid lines represent linear regressions fitted on pathways with negative or positive hub score assortativity, respectively. Dashed line represents a linear regression fitted on all data.

**Figure S6: Factors driving stochastic gene expression in the bone marrow-derived dendritic cells data set.** Legends as in **Figure 4**.

**Figure S7: Enriched pathways in the low-noise gene set of the bone marrow-derived dendritic cells data set.**

**Figure S8: Principal component analysis of pathways centrality measures of the bone marrow-derived dendritic cells data set.** Legends as in **Figure S2**.

**Figure S9: Robustness and power analysis.** A jackknife procedure was conducted by fitted linear models with all explanatory variables on a subset of cells taken randomly (**x-axis**). **A**, estimated coefficient of each effect. **B**, proportion of simulations where the coefficient is significant at the 5% level. Filled bars correspond to significant effect when the complete data set is used. PC: principal component. PPI: protein-protein interactions. TF: transcription factors.

## REFERENCES:

1. Coveney P, Highfield R. *The arrow of time: a voyage through science to solve time's greatest mystery*. 2015.
2. Solé R, Goodwin B. *Signs of life: how complexity pervades biology*. New York, NY: Basic Books; 2002.
3. May RM. Simple mathematical models with very complicated dynamics. *Nature*. 1976;261: 459–467. doi:10.1038/261459a0
4. Scheinerman ER. *Invitation to Dynamical Systems*. [Internet]. Dover Publications; 2013. Available: <http://www.totalboox.com/book/id-1437146502651819388>
5. Kauffman S. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Oxford University Press; 1996.
6. Kauffman S. *Evolution beyond Newton, Darwin and entailing law. Beyond mechanism: putting life back into biology* Plymouth, UK: Lexington Books. 2013; 1–24.
7. Wangersky PJ. Lotka-Volterra Population Models. *Annual Review of Ecology and Systematics*. 1978;9: 189–218.
8. Bacaër N. Lotka, Volterra and the predator–prey system (1920–1926). *A Short History of Mathematical Population Dynamics*. London: Springer London; 2011. pp. 71–76. doi:10.1007/978-0-85729-115-8\_13
9. Barroso GV, Luz DR. On the limits of complexity in living forms. *J Theor Biol*. 2015;379: 89–90. doi:10.1016/j.jtbi.2015.04.032
10. Gould SJ. *Wonderful life: the Burgess Shale and the nature of history*. New York: Norton; 2007.
11. Blount ZD, Lenski RE, Losos JB. Contingency and determinism in evolution: Replaying life's tape. *Science*. 2018;362: eaam5979. doi:10.1126/science.aam5979
12. Casillas S, Barbadilla A. Molecular Population Genetics. *Genetics*. 2017;205: 1003–1035. doi:10.1534/genetics.116.196493
13. Wright S. Evolution in Mendelian Populations. *Genetics*. 1931;16: 97–159.
14. Fisher RA. *The genetical theory of natural selection*. 2017.
15. Hey J, Hey J. The neutralist, the fly and the selectionist. *Trends in Ecology & Evolution*. 1999;14: 35–38. doi:10.1016/S0169-5347(98)01497-9
16. Nei M. Selectionism and Neutralism in Molecular Evolution. *Mol Biol Evol*. 2005;22: 2318–2342. doi:10.1093/molbev/msi242



17. S?ll T, Bengtsson BO. *Understanding Population Genetics*. John Wiley & Sons; 2017.
18. Lewontin RC, Hubby JL. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*. 1966;54: 595–609.
19. Charlesworth B, Charlesworth D. Population genetics from 1966 to 2016. *Heredity*. 2016; 1–8. doi:10.1038/hdy.2016.55
20. Evolutionary Rate at the Molecular Level | Nature [Internet]. [cited 20 Jan 2019]. Available: <https://www.nature.com/articles/217624a0>
21. Zuckerkandl E, Pauling L. Evolutionary Divergence and Convergence in Proteins. In: Bryson V, Vogel HJ, editors. *Evolving Genes and Proteins*. Academic Press; 1965. pp. 97–166. doi:10.1016/B978-1-4832-2734-4.50017-6
22. Ohta T. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*. 1992;23: 263–286. doi:10.1146/annurev.es.23.110192.001403
23. Kern AD, Hahn MW. The Neutral Theory in Light of Natural Selection. *Mol Biol Evol*. 2018;35: 1366–1371. doi:10.1093/molbev/msy092
24. Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. *Evolution*. 2019;73: 111–114. doi:10.1111/evo.13650
25. Stephan W. Selective Sweeps. *Genetics*. 2019;211: 5–13. doi:10.1534/genetics.118.301319
26. Payseur BA. Genetic Links between Recombination and Speciation. *PLOS Genetics*. 2016;12: e1006066. doi:10.1371/journal.pgen.1006066
27. Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature reviews Genetics*. 2013;14: 262–74. doi:10.1038/nrg3425
28. Lenormand T, Roze D, Rousset F. Stochasticity in evolution. *Trends in Ecology and Evolution*. 2009;24: 157–165. doi:10.1016/j.tree.2008.09.014
29. Tran TD, Hofrichter J, Jost J. An introduction to the mathematical structure of the Wright–Fisher model of population genetics. *Theory Biosci*. 2013;132: 73–82. doi:10.1007/s12064-012-0170-3
30. Kimura M, Crow JF. The number of alleles that can be maintained in a finite population. *Genetics*. 1964;49: 725–738.
31. Kimura M. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. *Genetics*. 1969;61: 893–903.
32. Walsh B, Lynch M. *Evolution and selection of quantitative traits*. 2018.

33. Nordborg M. Coalescent Theory. *Handbook of Statistical Genetics: Third Edition*. 2008;2: 843–877. doi:10.1002/9780470061619.ch25
34. Watterson GA. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*. 1975;7: 256–276. doi:10.1016/0040-5809(75)90020-9
35. Hein J, Schierup M, Wiuf C. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory*. Oxford, New York: Oxford University Press; 2004.
36. Kingman J. Genealogy Populations. *Journal of applied probability*. 1982;19: 27–43.
37. Tajima F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*. 1983;105: 437–460. doi:6628982
38. Hudson RR. Testing the Constant-Rate Neutral Allele Model with Protein Sequence Data. *Evolution*. 1983;37: 203–217. doi:10.1111/j.1558-5646.1983.tb05528.x
39. Hudson RR. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*. 1983;23: 183–201. doi:10.1016/0040-5809(83)90013-8
40. Kaplan N, Hudson RR. The use of sample genealogies for studying a selectively neutral m-loci model with recombination. *Theoretical Population Biology*. 1985;28: 382–396. doi:10.1016/0040-5809(85)90036-X
41. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*. 1985;111: 147–164.
42. Hunter N. Meiotic Recombination: The Essence of Heredity. *Cold Spring Harb Perspect Biol*. 2015;7: a016618. doi:10.1101/cshperspect.a016618
43. Otto SP, Lenormand T. Evolution of sex: Resolving the paradox of sex and recombination. *Nature Reviews Genetics*. 2002;3: 252–261. doi:10.1038/nrg761
44. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews Genetics*. 2008;9: 477–85. doi:10.1038/nrg2361
45. Kauppi L, Jeffreys AJ, Keeney S. Where the crossovers are: recombination distributions in mammals. *Nature reviews Genetics*. 2004;5: 413–424. doi:10.1038/nrg1346
46. †The International HapMap Consortium. The International HapMap Project. *Nature*. 2003;426: 789–796. doi:10.1038/nature02168
47. Griffiths RC, Marjoram P. An ancestral recombination graph. *Progress in population genetics and human evolution*. Springer; 1997. pp. 257–270. Available: <https://research.monash.edu/en/publications/an-ancestral-recombination-graph>
48. Griffiths RC, Marjoram P. Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol*. 1996;3: 479–502. doi:10.1089/cmb.1996.3.479

49. Wiuf C, Hein J. Recombination as a point process along sequences. *Theoretical population biology*. 1999;55: 248–59. doi:10.1006/tpbi.1998.1403
50. Griffiths RC. Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology*. 1981;19: 169–186. doi:10.1016/0040-5809(81)90016-2
51. Hudson RR. Two-Locus Sampling Distributions and Their Application. *Genetics*. 2001;159: 1805–1817.
52. Kamm JA, Spence JP, Chan J, Song YS. Two-Locus Likelihoods Under Variable Population Size and Fine-Scale Recombination Rate Estimation. *Genetics*. 2016;203: 1381–1399. doi:10.1534/genetics.115.184820
53. McVean GAT, Cardin NJ. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2005;360: 1387–1393. doi:10.1098/rstb.2005.1673
54. Marjoram P, Wall JD. Fast “coalescent” simulation. *BMC Genetics*. 2006;7: 16–16. doi:10.1186/1471-2156-7-16
55. Wakeley J. Coalescent theory has many new branches. *Theoretical Population Biology*. 2013;87: 1–4. doi:10.1016/j.tpb.2013.06.001
56. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Mol Biol Evol*. doi:10.1093/molbev/msy228
57. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011;475: 493–496. doi:10.1038/nature10231
58. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*. 2014;46: 919–925. doi:10.1038/ng.3015
59. Palamara PF, Terhorst J, Song YS, Price AL. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*. 2018;50: 1311–1317. doi:10.1038/s41588-018-0177-x
60. Raynal L, Marin J-M, Pudlo P, Ribatet M, Robert CP, Estoup A. ABC random forests for Bayesian parameter inference. *Bioinformatics*. 2018; doi:10.1093/bioinformatics/bty867
61. Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends Genet*. 2018;34: 301–312. doi:10.1016/j.tig.2017.12.005
62. Schrider DR, Ayroles J, Matute DR, Kern AD. Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLOS Genetics*. 2018;14: e1007341. doi:10.1371/journal.pgen.1007341
63. Yang Z. *Molecular evolution: a statistical approach* [Internet]. 2014. Available: <http://dx.doi.org/10.1093/acprof:oso/9780199602605.001.0001>
64. Pearl J, Mackenzie D. *The book of why: the new science of cause and effect*. 2018.

65. Ellegren H, Galtier N. Determinants of genetic diversity. *Nat Rev Genet.* 2016;17: 422–433. doi:10.1038/nrg.2016.58
66. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 2009;5: e1000695. doi:10.1371/journal.pgen.1000695
67. Edge MD, Coop G. Reconstructing the History of Polygenic Scores Using Coalescent Trees. *Genetics.* 2019;211: 235–262. doi:10.1534/genetics.118.301687
68. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* 2014.
69. López-Maury L, Marguerat S, Bähler J. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics.* 2008;9: 583–593. doi:10.1038/nrg2398
70. Bell G. Evolutionary Rescue. *Annu Rev Ecol Evol Syst.* 2017;48: 605–627. doi:10.1146/annurev-ecolsys-110316-023011
71. Healey D, Axelrod K, Gore J. Negative frequency-dependent interactions can underlie phenotypic heterogeneity in a clonal microbial population. *Molecular Systems Biology.* 2016;12: 877. doi:10.15252/msb.20167033
72. Spudich JL, Koshland DE Jr. Non-genetic individuality: chance in the single cell. *Nature.* 1976: 467–471.
73. Otwinowski J, Nemenman I. Genotype to Phenotype Mapping and the Fitness Landscape of the *E. coli* lac Promoter. *PLOS ONE.* 2013;8: e61570. doi:10.1371/journal.pone.0061570
74. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell.* 2017;169: 1177–1186. doi:10.1016/j.cell.2017.05.038
75. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic Gene Expression in a Single Cell. *Science.* 2002;297: 1183–1186.
76. Barkai N, Leibler S. Circadian clocks limited by noise. *Nature.* 1999;403: 267–268.
77. Thattai M, Oudenaarden AV. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America.* 2001;98: 8614–8619.
78. Furlong LI. Human diseases through the lens of network biology. *Trends in Genetics.* 2013;29: 150–159. doi:10.1016/j.tig.2012.11.004
79. Davidson EH. Emerging properties of animal gene regulatory networks. *Nature.* 2010;468: 911–920. doi:10.1038/nature09645
80. Barabási A-L, Oltvai ZN. Network biology: understanding the cell’s functional

- organization. *Nature reviews Genetics*. 2004;5: 101–113. doi:10.1038/nrg1272
81. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature*. 1999;402: C47–C52. doi:10.1038/35011540
  82. Blake WJ, Kærn M, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. *Nature*. 2003;422: 633–637. doi:10.1038/nature01546
  83. Elowitz MB. Stochastic Gene Expression in a Single Cell. 2014;1183: 1183–1187. doi:10.1126/science.1070919
  84. Wang J, Atolia E, Hua B, Savir Y, Escalante-Chong R, Springer M. Natural Variation in Preparation for Nutrient Depletion Reveals a Cost-Benefit Tradeoff. *PLoS biology*. 2015;13: e1002041–e1002041. doi:10.1371/journal.pbio.1002041
  85. Healey D, Axelrod K, Gore J, Acar M, Becskei A, Oudenaarden A van, et al. Negative frequency-dependent interactions can underlie phenotypic heterogeneity in a clonal microbial population. *Molecular Systems Biology*. 2016;12: 877–877. doi:10.15252/msb.20167033
  86. Pedraza JM, van Oudenaarden A. Noise propagation in gene networks. *Science*. 2005;307: 1965–1969. doi:10.1126/science.1109090
  87. Draghi J, Whitlock M. Robustness to noise in gene expression evolves despite epistatic constraints in a model of gene networks. *Evolution*. 2015;69: 2345–2358. doi:10.1111/evo.12732
  88. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB. Noise Minimization in Eukaryotic Gene Expression. *PLoS Biology*. 2004;2: 0834–0838. doi:10.1371/journal.pbio.0020137
  89. Frickel J, Feulner PGD, Karakoc E, Becks L. Population size changes and selection drive patterns of parallel evolution in a host–virus system. *Nature Communications*. 2018;9: 1706. doi:10.1038/s41467-018-03990-7
  90. Deutsch D. The fabric of reality [Internet]. 2011. Available: [https://nls.ldls.org.uk/welcome.html?ark:/81055/vdc\\_100048487759.0x000001](https://nls.ldls.org.uk/welcome.html?ark:/81055/vdc_100048487759.0x000001)
  91. Coelho MTP, Diniz-Filho JA, Rangel TF. A parsimonious view of the parsimony principle in ecology and evolution. *Ecography*. 0. doi:10.1111/ecog.04228
  92. Johnson NA. Speciation: Dobzhansky–Muller incompatibilities, dominance and gene interactions. *Trends in Ecology & Evolution*. 2000;15: 480–482. doi:10.1016/S0169-5347(00)01961-3
  93. Charlesworth B. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*. 2009;10: 195–205. doi:10.1038/nrg2526
  94. Fraser HB. Gene expression drives local adaptation in humans. *Genome Research*. 2013;23: 1089–1096. doi:10.1101/gr.152710.112

95. Runcie DE, Mukherjee S. Dissecting high-dimensional phenotypes with bayesian sparse factor analysis of genetic covariance matrices. *Genetics*. 2013;194: 753–767. doi:10.1534/genetics.113.151217
96. Huber CD, Durvasula A, Hancock AM, Lohmueller KE. Gene expression drives the evolution of dominance. *Nature Communications*. 2018;9: 2750. doi:10.1038/s41467-018-05281-7
97. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461: 747–753. doi:10.1038/nature08494
98. Nolte IM, van der Most PJ, Alizadeh BZ, de Bakker PI, Boezen HM, Bruinenberg M, et al. Missing heritability: is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study. *European Journal of Human Genetics*. 2017;25: 877–885. doi:10.1038/ejhg.2017.50
99. Keightley PD, Otto SP. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*. 2006;443: 89–92. doi:10.1038/nature05049
100. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. *Genet Res*. 1966;8: 269–294.
101. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. *Genet Res*. 2007;89: 391–403. doi:10.1017/S0016672308009579
102. Boulton a, Myers RS, Redfield RJ. The hotspot conversion paradox and the evolution of meiotic recombination. *Proceedings of the National Academy of Sciences of the United States of America*. 1997;94: 8058–8063. doi:10.1073/pnas.94.15.8058
103. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*. 2010;327: 876–879. doi:10.1126/science.1182363
104. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*. 2010;467: 1099–1103. doi:10.1038/nature09525
105. Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, et al. Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. *Mol Ecol*. 2017;26: 4158–4172. doi:10.1111/mec.14197
106. Dumont BL, Payseur BA. Genetic Analysis of Genome-Scale Recombination Rate Evolution in House Mice. *PLOS Genetics*. 2011;7: e1002116. doi:10.1371/journal.pgen.1002116
107. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*. 2010;327: 836–840. doi:10.1126/science.1183439

108. Auton A, Fedel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, et al. A fine-scale chimpanzee genetic map from population sequencing. *Science*. 2012;336: 193–198. doi:10.1126/science.1216872
109. Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, et al. Stable recombination hotspots in birds. *Science*. 2015;350: 928–932. doi:10.1126/science.aad0843
110. Heil S, S C, Ellison C, Dubin M, Noor MAF. Recombining without Hotspots: A Comprehensive Evolutionary Portrait of Recombination in Two Closely Related Species of *Drosophila*. *Genome Biol Evol*. 2015;7: 2829–2842. doi:10.1093/gbe/evv182
111. Brand CL, Cattani MV, Kingan SB, Landeen EL, Presgraves DC. Molecular Evolution at a Meiosis Gene Mediates Species Differences in the Rate and Patterning of Recombination. *Current Biology*. 2018;28: 1289-1295.e4. doi:10.1016/j.cub.2018.02.056
112. Kohl KP, Jones CD, Sekelsky J. Evolution of an MCM complex in flies that promotes meiotic crossovers by blocking BLM helicase. *Science*. 2012;338: 1363–1365. doi:10.1126/science.1228190
113. Wang J, Street NR, Scofield DG, Ingvarsson PK. Natural Selection and Recombination Rate Variation Shape Nucleotide Polymorphism Across the Genomes of Three Related *Populus* Species. *Genetics*. 2016;202: 1185–1200. doi:10.1534/genetics.115.183152
114. Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*. 2018; eaar3684. doi:10.1126/science.aar3684
115. Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, et al. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *Science*. 2017;358: 951–954. doi:10.1126/science.aao0960
116. Martin SH, Jiggins CD. Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev*. 2017;47: 69–74. doi:10.1016/j.gde.2017.08.007
117. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*. 2017;101: 5–22. doi:10.1016/j.ajhg.2017.06.005
118. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*. 2009;10: 285–311. doi:10.1146/annurev-genom-082908-150001
119. Kostka D, Hubisz MJ, Siepel A, Pollard KS. The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome. *Mol Biol Evol*. 2012;29: 1047–1057. doi:10.1093/molbev/msr279
120. Bolívar P, Mugal CF, Nater A, Ellegren H. Recombination Rate Variation Modulates Gene Sequence Evolution Mainly via GC-Biased Gene Conversion, Not Hill–Robertson

- Interference, in an Avian System. *Molecular Biology and Evolution*. 2016;33: 216–227. doi:10.1093/molbev/msv214
121. Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. Quantification of GC-biased gene conversion in the human genome. *Genome Res*. 2015;25: 1215–1228. doi:10.1101/gr.185488.114
  122. Stumpf MPH, McVean G a. T. Estimating recombination rates from population-genetic data. *Nature Reviews Genetics*. 2003;4: 959–968. doi:10.1038/nrg1227
  123. Rosenberg NA, Nordborg M. Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms. *Nature Reviews Genetics*. 2002;3: 380–390. doi:10.1038/nrg795
  124. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. *Science*. 2004;304: 581–584. doi:10.1126/science.1092500
  125. Auton A, McVean G. Recombination rate estimation in the presence of hotspots. *Genome Res*. 2007;17: 1219–1227. doi:10.1101/gr.6386707
  126. Chan AH, Jenkins PA, Song YS. Genome-Wide Fine-Scale Recombination Rate Variation in *Drosophila melanogaster*. *PLOS Genetics*. 2012;8: e1003090. doi:10.1371/journal.pgen.1003090
  127. Baudat F, Imai Y, de Massy B. Meiotic recombination in mammals: localization and regulation. *Nature Reviews Genetics*. 2013;14: 794–806. doi:10.1038/nrg3573
  128. Hudson RR, Kaplan NL. The coalescent process in models with selection and recombination. *Genetics*. 1988;120: 831–840.
  129. Wilton PR, Carmi S, Hobolth A. The SMC' Is a Highly Accurate Approximation to the Ancestral Recombination Graph. *Genetics*. 2015;200: 343–355. doi:10.1534/genetics.114.173898
  130. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole-genomes. *Nat Genet*. 2017;49: 303–309. doi:10.1038/ng.3748
  131. Munch K, Schierup MH, Mailund T. Unraveling recombination rate evolution using ancestral recombination maps. *Bioessays*. 2014;36: 892–900. doi:10.1002/bies.201400047
  132. Munch K, Mailund T, Dutheil JY, Schierup MH. A fine-scale recombination map of the human-chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion. *Genome Res*. 2014;24: 467–474. doi:10.1101/gr.158469.113
  133. Dutheil JY. Hidden Markov Models in Population Genomics. *Methods Mol Biol*. 2017;1552: 149–164. doi:10.1007/978-1-4939-6753-7\_11



134. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids [Internet]. Cambridge: Cambridge University Press; 1998. doi:10.1017/CBO9780511790492
135. McVean GAT. A genealogical interpretation of linkage disequilibrium. *Genetics*. 2002;162: 987–991.
136. Wirtz J, Rauscher M, Wiehe T. Topological linkage disequilibrium calculated from coalescent genealogies. *Theoretical Population Biology*. 2018; doi:10.1016/j.tpb.2018.09.001
137. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974;19: 716–723. doi:10.1109/TAC.1974.1100705
138. Staab PR, Zhu S, Metzler D, Lunter G. Scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*. 2015;31: 1680–1682. doi:10.1093/bioinformatics/btu861
139. Dapper AL, Payseur BA. Effects of Demographic History on the Detection of Recombination Hotspots from Linkage Disequilibrium. *Mol Biol Evol*. 2018;35: 335–353. doi:10.1093/molbev/msx272
140. Johnston HR, Cutler DJ. Population demographic history can cause the appearance of recombination hotspots. *Am J Hum Genet*. 2012;90: 774–783. doi:10.1016/j.ajhg.2012.03.011
141. Martin SH, Davey J, Salazar C, Jiggins C. Recombination rate variation shapes barriers to introgression across butterfly genomes. *bioRxiv*. 2018; 297531. doi:10.1101/297531
142. Grosse C, Keller L, Biebach I, Consortium TIGG, Croll D. Introgression from Domestic Goat Generated Variation at the Major Histocompatibility Complex of Alpine Ibex. *PLOS Genetics*. 2014;10: e1004438. doi:10.1371/journal.pgen.1004438
143. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of *de novo* mutations in humans. *Nature Genetics*. 2015;47: 822–826. doi:10.1038/ng.3292
144. Hartmann FE, McDonald BA, Croll D. Genome-wide evidence for divergent selection between populations of a major agricultural pathogen. *Molecular Ecology*. 27: 2725–2741. doi:10.1111/mec.14711
145. Grandaubert J, Dutheil JY, Stukenbrock EH. The genomic determinants of adaptive evolution in a fungal pathogen. *bioRxiv*. 2018; 176727. doi:10.1101/176727
146. Stukenbrock EH, Dutheil JY. Fine-Scale Recombination Maps of Fungal Plant Pathogens Reveal Dynamic Recombination Landscapes and Intragenic Hotspots. *Genetics*. 2018;208: 1209–1229. doi:10.1534/genetics.117.300502
147. Croll D, Lendenmann MH, Stewart E, McDonald BA. The impact of recombination hotspots on genome evolution of a fungal plant pathogen. *Genetics*. 2015;201: 1213–

1228. doi:10.1534/genetics.115.180968
148. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 2016; doi:10.1038/nature18964
149. Paigen K, Petkov PM. PRDM9 and Its Role in Genetic Recombination. *Trends Genet*. 2018;34: 291–300. doi:10.1016/j.tig.2017.12.017
150. Coop G, Myers SR. Live hot, die young: Transmission distortion in recombination hotspots. *PLoS Genetics*. 2007;3: 0377–0386. doi:10.1371/journal.pgen.0030035
151. Lartillot T, Duret L, Lartillot N. The Red Queen model of recombination hot-spot evolution: a theoretical investigation. *Phil Trans R Soc B*. 2017;372: 20160463. doi:10.1098/rstb.2016.0463
152. Lesecque Y, Glémin S, Lartillot N, Mouchiroud D, Duret L. The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes. *PLOS Genetics*. 2014;10: e1004790. doi:10.1371/journal.pgen.1004790
153. Slatkin M, Racimo F. Ancient DNA and human history. *PNAS*. 2016;113: 6380–6387. doi:10.1073/pnas.1524306113
154. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. 2013;505: 43–49. doi:10.1038/nature12886
155. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. *Science (New York, NY)*. 2010;328: 710–22. doi:10.1126/science.1188021
156. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*. 2012;338: 222–226. doi:10.1126/science.1224344
157. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*. 2014;514: 445–449. doi:10.1038/nature13810
158. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and Introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLOS Genetics*. 2014;10: e1004410. doi:10.1371/journal.pgen.1004410
159. Teng H, Zhang Y, Shi C, Mao F, Cai W, Lu L, et al. Population Genomics Reveals Speciation and Introgression between Brown Norway Rats and Their Sibling Species. *Mol Biol Evol*. 2017;34: 2214–2228. doi:10.1093/molbev/msx157
160. Van Belleghem SM, Baquero M, Papa R, Salazar C, McMillan WO, Counterman BA, et al. Patterns of Z chromosome divergence among *Heliconius* species highlight the importance of historical demography. *Mol Ecol*. 2018; doi:10.1111/mec.14560

161. Delmore KE, Lugo Ramos JS, Van Doren BM, Lundberg M, Bensch S, Irwin DE, et al. Comparative analysis examining patterns of genomic differentiation across multiple episodes of population divergence in birds. *Evolution Letters*. 2018;2: 76–87. doi:10.1002/evl3.46
162. Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Phil Trans R Soc B*. 2017;372: 20160455. doi:10.1098/rstb.2016.0455
163. Librado P, Gamba C, Gaunitz C, Sarkissian CD, Pruvost M, Albrechtsen A, et al. Ancient genomic changes associated with domestication of the horse. *Science*. 2017;356: 442–445. doi:10.1126/science.aam5298
164. Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *PNAS*. 2016;113: 5652–5657. doi:10.1073/pnas.1514696113
165. Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, et al. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol Biol Evol*. 2013;30: 1745–1750. doi:10.1093/molbev/mst097
166. Sand A, Kristiansen M, Pedersen CN, Mailund T. zipHMMLib: a highly optimised HMM library exploiting repetitions in the input to speed up the forward algorithm. *BMC Bioinformatics*. 2013;14: 339. doi:10.1186/1471-2105-14-339
167. Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, et al. The *Drosophila* Genome Nexus: A Population Genomic Resource of 623 *Drosophila melanogaster* Genomes, Including 197 from a Single Ancestral Range Population. *Genetics*. 2015;199: 1229–1241. doi:10.1534/genetics.115.174664
168. Hunter-Zinck H, Clark AG. Aberrant Time to Most Recent Common Ancestor as a Signature of Natural Selection. *Mol Biol Evol*. 2015;32: 2784–2797. doi:10.1093/molbev/msv142
169. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics*. 2014;10. doi:10.1371/journal.pgen.1004342
170. Agrawal AF, Hartfield M. Coalescence with Background and Balancing Selection in Systems with Bi- and Uniparental Reproduction: Contrasting Partial Asexuality and Selfing. *Genetics*. 2016;202: 313–326. doi:10.1534/genetics.115.181024
171. Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Veeramah K, Wittkopp PJ, Gronau I, editors. *eLife*. 2018;7: e36317. doi:10.7554/eLife.36317
172. Bhaskar A, Wang YXR, Song YS. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Res*.

- 2015;25: 268–279. doi:10.1101/gr.178756.114
173. Gossmann TI, Woolfit M, Eyre-Walker A. Quantifying the variation in the effective population size within a genome. *Genetics*. 2011;189: 1389–1402. doi:10.1534/genetics.111.132654
  174. Zeng K, Jackson BC. Methods for estimating demography and detecting between-locus differences in the effective population size and mutation rate. *Mol Biol Evol*. doi:10.1093/molbev/msy212
  175. Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD. Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends in Genetics*. 2014;30: 540–546. doi:10.1016/j.tig.2014.09.010
  176. Jonsson H, Sulem P, Arnadottir GA, Palsson G, Eggertsson HP, Kristmundsdottir S, et al. Recurrence of de novo mutations in families. *bioRxiv*. 2017; 221259. doi:10.1101/221259
  177. Besenbacher S, Hvilsum C, Marques-Bonet T, Mailund T, Schierup MH. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nature Ecology & Evolution*. 2019; 1. doi:10.1038/s41559-018-0778-x
  178. Smith TCA, Arndt PF, Eyre-Walker A. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLoS Genet*. 2018;14: e1007254. doi:10.1371/journal.pgen.1007254
  179. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet*. 2011;12: 756–766. doi:10.1038/nrg3098
  180. Barroso GV, Puzovic N, Duthel J. Inference of recombination maps from a single pair of genomes and its application to archaic samples. *bioRxiv*. 2018; 452268. doi:10.1101/452268
  181. Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*. 2019;363: eaau1043. doi:10.1126/science.aau1043
  182. Venables WN, Ripley BD. *Modern Applied Statistics with S* [Internet]. New York, NY: Springer New York; 2002. doi:10.1007/978-0-387-21706-2
  183. Groemping U. Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software*. 2006;17: 1–27. doi:10.18637/jss.v017.i01
  184. Gillespie DT. Exact Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*. 1977;81: 2340–2361.
  185. Shahrezaei V, Swain PS. The stochastic nature of biochemical networks. *Curr Opin Biotechnol*. 2008;19: 369–374. doi:10.1016/j.copbio.2008.06.011
  186. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian Genes

- Are Transcribed with Widely Different Bursting Kinetics. *Science*. 2011;332: 472–474. doi:10.1126/science.1198817
187. Mcadams HH, Arkin A. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*. 1997;94: 814–819.
  188. Becskei A, Kaufmann BB, van Oudenaarden A. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature Genetics*. 2005;37: 937–944. doi:10.1038/ng1616
  189. Raj A, Oudenaarden AV. Review Nature , Nurture , or Chance : Stochastic Gene Expression and Its Consequences. *Cell*. 2008;135: 216–226. doi:10.1016/j.cell.2008.09.050
  190. Raser JM, O’Shea EK. Noise in Gene Expression: Origins, Consequences, and Control. *Science*. 2005;309.
  191. Hebenstreit D. Are gene loops the cause of transcriptional noise? *Trends in Genetics*. 2013;29: 333–338. doi:10.1016/j.tig.2013.04.001
  192. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology*. 2015;16: 245–257. doi:10.1038/nrm3965
  193. Arkin A, Ross J, Mcadams HH. Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage L-Infected Escherichia coli Cells. *Genetics*. 1998;149: 1633–1648.
  194. Norman TM, Lord ND, Paulsson J, Losick R. Stochastic Switching of Cell Fate in Microbes. *Annual review of microbiology*. 2015;69: 381–403. doi:10.1146/annurev-micro-091213-112852
  195. Thattai M, Oudenaarden AV. Stochastic Gene Expression in Fluctuating Environments. *Genetics*. 2004;167: 523–530.
  196. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, Derisi JL, et al. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature*. 2006;441: 840–846. doi:10.1038/nature04785
  197. Lehner B. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Molecular systems biology*. 2008;4: 170–170. doi:10.1038/msb.2008.11
  198. Wang Z, Zhang J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proceedings of the National Academy of Sciences*. 2011;108: E67–E76. doi:10.1073/pnas.1100059108
  199. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, Oudenaarden AV. Regulation of noise in the expression of a single gene. *Nature genetics*. 2002;31: 69–73. doi:10.1038/ng869
  200. Chubb JR, Trcek T, Shenoy SM, Singer RH. Transcriptional Pulsing of a Developmental

- Gene. *Current Biology*. 2006;16: 1018–1025. doi:10.1016/j.cub.2006.03.092
201. Bar-even A, Paulsson J, Maheshri N, Carmi M, Shea EO, Pilpel Y, et al. Noise in protein expression scales with natural protein abundance. *Nature genetics*. 2006;38: 636–643. doi:10.1038/ng1807
  202. Kepler TB, Elston TC. Stochasticity in Transcriptional Regulation : Origins, Consequences, and Mathematical Representations. *Biophysical Journal*. 2001;81: 3116–3136.
  203. Batada NN, Hurst LD. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet*. 2007;39: 945–949. doi:10.1038/ng2071
  204. Kaufmann BB, van Oudenaarden A. Stochastic gene expression: from single molecules to the proteome. *Current opinion in genetics & development*. 2007;17: 107–112. doi:10.1016/j.gde.2007.02.007
  205. Sánchez A, Kondev J. Transcriptional control of noise in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105: 5081–5086. doi:10.1073/pnas.0707904105
  206. Wang G-Z, Lercher MJ, Hurst LD. Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. *Genome Biol Evol*. 2011;3: 320–331. doi:10.1093/gbe/evr025
  207. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. *Nature*. 2010;467: 167–173. doi:10.1038/nature09326
  208. Li J, Min R, Vizeacoumar FJ, Jin K, Xin X, Zhang Z. Exploiting the determinants of stochastic gene expression in *Saccharomyces cerevisiae* for genome-wide prediction of expression noise. *Proc Natl Acad Sci USA*. 2010;107: 10472–10477. doi:10.1073/pnas.0914302107
  209. Sauer U, Heineman M, Zamboni N. Getting Closer to the Whole Picture. *Science*. 2007;316: 550–551. doi:10.1126/science.1142502
  210. Taniguchi Y, Choi PJ, Li G, Chen H, Babu M, Hearn J, et al. Quantifying *E. coli* Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science (New York, NY)*. 2011;329: 533–539. doi:10.1126/science.1188308
  211. Tao Y, Zheng X, Sun Y. Effect of feedback regulation on stochastic gene expression. *J Theor Biol*. 2007;247: 827–836. doi:10.1016/j.jtbi.2007.03.024
  212. Pál C, Papp B, Hurst LD. Highly Expressed Genes in Yeast Evolve Slowly. *Genetics*. 2001;158: 927–931. doi:10.1080/13518040701205365
  213. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485: 376–380. doi:10.1038/nature11082

214. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. *Genome Biol.* 2015;16: 56. doi:10.1186/s13059-015-0621-5
215. Sharon E, Van Dijk D, Kalma Y, Keren L, Manor O, Yakhini Z, et al. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Research.* 2014;24: 1698–1706. doi:10.1101/gr.168773.113
216. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. *Cell.* 2007;129: 823–837. doi:10.1016/j.cell.2007.05.009
217. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature.* 2013;498: 236–240. doi:10.1038/nature12172
218. Hussain SP, Harris CC. p53 biological network: at the crossroads of the cellular-stress response pathway and molecular carcinogenesis. *J Nippon Med Sch.* 2006;73: 54–64.
219. Viney M, Reece SE. Adaptive noise. *Proc Biol Sci.* 2013;280. doi:10.1098/rspb.2013.1104
220. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L. The large-scale organization of metabolic networks. *Nature.* 2000;407: 651–654. doi:10.1038/35036627
221. Barabási A-L, Albert R. Emergence of Scaling in Random Networks. *Science.* 1999;286: 509–513.
222. Newman MEJ. The Structure and Function of Complex Networks. *SIAM Review.* 2003;45: 167–256. doi:10.1137/S003614450342480
223. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary Rate in the Protein Interaction Network. *Science.* 2002;296: 750–752. doi:10.1126/science.1068696
224. Hahn MW, Conant GC, Wagner A. Molecular Evolution in Large Genetic Networks: Does Connectivity Equal Constraint? *Journal of Molecular Evolution.* 2004;58: 203–211. doi:10.1007/s00239-003-2544-0
225. Jovelin R, Phillips PC. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome biology.* 2009;10: R35–R35. doi:10.1186/gb-2009-10-4-r35
226. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res.* 2016;44: D481-487. doi:10.1093/nar/gkv1351
227. Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. *Nature.* 2005;433: 895–900. doi:10.1038/nature03286.1
228. Joy MP, Brock A, Ingber DE, Huang S. High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine and Biotechnology.* 2005;2005: 96–103. doi:10.1155/JBB.2005.96

229. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*. 2007;3: 713–720. doi:10.1371/journal.pcbi.0030059
230. Hahn MW, Kern AD. Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. *Molecular Biology and Evolution*. 2004;22: 7–10. doi:10.1093/molbev/msi072
231. Vitkup D, Kharchenko P, Wagner A. Influence of metabolic network structure and function on enzyme evolution. *Genome biology*. 2006;7: R39–R39. doi:10.1186/gb-2006-7-5-r39
232. Maslov S, Sneppen K. Specificity and Stability in Topology of Protein Networks. *Science*. 2002;296: 910–913.
233. Hirsh A, Fraser H. Protein dispensability and rate of evolution. *Nature*. 2001;411: 1046–1049. doi:10.1038/35082561
234. Jacob F. Evolution and Tinkering. *Science*. 1977;196: 1161–1166.
235. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. *Nature reviews Genetics*. 2011;12: 692–702. doi:10.1038/nrg3053
236. Xie C, Zhang YE, Chen JY, Liu CJ, Zhou WZ, Li Y, et al. Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding RNAs. *PLoS Genetics*. 2012;8: e1002942.-e1002942. doi:10.1371/journal.pgen.1002942
237. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC genomics*. 2013;14: 117–117. doi:10.1186/1471-2164-14-117
238. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106: 7273–80. doi:10.1073/pnas.0901808106
239. Jeong H, Mason SP, Barabási a L, Oltvai ZN. Lethality and centrality in protein networks. *Nature*. 2001;411: 41–42. doi:10.1038/35075138
240. Becskei A, Serrano L. Engineering stability in gene networks by autoregulation. *Nature*. 2000;405: 590–593. doi:10.1038/35014651
241. Landgraf AJ, Lee Y. Dimensionality Reduction for Binary Data through the Projection of Natural Parameters. arXiv:151006112 [stat]. 2015; Available: <http://arxiv.org/abs/1510.06112>
242. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*. 2014;510: 363–9. doi:10.1038/nature13437



243. Pozhitkov, Alex E., Tautz D, Noble, Peter A. Oligonucleotide microarrays: widely appliedçpoorly understood. *B RIEFINGS IN FUNC TIONAL GENOMICS AND P ROTEOMICS* . 2007;6: 141–148.
244. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nat Methods*. 2014;11: 740–742. doi:10.1038/nmeth.2967
245. Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, et al. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*. 2016; 1–8. doi:10.1093/bioinformatics/btw202
246. Dublanche Y, Michalodimitrakis K, Kümmerer N, Foglierini M, Serrano L. Noise in transcription negative feedback loops: simulation and experimental analysis. *Molecular systems biology*. 2006;2: 41–41. doi:10.1038/msb4100081
247. Wolf L, Silander OK, van Nimwegen EJ. Expression noise facilitates the evolution of gene regulation. *eLife*. 2015;4: 1–48. doi:10.1101/007237
248. Metzger BPH, Yuan DC, Gruber JD, Dubeau F, Wittkopp PJ. Selection on noise constrains variation in a eukaryotic promoter. *Nature*. 2015;521: 344–347. doi:10.1038/nature14244
249. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq : a highly reproducible and sensitive single-cell RNA sequencing method , reveals non-genetic gene-expression heterogeneity. *Genome Biology*. 2013;14: R31–R31. doi:10.1186/gb-2013-14-4-r31
250. Dray S, Dufour A-B. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*. 2007;22. doi:10.18637/jss.v022.i04
251. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst*. 2016;12: 477–479. doi:10.1039/c5mb00663e
252. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal Complex Systems*. 2006;1695: 1695.
253. Sales G, Calura E, Cavalieri D, Romualdi C. graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*. 2012;13: 20. doi:10.1186/1471-2105-13-20
254. Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*. 2008;9: 405. doi:10.1186/1471-2105-9-405
255. Mora A, Donaldson IM. iRefR: an R package to manipulate the iRefIndex consolidated protein interaction database. *BMC Bioinformatics*. 2011;12: 455. doi:10.1186/1471-2105-12-455
256. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis.

Bioinformatics. 2007;23: 3024–3031. doi:10.1093/bioinformatics/btm440

257. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 2006;22: 1600–1607. doi:10.1093/bioinformatics/btl140
258. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57: 289–300.
259. Harrell FE. *Regression Modeling Strategies* [Internet]. Cham: Springer International Publishing; 2015. doi:10.1007/978-3-319-19425-7
260. Lindeman RH, Merenda PF, Gold RZ. *Introduction to Bivariate and Multivariate Analysis*. Glenview, Ill: Scott Foresman & Co; 1979.

## **Appendix 1 – Supplemental Material**

Digital Object Identifiers (DOI) for Supplemental Tables.  
Their corresponding legends can be found at the end of each chapter.

### **CHAPTER 1**

10.6084/m9.figshare.7667717

### **CHAPTER 2**

10.6084/m9.figshare.7667732

### **CHAPTER 3**

10.6084/m9.figshare.7667741

## **Appendix 2 – Book Chapter**

This appendix contains a chapter to be published in the upcoming book

*Statistical Population Genomics*

This text is co-authored with Ana Filipa Moutinho and Julien Yann Dutheil

## A population genomics lexicon

Gustavo Valadores Barroso<sup>1,\*</sup>, Ana Filipa Moutinho<sup>1,\*</sup> and Julien Y Dutheil<sup>1</sup>

<sup>1</sup> Max Planck Institute of Evolutionary Biology, Department of Evolutionary Genetics, GERMANY, [dutheil@evolbio.mpg.de](mailto:dutheil@evolbio.mpg.de). \* these authors contributed equally to this work.



---

## Contents

### **1 A population genomics lexicon**

*Gustavo Valadores Barroso, Ana Filipa Moutinho and Julien Y Dutheil* . 1

**Summary.** Population genomics is a growing field stemming from more than thirty years of developments in population genetics. Here, we summarize the main theoretical concepts and terminology underlying both theoretical and empirical statistical population genomics studies. We provide the reader with pointers toward the original literature as well as methodological and historical reviews.

Keywords: Population genetics, Neutral theory, Coalescent theory, Mutation, Recombination, Selection, Lexicon

Running head: Population Genomics Lexicon

## 1 Genomic variation

### 1.1 Loci, alleles, and polymorphism

Population genetics studies the evolution of genome variants in populations. A *locus* (*pl. loci*) refers to a given location in the genome, defined by a chromosome id, as well as begin and end positions. The particular sequence at a given locus may vary between individuals, each variant being termed an *allele*. We call loci with at least two alleles *polymorphic* and invariant loci *monomorphic*. The term *polymorphism* refers to the presence of multiple alleles but is commonly used as a countable noun as a substitute for “polymorphic locus” (*one polymorphism, several polymorphisms*).

Alleles may differ because of the nucleotide content, but also in length, as a result of nucleotide insertions or deletions (*a.k.a. indels*). Variable loci of length one can have up to four distinct alleles (A, C, G or T) and are termed *single nucleotide polymorphisms (SNPs)*. SNPs constitute so far the majority of the data accounted for by population genetic models.



## 1.2 Mutations

Molecular events altering the genome are termed *mutations*. Mutations include substitution of a nucleotide into another one, removal or addition of one or several nucleotides, as well as multiplication of some part of the genome. Mutation is the process by which new alleles are formed. The *infinite site model* assumes that during the timeframe of evolution modeled, each locus have undergone at most one mutation [1, 2, 3]. This model also implies that each mutation creates a new allele in the population and that there is no “backward” or “reverse” mutation. The infinite site model is a generally reasonable assumption as the mutation rate is typically low and genomes are large. It might be locally invalidated, however, in case of mutation hotspots or when larger evolutionary timescales are considered. Under this premise, at most two alleles are expected per locus. Loci with two alleles are termed *diallelic* or *biallelic*, the first term having historical precedence and being more accurate [4], while the second is more commonly used since the 1990s. Furthermore, in a population genomic dataset, a sampled diallelic locus is called a *singleton* if one of the two alleles is present in only one individual, and a *doubleton* if it is present in precisely two individuals.

## 1.3 The Wright-Fisher model

The most straightforward process of allele evolution within a single population is named the *Wright-Fisher model*. It describes the evolution of alleles in a population of fixed and constant size, where all alleles have the same fitness, and therefore the same chance to be transmitted to the next generation (*neutral evolution*). The population is assumed to be *panmictic*, that is, individuals are randomly mating. Time is discretized in *non-overlapping generations* so that the alleles in the current generation are a random sample

of the alleles from the previous generation. Under such conditions, allelic frequencies evolve only because of the stochasticity in the sampling of gametes that will contribute to the next generation, a process termed *genetic drift*. Because populations are of finite size, alleles will be sampled at their actual frequencies on average only and the ultimate fate of any allele is either to reach frequency zero in the population and be lost, when by chance no individual carrying this allele has any descendant in the next generation or to become fixed when all other alleles have been lost. When genetic drift is the only force acting on a population, the number of alleles at a given locus is necessarily decreasing over time.

The *Wright-Fisher model with mutation* extends the Wright-Fisher model by introducing new alleles in the population, at a given rate. Mutation and drift act in opposite direction and a *mutation-drift equilibrium* is reached when the rate of allele creation by mutation equals the rate of allele loss by drift. The genetic diversity is then determined by the sole product of the population size  $N$  and the mutation rate  $u$ . Under the infinite site model, the expected heterozygosity at a locus in a population of diploid individuals is approximated by

$$\hat{h} = \frac{4 \cdot N \cdot u}{4 \cdot N \cdot u + 1} \quad [1]$$

while the expected number of distinct alleles and their respective frequencies can be estimated using *Ewens's sampling formula* [5].

A *substitution* occurs when a new mutation has spread in the population, increasing from frequency  $1/(2N)$  to 1 (see Note 1). Kimura showed that the average time to fixation of a new mutation is  $4N$  in a population of diploid individuals [6]. Furthermore, as a neutral mutation has a probability of reaching fixation equal to  $1/(2N)$  and given that there are  $2N \cdot u$  new mutations per

generation, in a purely neutrally evolving population, the expected number of substitutions per generation is equal to  $2N \cdot u \cdot 1/(2N) = u$ . The substitution rate is therefore independent of the population size, and the number of substitutions between two populations is a direct measure of the number of generations separating them, a phenomenon termed *molecular clock* [7].

#### 1.4 The backward Wright-Fisher model: the standard coalescent

While the Wright-Fisher process naturally describes the evolution of sequences within populations one generation after the other, population genetic data typically represent individuals sampled at a given time point. For inference purposes, it is therefore convenient to model the history of the genetic material that gave rise to the sample. The modelization of the ancestry of a sample (also known as the *genealogy*) is typically done backward in time, as every locus find a common ancestor in the past, until the *most recent common ancestor (MRCA)* of the sample. The merging of two lineages in the past is called a *coalescence event*, and the set of mathematical tools describing this process under a variety of demographic models is referred to as the *coalescence theory*. Kingman [8] first described the *standard coalescent*, the genealogical model corresponding to the Wright-Fisher model (but see [9] and [10] for a historical perspective). The standard coalescent is sometimes referred to as the *Kingman's coalescent*.

## 2 Beyond the Wright-Fisher model

The Wright-Fisher model has been extended in several ways to include more realistic assumptions on the underlying evolutionary process. These extensions led to the concept of *Effective population size ( $N_e$ )*, originally defined as

the number of individuals contributing to the gene pool. When a population deviates from the assumptions of the Wright-Fisher model,  $N_e$  is no longer equal to the census population size ( $N$ ). Often (but not always) in such cases,  $N_e$  can be obtained by a linear scaling of  $N$  such that it reflects the number of individuals from an idealized Wright-Fisher population that would display the same genetic diversity as the actual population under study [11].

## 2.1 Demography

A possible deviation from the Wright-Fisher assumptions happens when the population size is not constant across generations. The term *demographic history* generally refers to the collection of demographic parameters (effective sizes, growth rates) that describes the history of the population until its most recent common ancestor [12]. When population size varies in a cyclic manner with relatively small period  $n$ , the resulting genealogies can be modeled by a Wright-Fisher process with a population size equal to the harmonic mean of the historical population sizes, so that

$$N_e = \frac{n}{\sum_i^n \frac{1}{N_i}},$$

where the  $N_i$  refer to the  $i$ th population size [13]. More drastic demographic effects include *genetic bottlenecks*, corresponding to a sharp decrease (shrinkage) in population size.

## 2.2 Population structure

In the absence of *panmixia*, genetic exchanges occur more often between certain individuals, resulting in *population structure* with several subpopulations (Fig. 1). Population structure may occur for different reasons such as overlap-

ping generations, assortative mating or geographic isolation [12]. *Assortative mating* occurs when individuals choose their mates according to some similarity between their phenotypes. If the phenotype is genetically determined, assortative mating can influence the level of heterozygosity in the population [14].

*Gene flow* describes the migration of genetic variants between subpopulations under a scenario of population structure. It reduces genetic differentiation among subpopulations [15]. Ultimately, subpopulations can diverge and become genetically isolated, a process called *speciation*. The simplest speciation processes involve spontaneous isolation (*isolation model*) or spontaneous isolation followed by a period of gene flow (*isolation with migration model*) [16]. When speciation events occur in a short timeframe and ancestral population sizes are large, ancestral polymorphism may persist in the ancestral species, a phenomenon called *incomplete lineage sorting (ILS)* (Fig. 1) [17]. The expected amount of incomplete lineage sorting depends on the number of generations between two isolation events ( $\Delta_T$ ) and the ancestral effective population size  $Ne_A$  [18]:

$$\Pr(ILS) = \frac{2}{3} e^{\left(-\frac{2 \cdot \Delta_T}{Ne_A}\right)}$$

The term *introgression* is used to depict the transfer of genetic material between diverged populations or species through secondary contact (hybridization, see Fig. 1) [19]. As a result, extant lineages share a common ancestor that predates the two isolation or speciation events. The resulting genealogy may, therefore, be incongruent with the phylogeny defined by the two splits, depending on the order of coalescence events between lineages [20].

### 3 Statistics on nucleotide diversity

Statistics are needed to infer population genetics parameters from polymorphism data. The *site frequency spectrum (SFS)* describes the empirical distribution of allele frequencies across segregating sites of a given (set of) loci in a population sample. For a sample of  $n$  chromosomes, the unfolded SFS is the set of counts of derived alleles  $X = (X_1, X_2, \dots, X_{n-1})$ , where sample configurations  $X_i$  denote the number of sites that have  $n - i$  ancestral and  $i$  derived alleles. The ancestral state is usually estimated using an outgroup sequence. In cases where we cannot assess the ancestral allele, the folded site frequency spectrum,  $X'$ , may be calculated instead.  $X'$  represents the distribution of the minor allele frequencies, such as  $X'_i = X_i + X_{n-i}$  for  $i < n/2$  and  $X'_{n/2} = X_{n/2}$  [13, 21, 22]. The shape of the SFS is affected by underlying population genetic processes, such as demography and selection, and therefore serves as the input of many population genetics methods [23].

*Watterson's theta* is an estimator of the population mutation rate  $\theta = 4Ne \cdot u$ , where  $Ne$  is the (diploid) effective population size and  $u$  the mutation rate. It is derived from the number of segregating sites  $S_n$  of a sample of size  $n$  [24]. Assuming an infinite sites model,  $S_n$  is equal to the product of  $u$  and the expected time to coalescence:

$$E[S_n] = u \cdot 4 \cdot Ne \sum_{i=1}^{n-1} i.$$

Since  $4Ne \cdot u = \theta$  the equation may be written as  $E[S_n] = \theta \cdot a_n$ , where  $a_n = \sum_{i=1}^{n-1} i$ . The proposed estimator of  $\theta$  is

$$\hat{\theta} = \frac{\hat{S}_n}{a_n} = \frac{\hat{S}_n}{\left(1 + \frac{1}{2} + \dots + \frac{1}{n-1}\right)},$$

where  $\hat{S}_n$  is the observed number of segregating sites in the sample. This estimator is unbiased when the data is generated from a Wright-Fisher process but is not robust to deviations from it [25].

Tajima's  $\pi$ , the *average pairwise heterozygosity* is a measure of nucleotide diversity defined as the number of pairwise differences between a set of sequences [26]. Under the infinite sites model, the number of mutations separating two orthologous chromosomes  $D_{ij}$  is equal to the number of nucleotide differences between sequences  $i$  and  $j$ . As the expectation of the average pairwise nucleotide differences between all pairs of sequences in a sample is equal to  $\theta = 4Ne \cdot u$  [27], Tajima's estimator of  $\theta$  is:

$$\hat{\pi} = \frac{2}{n(n-1) \cdot L} \sum_{i=1}^{n-1} \sum_{j=i+1}^n D_{ij},$$

where  $L$  is the total sequence length.

## 4 Selective processes

### 4.1 Protein-coding genes

The coding region of a protein-coding gene, also known as *Coding DNA Sequence (CDS)* is the portion of DNA, or RNA, that encodes a protein. A start and stop codons limit the coding region at the five-prime and three-prime end, respectively. In mRNAs, the CDS is bounded by the five-prime untranslated region (5'-UTR) and the three-prime untranslated region (3'-UTR), also included in the exons. Mutations within coding regions are expected to be of distinct types: *synonymous mutations* lead to no change of amino-acid at the protein level due to the redundancy of the genetic code, as opposed to *non-synonymous mutations*. Non-synonymous mutations can further be classified

as *conservative* and *non-conservative* (= *radical*), whether they replace an amino-acid by a biochemically similar one or not.

## 4.2 Fitness effect

The resulting change of fitness at the organism level characterizes the type of mutations: neutral mutations have no impact on the fitness, while harmful or deleterious mutations induce a lower fitness. Conversely, advantageous mutations increase the fitness of the organism compared to the non-mutated genotype. There is, however, a wide range of selective effects, which extends the categorization of mutations from strongly deleterious, through weakly deleterious, neutral to mildly and highly adaptive mutations. The relative frequencies of these types of mutations represent the distribution of fitness effects [28, 29].

The *selection coefficient* ( $s$ ) is a measure of differences in fitness, which determines the changes in genotype frequencies that occur due to selection. It is commonly expressed as a relative fitness. If one considers a single locus with two alleles  $A$  and  $a$ , a standard parametrization is to attribute a fitness of 1 to the homozygote  $AA$  and relative fitness of  $1 + s$  for the homozygote  $aa$ . The heterozygote  $Aa$  is attributed a fitness of  $1 + h \cdot s$ , where  $h$  is the so-called *coefficient of dominance*. The  $s$  parameter varies between  $-1$  and  $+\infty$  (but see Note 2), wherein values comprised among  $-1$  and  $0$  are indicative of negative selection, while positive values correspond to positive selection [13, 30].

## 4.3 Types of selection

*Positive selection* acts on alleles that increase fitness, raising their frequency in the population over time, while *negative selection* (= *purifying selection*)



decrease the frequency of alleles that impair fitness. Both positive and negative selection decrease genetic diversity. Conversely, *balancing selection* acts by maintaining multiple alleles in the gene pool of a population at frequencies higher than expected by drift alone. Three mechanisms are generally acknowledged: *heterozygous advantage*, where heterozygotes have a higher fitness than homozygotes and maintain genetic polymorphism; *frequency-dependent selection*, where the fitness of the genotype is inversely proportional to its frequency in the population; and *environment-dependent fitness* of genotypes [30, 31].

Under positive selection, a new beneficial mutation will rise in frequency in a population. As the new positively-selected allele increases its frequency, nearby linked alleles on the chromosome will “hitchhike” along with it, also growing in frequency, thus producing a *selective sweep* of genetic diversity. *Hard sweeps* occur when a new mutation is positively selected and is therefore exclusively associated with the genetic background where it arose. Conversely, *soft sweeps* occur when a mutation is already segregating in the population at the onset of selection. This mutation may exist in several genetic backgrounds and therefore does not prompt a complete loss of genetic variation after the selective sweep [32].

#### 4.4 Inference of selection

The strength and direction of selection acting on protein-coding regions may be assessed by contrasting the rate of non-synonymous (potentially under selection,  $dN$ ) to synonymous (assumed to be neutral,  $dS$ ) substitutions between species. In a population of sequences evolving neutrally, all substitutions are neutral and the two rates are equal, leading to a  $dN/dS$  ratio equal to one on average. Assuming non-synonymous mutations are either neutral or deleterious while synonymous mutations are always neutral, the rate of

non-synonymous substitutions will be lower than the rate of synonymous substitutions, and the  $dN/dS$  ratio will be lower than one. Conversely, if non-synonymous mutations are positively selected, their rate of fixation may exceed the rate of synonymous mutation, leading to a higher substitution rate and a  $dN/dS$  ratio higher than one.

At the population level, the ratio of non-synonymous ( $pN$ ) and synonymous ( $pS$ ) polymorphism is indicative of the strength of purifying selection acting on a protein. Because non-synonymous mutations are more likely to have a negative effect and be counter-selected, they will be removed from the population or segregate at low-frequency. We can estimate the synonymous and non-synonymous genetic diversity by computing the average pairwise heterozygosity  $\pi$  separately for non-synonymous and synonymous mutations, noted  $\pi_N$  and  $\pi_S$ , respectively. The  $\pi_N/\pi_S$  ratio is therefore generally below one, the stronger the purifying selection, the closer the ratio is to zero.

Contrasting the  $dN/dS$  and  $pN/pS$  ratios allows to test the selection regime acting on the sequences [33]. If mutations are all neutral, we expect the ratios  $dN/dS$  and  $pN/pS$  to be equal. Positively selected mutations will tend to quickly rise to fixation and will not be observed as polymorphism, leading to an increased  $dN/dS$  ratio higher than  $pN/pS$ . Conversely, balancing selection will lead to an excess of polymorphism detectable as  $dN/dS < pN/pS$  [34]. A simple measure of the proportion of amino-acid substitutions resulting from positive selection ( $\alpha$ ) is given by  $1 - (dS \cdot pN/dN \cdot pS)$  [35]. Using the complete synonymous and non-synonymous site-frequency spectra, it is further possible to estimate the distribution of fitness effects and account for slightly deleterious and slightly advantageous mutations when estimating the rate of adaptive substitutions (see Chapter 5) [36].

## 5 Linkage and recombination

### 5.1 The coalescent with recombination

In sexually reproducing species, *recombination* refers to both the shuffling of non-homologous chromosomes and the rearrangement of homologous chromosomes during meiosis. Such cross-over events cause each chromosome to have two parent chromosomes in the previous generation, which are themselves the products of recombination events in the previous generations. Therefore, any chromosome in the current generation can be viewed as a mosaic of chromosomes that existed in the past (Fig. 3) [37]. The collection of coalescence and recombination events that describes the history of sampled chromosomes until the most recent common ancestor of each non-recombining block is reached (Fig. 3) is called the *ancestral recombination graph (ARG)* [40]. Compared to a tree-like genealogy of a sample without recombination, whose complexity depends only on the sample size, the complexity of the ARG grows with the sample size and the number of recombination events in the ancestry of the sample.

Backward-in-time, the *most recent common ancestor (MRCA)* denotes the first individual where the entire sample (population) coalesces for a particular non-recombining block. The *TMRCA* notes the timing of such event. DNA sequences provide no information beyond the MRCA in a sample of genomes since all individuals will share any mutation that happens further back in time [39]. In the presence of recombination, different parts of the genome will have different MRCA. In this case, all ancestral material is eventually found as a contiguous sequence in the *grand most recent common ancestor (GMRCA)* of the sample (Fig. 3). If the GMRCA is not an MRCA for any nucleotide, this individual does not have any significance for DNA sequences [40].

In the ARG, nucleotide segments that are found both in past chromosomes and in contemporary samples are termed *ancestral genetic material* (see Fig. 2). Conversely, *non-ancestral genetic material* refers to segments that are found in past chromosomes but not in contemporary samples. Furthermore, non-ancestral genetic material flanked on both sides by ancestral genetic material is referred to as *trapped genetic material*. In this setting, recombination events that happen in trapped genetic material can affect linkage disequilibrium between present-day nucleotides (Fig. 2). Thus the existence of trapped genetic material introduces long-range correlations between genealogies rendering the coalescent with recombination a non-Markovian process along chromosomes [41]. The *Sequentially Markov coalescent (SMC)* is an approximation to the coalescent with recombination whereby recombination events are assumed to happen only within ancestral material. This approximation allows the use of efficient algorithms in both simulation and data analysis [42, 43].

## 5.2 Impact of linkage on selection

An excess of linkage between loci compared to a random association is termed *linkage disequilibrium (LD)*. LD arises from genetic drift, population admixture, and selection, but is removed by recombination. It is, therefore, higher between close loci and decays with increasing physical distance [44].

*Linked selection* refers to the reduction of diversity at neutral sites that happens as a result of their physical linkage to variants under selection. In the absence of recombination, all variants segregating in a chromosome would undergo the same shift in frequency as the selected variant. However, recombination creates new allelic combinations and reduces this correlation as the physical distance from the selected locus increases. In some cases, we can

model the local reduction in diversity as a result of linked selection as a local reduction in the effective population size [45]. *Background selection* refers to a form of linked selection where the reduction of diversity at neutral loci results from linkage to a locus under purifying selection [46], and *genetic hitchhiking* is commonly used to depict linked selection due to linkage to a locus under positive directional selection [32]. Linkage of two or more loci can also impair the efficacy of selection. In the absence of recombination between selected loci, only the unlikely event of recurrent mutations can generate the optimal haplotypic combination [47], a phenomenon termed *Hill-Robertson Interference (HRI)*.

## References

1. Kimura, M., Crow, J.F. (1964), The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738
2. Kimura, M. (1969), The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**(4), 893–903
3. Crow, J.F. (1989), Twenty-five years ago in genetics: the infinite allele model. *Genetics* **121**(4), 631–634
4. Elston, R.C., Satagopan, J., Sun, S. (2017), Statistical Genetic Terminology. *Methods Mol. Biol.* **1666**, 1–9. DOI 10.1007/978-1-4939-7274-6\_1
5. Ewens, W.J. (1972), The sampling theory of selectively neutral alleles. *Theor Popul Biol* **3**(1), 87–112
6. Kimura, M. (1970), The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. *Genet. Res.* **15**(1), 131–133
7. Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge. URL <http://ebooks.cambridge.org/ref/id/CB09780511623486>

8. Kingman, J.F.C. (1982), The coalescent. *Stochastic Processes and their Applications* **13**(3), 235–248. DOI 10.1016/0304-4149(82)90011-4. URL <http://www.sciencedirect.com/science/article/pii/0304414982900114>
9. Barton, N.H. (2016), Richard Hudson and Norman Kaplan on the Coalescent Process. *Genetics* **202**(3), 865–866. DOI 10.1534/genetics.116.187542
10. Kingman, J.F.C. (2000), Origins of the Coalescent: 1974-1982. *Genetics* **156**(4), 1461–1463. URL <http://www.genetics.org/content/156/4/1461>
11. Sjödin, P., Kaj, I., Krone, S., Lascoux, M., Nordborg, M. (2005), On the meaning and existence of an effective population size. *Genetics* **169**(2), 1061–1070. DOI 10.1534/genetics.104.026799
12. Wakeley, J. (2008). *Coalescent Theory: An Introduction*, 1st edition edition edn. Roberts and Company Publishers, Reading, Pa. : Bloxham
13. Wright, S. (1938), The Distribution of Gene Frequencies Under Irreversible Mutation. *Proc. Natl. Acad. Sci. U.S.A.* **24**(7), 253–259
14. Jiang, Y., Bolnick, D.I., Kirkpatrick, M. (2013), Assortative mating in animals. *Am. Nat.* **181**(6), E125–138. DOI 10.1086/670160
15. Sousa, V., Hey, J. (2013), Understanding the origin of species with genome-scale data: modelling gene flow. *Nat. Rev. Genet.* **14**(6), 404–414. DOI 10.1038/nrg3446
16. Hey, J., Nielsen, R. (2004), Multilocus Methods for Estimating Population Sizes, Migration Rates and Divergence Time, With Applications to the Divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**(2), 747–760. DOI 10.1534/genetics.103.024182. URL <http://www.genetics.org/content/167/2/747>
17. Dutheil, J.Y., Hobolth, A. (2012), Ancestral population genomics. *Methods Mol. Biol.* **856**, 293–313. DOI 10.1007/978-1-61779-585-5\_12
18. Hobolth, A., Christensen, O.F., Mailund, T., Schierup, M.H. (2007), Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* **3**(2), e7. DOI 10.1371/journal.pgen.0030007

19. Martin, S.H., Jiggins, C.D. (2017), Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.* **47**, 69–74. DOI 10.1016/j.gde.2017.08.007
20. Mailund, T., Munch, K., Schierup, M.H. (2014), Lineage Sorting in Apes. *Annu. Rev. Genet.* DOI 10.1146/annurev-genet-120213-092532
21. Bustamante, C.D., Wakeley, J., Sawyer, S., Hartl, D.L. (2001), Directional selection and the site-frequency spectrum. *Genetics* **159**(4), 1779–1788
22. Wright, S. (1968). *Evolution and the Genetics of Populations: Volume 2, The Theory of Gene Frequencies*. Univ of Chicago Pr, Chicago
23. Schraiber, J.G., Akey, J.M. (2015), Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* **16**(12), 727–740. DOI 10.1038/nrg4005
24. Watterson, G.A. (1975), On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**(2), 256–276
25. Tajima, F. (1989), Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3), 585–595
26. Nei, M., Tajima, F. (1981), Genetic Drift and Estimation of Effective Population Size. *Genetics* **98**(3), 625–640. URL <http://www.genetics.org/content/98/3/625>
27. Tajima, F. (1983), Evolutionary Relationship of DNA Sequences in Finite Populations. *Genetics* **105**(2), 437–460
28. Eyre-Walker, A., Keightley, P.D. (2007), The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**(8), 610–618. DOI 10.1038/nrg2146
29. Orr, H.A. (2009), Fitness and its role in evolutionary genetics. *Nat. Rev. Genet.* **10**(8), 531–539. DOI 10.1038/nrg2603
30. Gillespie, J.H. (2004). *Population Genetics: A Concise Guide*. JHU Press
31. Nielsen, R. (2005), Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218. DOI 10.1146/annurev.genet.39.073003.112420
32. Maynard Smith, J., Haigh, J. (1974), The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**(1), 23–35
33. McDonald, J.H., Kreitman, M. (1991), Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**(6328), 652–654. DOI 10.1038/351652a0

34. Parsch, J., Zhang, Z., Baines, J.F. (2009), The influence of demography and weak selection on the McDonald–Kreitman test: an empirical study in *Drosophila*. *Mol Biol Evol* **26**(3), 691–698. DOI 10.1093/molbev/msn297. URL <https://academic.oup.com/mbe/article/26/3/691/979205>
35. Smith, N.G.C., Eyre-Walker, A. (2002), Adaptive protein evolution in *Drosophila*. *Nature* **415**(6875), 1022–1024. DOI 10.1038/4151022a
36. Keightley, P.D., Eyre-Walker, A. (2007), Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**(4), 2251–2261. DOI 10.1534/genetics.107.080663
37. Stumpf, M.P.H., McVean, G.A.T. (2003), Estimating recombination rates from population-genetic data. *Nat. Rev. Genet.* **4**(12), 959–968. DOI 10.1038/nrg1227
38. Griffiths, R., Marjoram, P.: An ancestral recombination graph. In: *Progress in population genetics and human evolution*, pp. 257 – 270. Springer (1997)
39. Rosenberg, N.A., Nordborg, M. (2002), Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3**(5), 380–390. DOI 10.1038/nrg795
40. Hein, J., Schierup, M.H., Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press
41. Rasmussen, M.D., Hubisz, M.J., Gronau, I., Siepel, A. (2014), Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**(5), e1004342. DOI 10.1371/journal.pgen.1004342
42. McVean, G.A.T., Cardin, N.J. (2005), Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **360**(1459), 1387–1393. DOI 10.1098/rstb.2005.1673
43. Marjoram, P., Wall, J.D. (2006), Fast "coalescent" simulation. *BMC Genet.* **7**, 16. DOI 10.1186/1471-2156-7-16
44. Slatkin, M. (2008), Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**(6), 477–485. DOI 10.1038/nrg2361



45. Cutter, A.D., Payseur, B.A. (2013), Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* **14**(4), 262–274. DOI 10.1038/nrg3425
46. Charlesworth, B., Morgan, M.T., Charlesworth, D. (1993), The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**(4), 1289–1303
47. Roze, D., Barton, N.H. (2006), The Hill-Robertson effect and the evolution of recombination. *Genetics* **173**(3), 1793–1811. DOI 10.1534/genetics.106.058586
48. Kelleher, J., Etheridge, A.M., McVean, G. (2016), Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12**(5), e1004842. DOI 10.1371/journal.pcbi.1004842

## 6 Notes

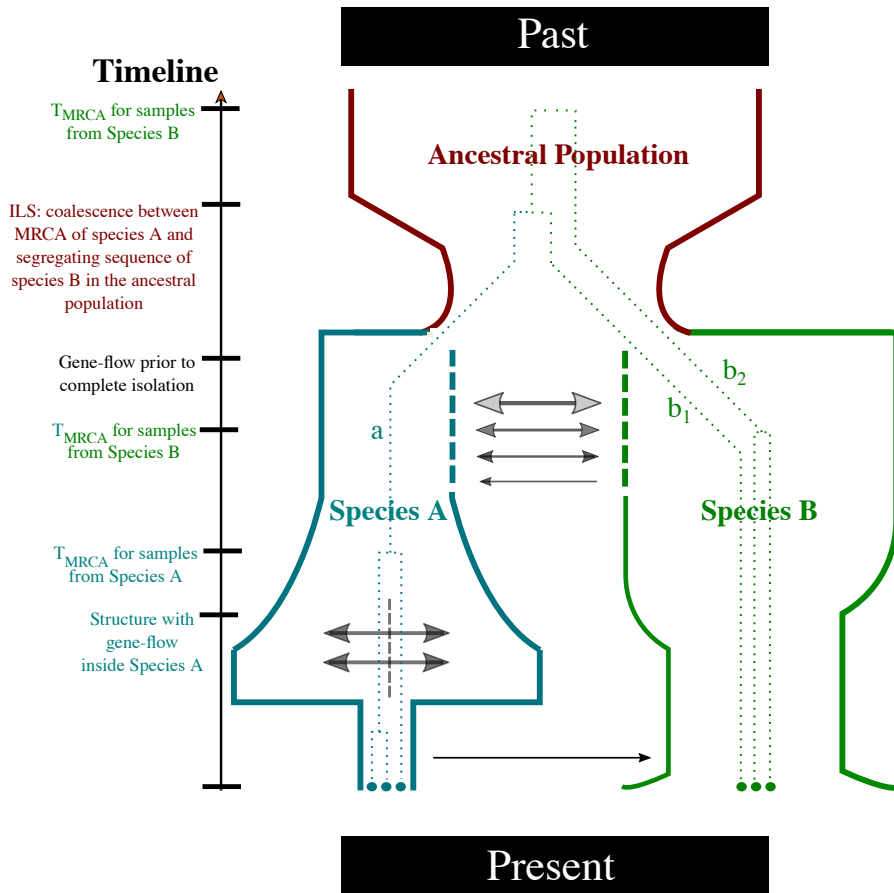
### 6.1 Note 1

The use of the term *substitution* differs in population genetics and molecular biology. In the latter case, it describes a particular type of mutation where a single nucleotide replaces a distinct one (as opposed to insertions/deletions for instance).

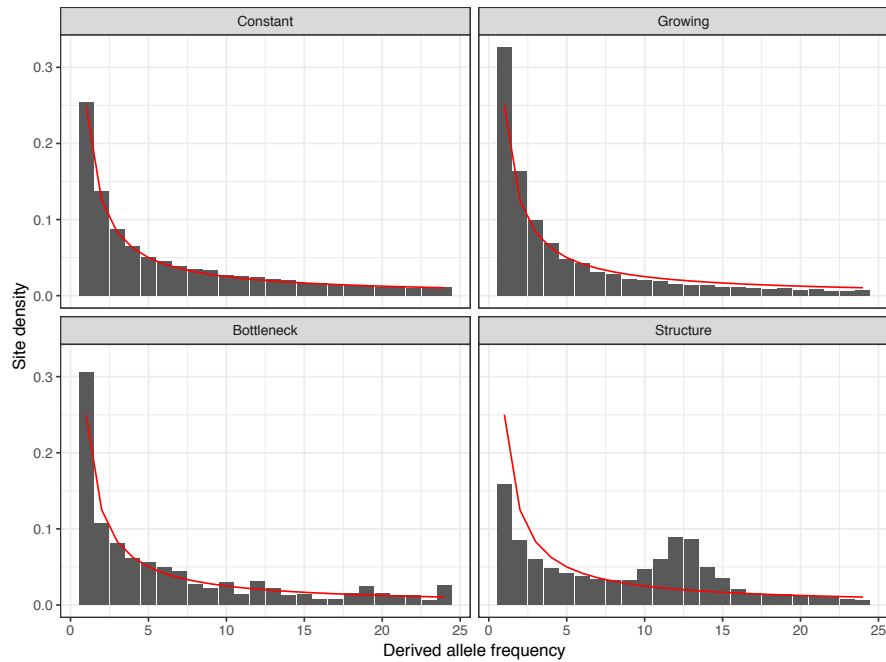
### 6.2 Note 2

In some instances,  $s$  is substituted by  $-s$ , so that the relative fitnesses become  $\omega_{AA} = 1$ ,  $\omega_{Aa} = 1 - h \cdot s$  and  $\omega_{aa} = 1 - s$ .

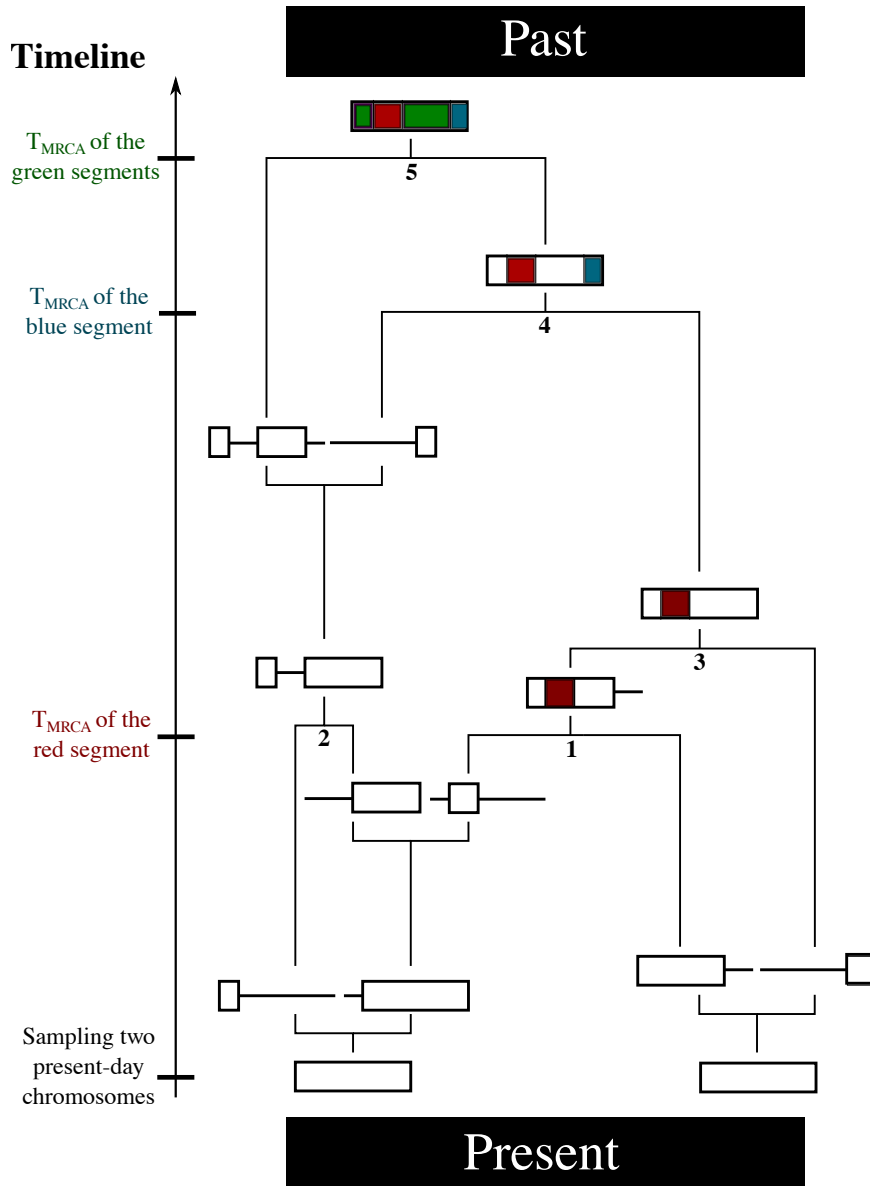
## 7 Figures



**Fig. 1.** Complex demographic history involving growth, shrinkage, structure, migration, and introgression. Colored lines represent species boundaries. Dashed lines denote structure with gene-flow whereas solid lines denote complete isolation. Arrows represent the direction of gene-flow; their thickness represents its magnitude. Dotted lines represent the genealogy of a non-recombining chromosomal segment experiencing incomplete lineage sorting. In this case, the MRCA of species A coalesces with one of the segregating sequences from species B, before (backward-in-time) the two segregating sequences from species B coalesce with each other (*i.e.*, before the MRCA of sampled B sequences is found).



**Fig. 2.** Effect of demography on the shape of the site frequency spectrum (SFS). The figure depicts four scenarios: constant population size, exponential growth, genetic bottleneck, and population structure. The red curve shows the expectation under a constant population size. In the case of exponential growth or a genetic bottleneck, the SFS displays an excess of low-frequency variants. Population structure, here simulated as two subpopulations exchanging migrants at a low rate, results in an excess of intermediate frequency variant when we reconstruct a single SFS from the two subpopulations. Simulations were performed using the `msprime` software [48], see Chapter 9 for more details on the `msprime` software and the online companion material.



**Fig. 3.** An Ancestral Recombination Graph. The ancestry of two chromosomes from the present population is traced back until its GMRCA. Thick bars represent ancestral material. Thin lines represent non-ancestral material. Yellow stars denote trapped non-ancestral material. 1:5 represent coalescence events. When a particular segment of the chromosome finds its MRCA, this region is colored. Note that the presence of trapped material causes two non-contiguous segments (green) to have the same MRCA (which also happens to be the GMRCA of the sample), resulting in long-distance Linkage Disequilibrium.

## **Appendix 3 – Published Article**

This appendix contains a version of chapter 3 as published by the journal *Genetics* in January 2018

# The Evolution of Gene-Specific Transcriptional Noise Is Driven by Selection at the Pathway Level

Gustavo Valadares Barroso,<sup>\*1</sup> Natasa Puzovic,<sup>\*</sup> and Julien Y. Dutheil<sup>\*,†</sup>

<sup>\*</sup>Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany and <sup>†</sup>Unité mixte de recherche 5554, Institut des Sciences de l'Évolution, Université de Montpellier, 34095, France

ORCID IDs: 0000-0002-1943-9297 (G.V.B.); 0000-0001-7753-4121 (J.Y.D.)

**ABSTRACT** Biochemical reactions within individual cells result from the interactions of molecules, typically in small numbers. Consequently, the inherent stochasticity of binding and diffusion processes generates noise along the cascade that leads to the synthesis of a protein from its encoding gene. As a result, isogenic cell populations display phenotypic variability even in homogeneous environments. The extent and consequences of this stochastic gene expression have only recently been assessed on a genome-wide scale, owing, in particular, to the advent of single-cell transcriptomics. However, the evolutionary forces shaping this stochasticity have yet to be unraveled. Here, we take advantage of two recently published data sets for the single-cell transcriptome of the domestic mouse *Mus musculus* to characterize the effect of natural selection on gene-specific transcriptional stochasticity. We show that noise levels in the mRNA distributions (also known as transcriptional noise) significantly correlate with three-dimensional nuclear domain organization, evolutionary constraints on the encoded protein, and gene age. However, the position of the encoded protein in a biological pathway is the main factor that explains observed levels of transcriptional noise, in agreement with models of noise propagation within gene networks. Because transcriptional noise is under widespread selection, we argue that it constitutes an important component of the phenotype and that variance of expression is a potential target of adaptation. Stochastic gene expression should therefore be considered together with the mean expression level in functional and evolutionary studies of gene expression.

**KEYWORDS** evolution of gene expression; systems biology; expression noise; biological networks; *Mus musculus*

ISOGENIC cell populations display phenotypic variability even in homogeneous environments (Spudich and Koshland 1976). This observation challenged the clockwork view of the intracellular molecular machinery and led to the recognition of the stochastic nature of gene expression. Since biochemical reactions result from the interactions of individual molecules in small numbers (Gillespie 1977), the inherent stochasticity of binding and diffusion processes generates noise along the biochemical cascade leading to the synthesis of a protein from its encoding gene (Figure 1). The study of stochastic gene expression (SGE) classically recognizes two sources of expression noise. Following the definition introduced by Elowitz *et al.* (2002), extrinsic noise

results from variation in the concentration, state, and location of shared key molecules involved in the reaction cascade from transcription initiation to protein folding. This is because molecules that are shared among genes, such as ribosomes and RNA polymerases, are typically present in low copy numbers relative to the number of genes that are actively transcribed (Shahrezaei and Swain 2008). Extrinsic factors also include physical properties of the cell such as size and growth rate, which are likely to impact the diffusion process of all molecular players. Extrinsic factors therefore affect every gene in a cell equally. Conversely, intrinsic factors generate noise in a gene-specific manner. They involve, for example, the strength of *cis*-regulatory elements (Suter *et al.* 2011), as well as the stability of the mRNA molecules that are transcribed (McAdams and Arkin 1997; Thattai and Oudenaarden 2001). Every gene is affected by both sources of stochasticity and the relative importance of each has been discussed in the literature (Beckstein *et al.* 2005; Raj and Oudenaarden 2008). Shahrezaei and Swain (2008) proposed a more general, systemic definition for any organization level, where intrinsic stochasticity is “generated by the dynamics of

Copyright © 2018 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.117.300467>

Manuscript received May 24, 2017; accepted for publication October 13, 2017; published Early Online November 2, 2017.

Available freely online through the author-supported open access option.

Supplemental material is available online at [www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300467/-/DC1](http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300467/-/DC1).

<sup>1</sup>Corresponding author: Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306 Plön, Germany. E-mail: [gvalbarroso@evolbio.mpg.de](mailto:gvalbarroso@evolbio.mpg.de)

the system from the random timing of individual reactions” and extrinsic stochasticity is “generated by the system interacting with other stochastic systems in the cell or its environment.” This generic definition therefore includes Raser and O’Shea’s suggestion to further distinguish extrinsic noise occurring “within pathways” and “between pathways” (Raser and O’Shea 2005). Other organization levels of gene expression are also likely to affect expression noise, such as chromatin structure (Blake *et al.* 2003; Hebenstreit 2013) and three-dimensional (3D) genome organization (Pombo and Dillon 2015).

Pioneering work by Fraser *et al.* (2004) has shown that SGE is an evolvable trait that is subject to natural selection. First, genes involved in core functions of the cell are expected to behave more deterministically (Barkai and Leibler 1999) because temporal oscillations in the concentration of their encoded proteins are likely to have a deleterious effect. Second, genes involved in the immune response (Arkin *et al.* 1998; Norman *et al.* 2015) and responses to environmental conditions can benefit from being unpredictably expressed in the context of selection for bet-hedging (Thattai and Oudenaarden 2004). As the relationship between fitness and stochasticity depends on the function of the underlying gene, selection on SGE is expected to act mostly at the intrinsic level (Newman *et al.* 2006; Lehner 2008; Wang and Zhang 2011). However, the molecular mechanisms by which natural selection operates to regulate expression noise remain to be elucidated.

Due to methodological limitations, seminal studies on SGE (both at the mRNA and protein levels) have focused on only a handful of genes (Elowitz *et al.* 2002; Ozbudak *et al.* 2002; Chubb *et al.* 2006). The canonical approach consists of selecting genes of interest and recording the change of their noise levels in a population of clonal cells as a function of either: (1) the concentration of the molecule that controls the affinity of the transcription factor (TF) to the promoter region of the gene (Blake *et al.* 2003; Bar-Even *et al.* 2006) or (2) mutations artificially imposed in regulatory sequences (Ozbudak *et al.* 2002). In parallel with theoretical work (Kepler and Elston 2001; Batada and Hurst 2007; Kaufmann and van Oudenaarden 2007; Sánchez and Kondev 2008), these pioneering studies have provided the basis of our current understanding of the proximate molecular mechanisms behind SGE, namely complex regulation by TFs, architecture of the upstream region (including the presence of the TATA box), gene orientation (Wang *et al.* 2011), translation efficiency, mRNA/protein stability (Eldar and Elowitz 2010), and properties of the protein–protein interaction (PPI) network (Li *et al.* 2010). However, measurements at the genome scale coupled with rigorous statistical analyses are needed to go beyond gene idiosyncrasies and particular histories, and test hypotheses about the evolutionary forces shaping SGE (Sauer *et al.* 2007).

The recent advent of single-cell RNA sequencing makes it possible to sequence the transcriptome of each individual cell in a collection of clones, and to observe the variation of gene-specific mRNA quantities across cells. This provides a genome-wide assessment of transcriptional noise. While not accounting for putative noise resulting from the process of translation of mRNAs into proteins, transcriptional noise accounts for noise generated

by both the synthesis and degradation of mRNA molecules (Figure 1). However, previous studies have shown that transcription is a limiting step in gene expression and that transcriptional noise is therefore a good proxy for expression noise (Newman *et al.* 2006; Taniguchi *et al.* 2011). Here, we used publicly available single-cell transcriptomics data sets to quantify gene-specific transcriptional noise and relate it to other genomic factors to uncover the molecular basis of selection on SGE.

## Materials and Methods

### Single-cell gene expression data set

We used the data set generated by Sasagawa *et al.* (2013) retrieved from the Gene Expression Omnibus repository (accession number GSE42268). We analyzed expression data corresponding to embryonic stem cells (ESC) in G1 phase, for which more individual cells were sequenced. A total of 17,063 genes had non-zero expression in at least one of the 20 single cells. Similar to Shalek *et al.* (2014), a filtering procedure was performed where only genes whose expression level satisfied  $\log[\text{fragments per kilobase of transcripts per million mapped fragments (FPKM)} + 1] > 1.5$  in at least one single cell were kept for further analyses. This filtering step resulted in a total of 13,660 appreciably expressed genes for which transcriptional noise was evaluated.

### Measure of transcriptional noise

The expression mean ( $\mu$ ) and variance ( $\sigma^2$ ) of each gene over all single cells were computed. We measured SGE as the ratio  $F^* = \sigma^2 / \sigma^2(\mu)$ , where  $\sigma^2(\mu)$  is the expected variance given the mean expression. To compute  $\sigma^2(\mu)$ , we performed several polynomial regressions with  $\log(\sigma^2)$  as a function of  $\log(\mu)$ , with degrees between 1 and 5. We then tested the resulting  $F^*$  measures for residual correlation with mean expression using Kendall’s rank correlation test. We find that a degree 3 polynomial regression was sufficient to remove any residual correlation with  $F^*$  (Kendall’s  $\tau = 0.0037$ ,  $P$ -value = 0.5217).  $F^*$  can be seen as a general expression for the Fano factor and noise measure: when using a polynomial of degree 1, the expression of  $F^*$  becomes  $F^* = \sigma^2 / \exp(a + b \cdot \log(\mu)) = \sigma^2 / \exp(a) \cdot \mu^b$ , and is therefore equivalent to the Fano factor when  $a = 0$  and  $b = 1$ , and equivalent to noise when  $a = 0$  and  $b = 2$ .

### Genome architecture

The mouse proteome from Ensembl (genome version: mm9) was used to get coordinates of all genes. The Hi-C data set for ESCs from Dixon *et al.* (2012) was used to get 3D domain information. Two genes were considered in proximity in one dimension (1D) if they are on the same chromosome and no protein-coding gene was found between them. The primary distance (in number of nucleotides) between their midpoint coordinates was also recorded as 1D a distance measure between the genes. Two genes were considered in proximity in 3D if the normalized contact number between

the two windows that the genes belonged to was non-null. Two genes belonging to the same window were considered to be in proximity. We further computed the relative difference of SGE between two genes by computing the ratio  $(F_2^* - F_1^*) / (F_2^* + F_1^*)$ . For each chromosome, we independently tested whether there was a correlation between the primary distance and the relative difference in SGE with a Mantel test, as implemented in the *ade4* package (Dray and Dufour 2007). To test whether genes in proximity (1D and 3D) had more similar transcriptional noise than distant genes, we contrasted the relative differences in transcription noise between pairs of genes in proximity and pairs of distant genes. As we test all pairs of genes, we performed a randomization procedure to assess the significance of the observed differences by permuting the rows and columns in the proximity matrices 10,000 times. Linear models accounting for “spatial” interactions with genes were fitted using the generalized least squares (GLS) procedure, as implemented in the *nlme* package for R. A correlation matrix between all tested genes was defined as  $G = \{g_{i,j}\}$ , where  $g_{i,j}$  is the correlation between genes  $i$  and  $j$ . We defined  $g_{i,j} = 1 - \exp(-\lambda \delta_{i,j})$ , where  $\delta_{i,j}$  takes 1 if genes  $i$  and  $j$  are in proximity, and 0 otherwise (binary model). Alternatively,  $\delta_{i,j}$  can be defined as the actual number of contacts between the two 20-kb regions [as defined by Dixon *et al.* (2012)] to which the genes belong (proportional model). Parameter  $\lambda$  was estimated jointly with other model parameters, it measures the strength of the genome spatial correlation. Models were compared using Akaike’s information criterion (AIC). We find that the proportional correlation model fitted the data better and therefore selected it for further analyses.

### TFs and histone marks

TF mapping data from the Ensembl regulatory build (Zerbino *et al.* 2015) were obtained via the *biomaRt* package for R. We used the Grch37 build as it contained data for stem cell epigenomes. Genes were considered to be associated with a given TF when at least one binding evidence was present in the 3-kb upstream flanking region. TFs associated with more than five genes for which transcriptional noise could be computed were not considered further. A similar mapping was performed for histone marks by counting the evidence of histone modifications in the 3-kb upstream and downstream regions of each gene. A logistic principal component analysis (PCA) was conducted on the resulting binary contingency tables using the *logisticPCA* package for R (Landgraf and Lee 2015), for TF and histone marks separately. Principal components (PCs) were used to define synthetic variables for further analyses.

### Biological pathways, PPIs, and network topology

We defined genes either in the top 10% least noisy or in the top 10% most noisy as candidate sets, and used the *Reactome PA* package (Yu and He 2016) to search the mouse Reactome database for overrepresented pathways with a 1% false discovery rate (FDR).

Centrality measures were computed using a combination of the *igraph* (Csardi and Nepusz 2006) and *graphite* (Sales *et al.* 2012) packages for R. As the calculation of assortativity does not handle missing data (that is, nodes of the pathway for which no value could be computed), we computed assortativity on the subnetwork with nodes for which data were available. Reactome centrality measures could be computed for a total of 4454 genes with expression data.

PPIs were retrieved from the iRefIndex database (Razick *et al.* 2008) using the *iRefR* package for R (Mora and Donaldson 2011). Interactions were converted to a graph using the dedicated R functions in the package, and the same methods were used to compute centrality measures as for the pathway analysis. Because the PPI-based graph was not oriented, authority scores were not computed for this data (as this gave identical results to hub scores). Furthermore, as most genes are part of a single graph structure in the case of PPIs, closeness values were not further analyzed as they were virtually identical for all genes.

### Gene ontology enrichment

Of the 13,660 genes, 8325 were associated with Gene Ontology (GO) terms. We tested genes for GO term enrichment at both ends of the  $F^*$  spectrum using the same threshold percentile of 10% low/high-noise genes as we did for the Reactome analysis. We carried out GO enrichment analyses using two different algorithms implemented in the */topGO/* R package.: “Parent-child” (Grossmann *et al.* 2007) and “Weight01,” a mixture of two algorithms developed by Alexa *et al.* (2006). We kept only the terms that appeared simultaneously on both Parent-child and Weight01 at under a 1% significance level, controlling for multiple testing using the FDR method (Benjamini and Hochberg 1995).

### Sequence divergence

Ensembl’s Biomart interface was used to retrieve the proportion of nonsynonymous ( $K_a$ ) and synonymous ( $K_s$ ) divergence estimates for each mouse gene relative to the human ortholog. This information was available for 13,124 genes.

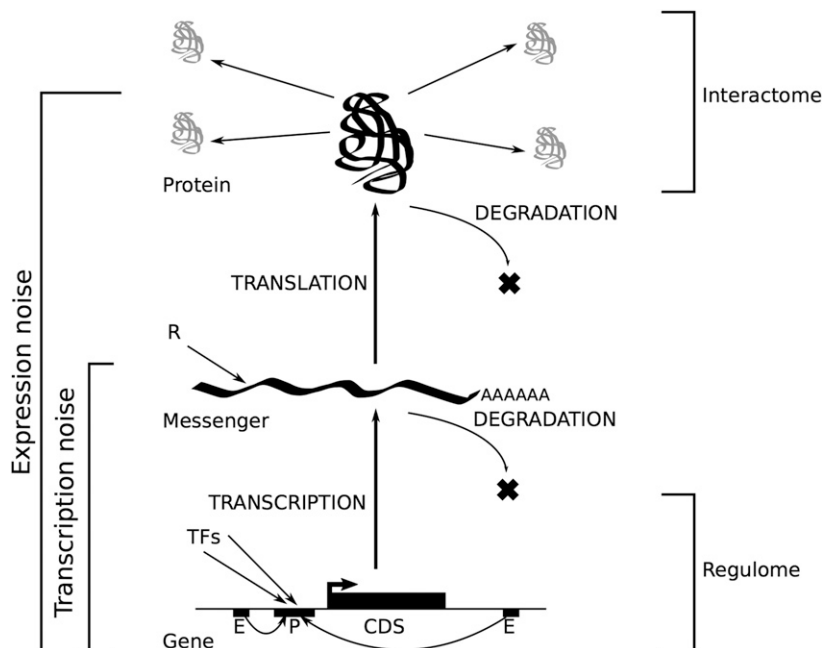
### Gene age

The relative taxonomic ages of the mouse genes have been computed and are available in the form of 20 phylostrata (Neme and Tautz 2013). Each phylostratum corresponds to a node in the phylogenetic tree of life. Phylostratum 1 corresponds to “All cellular organisms” whereas phylostratum 20 corresponds to “*Mus musculus*,” with other levels in between. We used this published information to assign each of our genes to a specific phylostratum and used this as a relative measure of gene age:  $Age = 21 - phylostratum$ , so that an age of 1 corresponds to genes specific to *M. musculus* and genes with an age of 20 are found in all cellular organisms.

### Linear modeling

We simultaneously assessed the effect of different factors on transcriptional noise by fitting linear models to the gene-specific  $F^*$  estimates. To avoid collinearity issues of intrinsically





**Figure 1** A systemic view of gene expression. CDS, coding sequence; TFs, transcription factors.

correlated explanatory variables, we conducted a data reduction procedure using multivariate analysis. We used variants of PCA on explanatory variables in three groups: network centrality measures, Ka/Ks and gene age with standard PCA, and TF-binding evidence and histone methylation patterns using logistic PCA, a generalization of PCA for binary variables (Landgraf and Lee 2015). In each case, we used the most representative components (totaling  $\geq 75\%$  of the total deviance) as synthetic variables. PCA analysis was conducted using the *ade4* package for R (Dray and Dufour 2007) and logistic PCA was performed using the *logisticPCA* package (Landgraf and Lee 2015).

We built a linear model with  $F^*$  as a response variable and 13 synthetic variables as explanatory variables. As the synthetic variables are PCs, they are orthogonal by construction. The fitted model displayed a significant departure to normality and was further transformed using the Box-Cox procedure [“boxcox” function from the *MASS* package for R (Venables and Ripley 2002)]. Residues of the selected model had normal, independent residue distributions (Shapiro–Wilk test of normality,  $P$ -value = 0.121; Ljung–Box test of independence,  $P$ -value = 0.2061) but still displayed significant heteroscedasticity (Harrison–McCabe test,  $P$ -value = 0.003). To ensure that this departure from the Gauss–Markov assumptions does not bias our inference, we used the “robcov” function of the *rms* package to get robust estimates of the effect significance (Harrell 2015). The relative importance of each explanatory factor was assessed using the method of Lindeman, Merenda, and Gold (Lindeman *et al.* 1979), as implemented is the R package *relaimpo*. The significance of the level of variance explained by each factor was computed using a standard ANOVA procedure.

#### Additional data sets

The aforementioned analyses were additionally conducted on the bone marrow-derived dendritic cell (BMDC) data set of

Shalek *et al.* (2014). Following the filtering procedure established by the authors in the original paper, genes that did not satisfied the condition of being expressed by an amount such that  $\log(\text{TPM} + 1) > 1$  in at least one of the 95 single cells were further discarded, where TPM stands for transcripts per million. This cut-off threshold resulted in 11,640 genes being kept for investigation. The rest of the analyses were conducted in the same way as for the ESC data set.

#### Jackknife procedure

A jackknife procedure was conducted to assess: (1) the robustness of our results to the choice of actual cells used to estimate mean and variance in gene expression and (2) the power of the pooled RNA sequencing analysis for which only three replicates were available. This analysis was conducted by sampling 3, 5, 10, and 15 of the original 20 single cells of the ESC data set (Sasagawa *et al.* 2013), 1000 times in each case. The exact same analysis was conducted on each random sample as for the complete data set, and model coefficients and their associated  $P$ -values were recorded.

#### Data availability

All data sets and scripts to reproduce the results of this study are available under the DOI 10.6084/m9.figshare.4587169.

## Results

### A new measure of noise to study genome-wide patterns of SGE

We used the data set generated by Sasagawa *et al.* (2013), which quantifies gene-specific amounts of mRNA as FPKM values for each gene and each individual cell. Among these, we selected all genes in a subset containing 20 ESCs in G1 phase to avoid recording variance that is due to different cell

types or cell-cycle phases. The Quartz-Seq sequencing protocol captures every poly-A RNA present in the cell at one specific moment, allowing the assessment of transcriptional noise. Following Shalek *et al.* (2014), we first filtered out genes that were not appreciably expressed to reduce the contribution of “technical” noise to the total noise. For each gene, we further calculated the mean  $\mu$  in FPKM units and variance  $\sigma^2$  in FPKM<sup>2</sup> units, as well as two previously published measures of stochasticity: the Fano factor, usually referred to as the bursty parameter, defined as  $\sigma^2/\mu$ , and noise, defined as the coefficient of variation squared ( $\sigma^2/\mu^2$ ). Both the variance and Fano factor are monotonically increasing functions of the mean (Figure 2A). Noise is inversely related to mean expression (Figure 2A), in agreement with previous observations at the protein level (Bar-Even *et al.* 2006; Taniguchi *et al.* 2011). While this negative correlation was theoretically predicted (Tao *et al.* 2007), it may confound the analyses of transcriptional noise at the genome level, because mean gene expression is under specific selective pressure (Pál *et al.* 2001). To disentangle these effects, we developed a new quantitative measure of noise, independent of the mean expression level of each gene. To achieve this, we performed polynomial regressions in the log-space plot of variance vs. mean. We defined  $F^*$  as  $\sigma_{obs}^2/\sigma_{pred}^2$  (see *Materials and Methods*), that is, the ratio of the observed variance over the variance component predicted by the mean expression level. We selected the simplest model for which no correlation between  $F^*$  and mean expression was observed, and found that a degree 3 polynomial model was sufficient to remove further correlation (Kendall’s  $\tau = -0.0037$ ,  $P$ -value = 0.5217, Figure 2A). Genes with  $F^* < 1$  have a variance lower than expected according to their mean expression, whereas genes with  $F^* > 1$  behave the opposite way (Figure 2B). This approach fulfills the same goal as the running median approach of Newman *et al.* (2006), while it includes the effect of mean expression directly into the measure of stochasticity instead of correcting *a posteriori* a dependent measure (in that case, the Fano factor). We therefore use  $F^*$  as a measure of SGE throughout this study.

### **SGE correlates with the 3D structure of the genome**

We first sought to investigate whether genome organization significantly impacts the patterns of SGE. We assessed whether genes in proximity along chromosomes display more similar amounts of transcriptional noise than distant genes. We tested this hypothesis by computing the primary distance on the genome between each pair of genes, that is, the number of base pairs separating them on the chromosome, as well as the relative difference in their transcriptional noise (see *Materials and Methods*). We found no significant association between the two distances (Mantel tests, each chromosome tested independently). However, contiguous genes had significantly more similar transcriptional noise than noncontiguous genes (permutation test,  $P$ -value  $< 1 \times 10^{-04}$ , Figure S1). Using Hi-C data from mouse embryonic cells (Dixon *et al.* 2012), we report that genes in contact in three dimensions

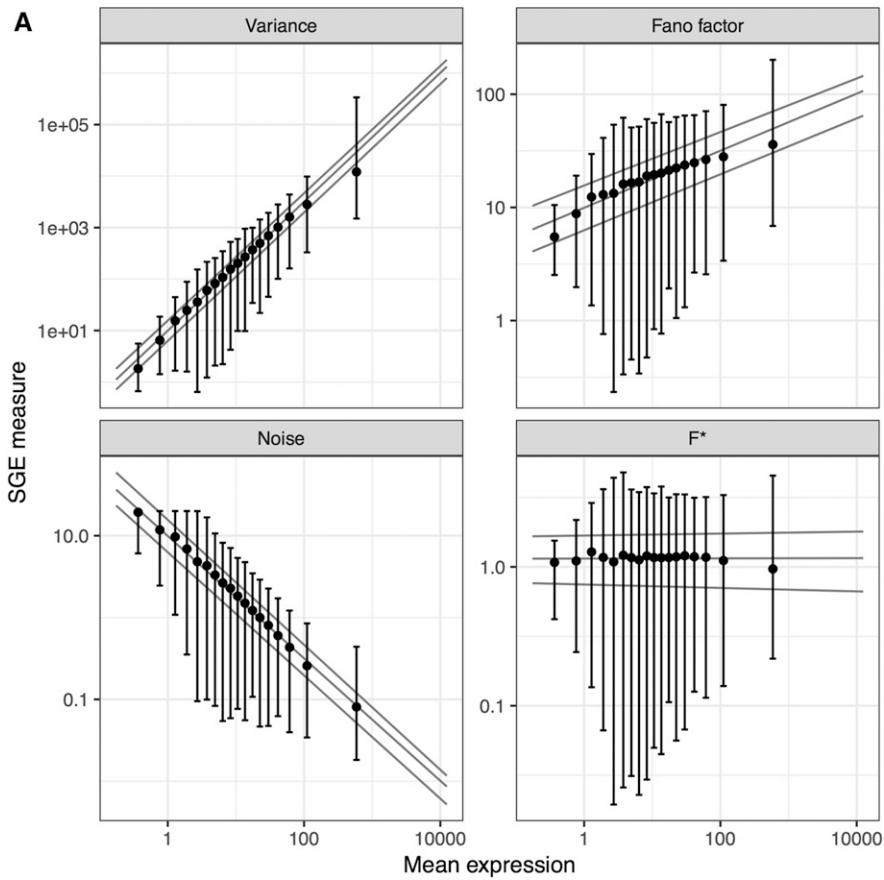
have significantly more similar transcriptional noise than genes not in contact (permutation test,  $P$ -value  $< 1 \times 10^{-03}$ , Figure S1). Most contiguous genes in one dimension also appear to be close in three dimensions, and the effect of 3D contact is stronger than that of 1D contact. These results therefore suggest that the 3D structure of the genome has a stronger impact on SGE than the position of the genes along the chromosomes. We further note that while highly significant, the size of this effect is small, with a mean difference in relative expression of  $-1.10\%$  (Figure S1).

### **TF binding and histone methylation impact SGE**

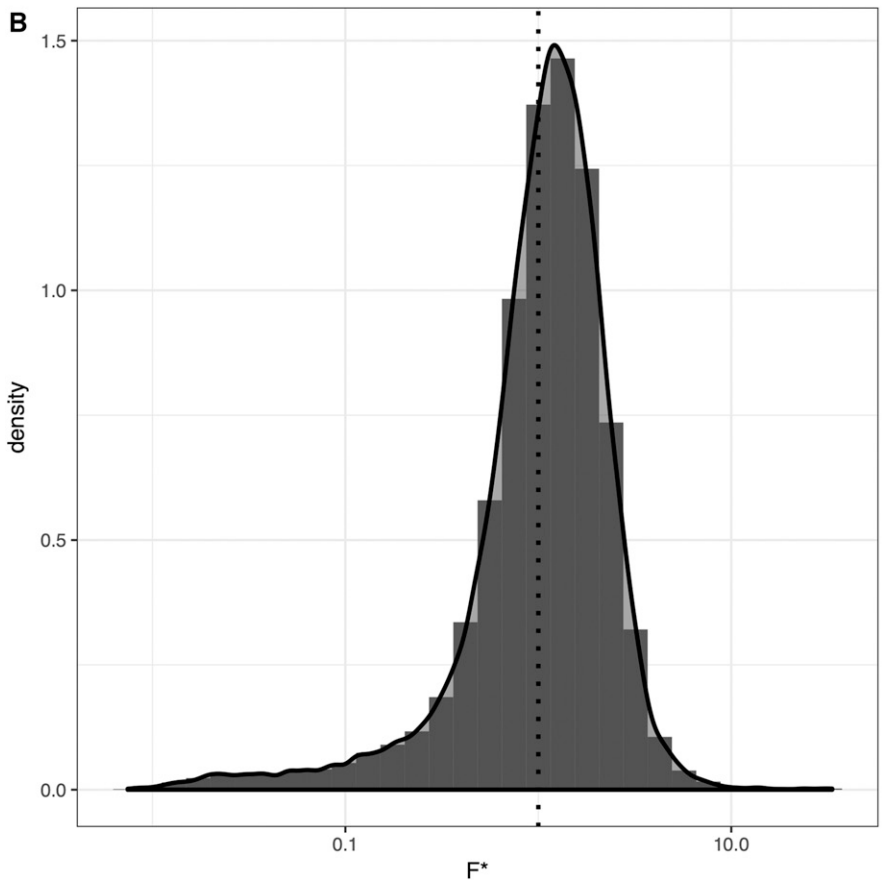
The binding of TFs to promoters constitutes one notable source of transcriptional noise (Figure 1) (Blake *et al.* 2003; Newman *et al.* 2006). In eukaryotes, the accessibility of promoters is determined by the chromatin state, which is itself controlled by histone methylation. We assessed the extent to which transcriptional noise is linked to particular TFs and histone marks by using data from the Ensembl regulatory build (Zerbino *et al.* 2015), which summarizes experimental evidence of TF binding and methylation sites along the genome. First, we contrasted the  $F^*$  values of genes with binding evidence for each annotated TF independently. Among 13 TFs represented by at least five genes in our data set, we found that four of them significantly influence  $F^*$  after adjusting for a global FDR of 5%: the transcription repressor *CTFC* (adjusted  $P$ -value = 0.0321), the TF CP2-like 1 (*Tcfcp2l1*, adjusted  $P$ -value = 0.0087), the X-linked Zinc Finger Protein (*Zfx*, adjusted  $P$ -value = 0.0284), and the *Myc* TF (*MYC*, adjusted  $P$ -value = 0.0104). Interestingly, association with each of these four TFs led to an increase in transcriptional noise. We also report a weak but significant positive correlation between the number of TFs associated with each gene and the amount of transcriptional noise (Kendall’s  $\tau = 0.0238$ ,  $P$ -value = 0.0007). This observation is consistent with the idea that noise generated by each TF is cumulative (Sharon *et al.* 2014). We then tested if particular histone marks are associated with transcriptional noise. Among five histone marks represented in our data set, three were found to be highly significantly associated to a higher transcriptional noise: H3K4me3 (adjusted  $P$ -value =  $2.0 \times 10^{-146}$ ), H3K4me2 (adjusted  $P$ -value =  $5.5 \times 10^{-121}$ ), and H3K27me3 (adjusted  $P$ -value =  $5.3 \times 10^{-34}$ ). Methylation on the fourth lysine of histone H3 is associated with gene activation in humans, while trimethylation on lysine 27 is usually associated with gene repression (Barski *et al.* 2007). These results suggest that both gene activation and silencing contribute to the stochasticity of gene expression, in agreement with the view that bursty transcription leads to increased noise (Blake *et al.* 2003; Newman *et al.* 2006; Batada and Hurst 2007).

### **Low noise genes are enriched for housekeeping functions**

We investigated the function of genes at both ends of the  $F^*$  spectrum. We defined as candidate gene sets the top 10%



**Figure 2** Transcriptional noise and mean gene expression. (A) Measures of noise plotted against the mean gene expression for each gene, in logarithmic scales: Variance, Fano factor (variance/mean), noise (square of the coefficient of variation, variance/mean<sup>2</sup>), and F\* (this study). Lines represent quantile regression fits (median, first, and third quartiles). Point and bars represent median, first, and third quartiles for each category of mean expression obtained by discretization of the x-axis. (B) Distribution of F\* over all genes in this study. Vertical line corresponds to F\* = 1. SGE, stochastic gene expression.



**Table 1 GO terms significantly enriched in the 10% genes with lowest transcriptional noise**

Ontology	GO ID	GO term	FDR Fisher “parent–child”	FDR Fisher “weight01”
MF	GO:0003735	Structural constituent of ribosome	$2.28 \times 10^{-07}$	$6.81 \times 10^{-20}$
MF	GO:0003676	Nucleic acid binding	$8.16 \times 10^{-06}$	$6.06 \times 10^{-04}$
BP	GO:0006412	Translation	$4.08 \times 10^{-08}$	$7.15 \times 10^{-12}$
BP	GO:0002227	Innate immune response in mucosa	$6.49 \times 10^{-04}$	$6.22 \times 10^{-03}$
CC	GO:0022625	Cytosolic large ribosomal subunit	$4.48 \times 10^{-03}$	$1.40 \times 10^{-12}$

GO, Gene Ontology; ID, identifier; FDR, False Discovery Rate; MF, Molecular Function; BP, Biological Process; CC, Cellular Compartment.

least noisy or the top 10% most noisy genes in our data set, and tested for enrichment of GO terms and Reactome pathways (see *Materials and Methods*). It is expected that genes encoding proteins participating in housekeeping pathways are less noisy because fluctuations in the concentrations of their products might have stronger deleterious effects (Pedraza and van Oudenaarden 2005). On the other hand, SGE could be selectively advantageous for genes involved in immune and stress responses, as part of a bet-hedging strategy (e.g., Arkin *et al.* 1998; Shalek *et al.* 2013). A GO terms enrichment test revealed significant categories enriched in the low-noise gene set only: molecular functions “nucleic acid binding” and “structural constituent of ribosome;” the biological processes “nucleosome assembly,” “innate immune response in mucosa,” and “translation;” and the cellular component “cytosolic large ribosomal subunit” (Table 1). All these terms but one relate to gene expression, in agreement with previously reported findings in yeast (Newman *et al.* 2006). We further find a total of 41 Reactome pathways significantly overrepresented in the low-noise gene set (FDR set to 1%). Interestingly, the most significant pathways belong to modules related to translation (RNA processing, initiation of translation, and ribosomal assembly), as well as several modules relating to gene expression, including chromatin regulation and mRNA splicing (Figure 3). Only one pathway was found to be enriched in the high-noise set: *TP53* regulation of transcription of cell cycle genes ( $P$ -value = 0.0079). This finding is interesting because *TP53* is a central regulator of the stress response in the cell (Hussain and Harris 2006). These results therefore corroborate previous findings that genes involved in the stress response might be evolving under selection for high noise as part of a bet-hedging strategy (Shalek *et al.* 2013; Viney and Reece 2013). The small amount of significantly enriched Reactome pathways by high-noise genes can potentially be explained by the nature of the data set: as the original experiment was based on unstimulated cells, genes that directly benefit from high SGE might not be expressed under these experimental conditions.

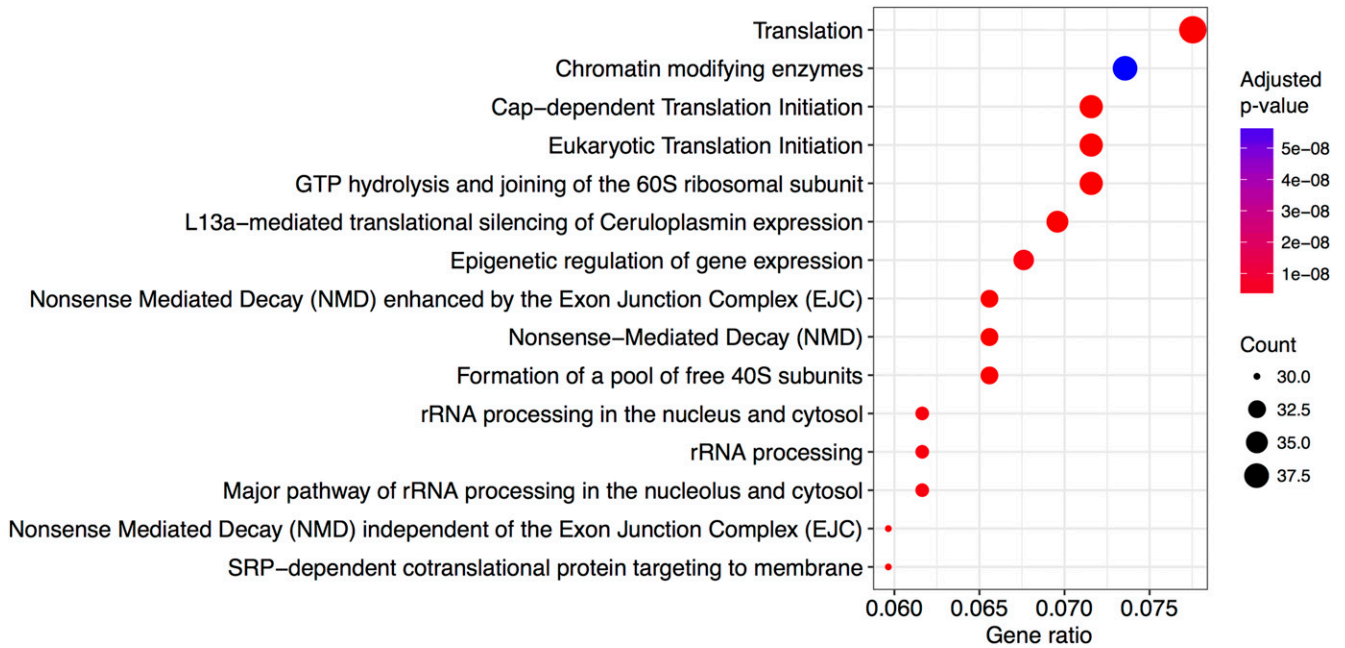
#### **Highly connected proteins are synthesized by low-noise genes**

The structure of the interaction network of proteins inside the cell can greatly impact the evolutionary dynamics of genes (Jeong *et al.* 2000; Barabási and Oltvai 2004). Furthermore, the contribution of each constitutive node within a given network varies. This asymmetry is largely reflected in the

power-law-like degree distribution that is observed in virtually all biological networks (Barabási and Albert 1999), with a few genes displaying many connections and a majority of genes displaying only a few. The individual characteristics of each node in a network can be characterized by various measures of centrality (Newman 2003). Following previous studies on protein evolutionary rate (Fraser *et al.* 2002; Hahn *et al.* 2004; Jovelin and Phillips 2009) and PPI networks (Li *et al.* 2010), we asked whether, at the gene level, there is a link between the centrality of a protein and the amount of transcriptional noise. We study six centrality metrics measured on two types of network data: (1) pathway annotations from the Reactome database (Fabregat *et al.* 2016) and (2) PPI data from the iRefIndex database. PPI data are typically more complete (5553 genes with gene expression data) but do not include information on functional interactions. The Reactome database is based on published functional evidence, but encompasses less genes (4454 genes for which expression data are available). In addition, graphs representing PPI networks are not oriented while graphs representing Pathway annotations are, implying that distinct statistics can be computed on both types of networks.

We first estimated the pleiotropy index of each gene by counting how many different pathways the corresponding proteins are involved in. We then computed centrality measures as averages over all pathways in which each gene is involved. These measures include: (1) node degree, which corresponds to the number of other nodes a given node is directly connected with; (2) hub score, which estimates the extent to which a node links to other central nodes; (3) authority score, which estimates the importance of a node by assessing how many hubs link to it; (4) transitivity, or clustering coefficient, defined as the proportion of neighbors that also connect to each other; (5) closeness, a measure of the topological distance between a node and every other reachable node (the fewer edge hops it takes for a protein to reach every other protein in a network, the higher its closeness); and (6) betweenness, a measure of the frequency with which a protein belongs to the shortest path between every pair of nodes.

We find that node degree, hub score, authority score; and transitivity are all significantly negatively correlated with transcriptional noise on pathway-based networks: the more central a protein is, the less transcriptional noise it displays (Figure 4, A–D and Table 2). We also observed that pleiotropy is negatively correlated with  $F^*$  (Kendall’s  $\tau = -0.0514$ ,



**Figure 3** Enriched pathways in the low-noise gene set. Depicted pathways are the 15 most significant in the 10% of genes with lowest transcriptional noise.

$P$ -value =  $8.31 \times 10^{-07}$ , Figure 4E and Table 2), suggesting that a protein that potentially performs multiple functions at the same time needs to be less noisy. As pleiotropic genes are themselves more central (e.g., correlation of pleiotropy and node degree: Kendall's  $\tau = 0.2215$ ,  $P$ -value  $< 2.2 \times 10^{-16}$ ) and evolve more slowly (correlation of pleiotropy and  $K_a/K_s$  ratio: Kendall's  $\tau = -0.1060$ ,  $P$ -value  $< 2.2 \times 10^{-16}$ ), we controlled for these variables and found consistent results (partial correlation of pleiotropy and  $F^*$ , accounting for centrality measures and  $K_a/K_s$ : Kendall's  $\tau = -0.0254$ ,  $P$ -value =  $7.45 \times 10^{-06}$ ). Closeness and betweenness, on the other hand, show a negative correlation with  $F^*$ , yet this was much less significant (Kendall's  $\tau = -0.0254$ ,  $P$ -value = 0.0109 for closeness and  $\tau = -0.0175$ ,  $P$ -value = 0.0865 for betweenness, see Figure 4, F and G and Table 2). In modular networks (Hartwell *et al.* 1999), nodes that connect different modules are extremely important to the cell (Guimera and Amaral 2005) and show high betweenness scores. In yeast, high betweenness proteins tend to be older and more essential (Joy *et al.* 2005), an observation also supported by our data set (betweenness *vs.* gene age, Kendall's  $\tau = 0.0619$ ,  $P$ -value =  $1.09 \times 10^{-07}$ ; betweenness *vs.*  $K_a/K_s$ , Kendall's  $\tau = -0.0857$ ,  $P$ -value =  $3.83 \times 10^{-16}$ ). However, it has been argued that in PPI networks, high betweenness proteins are less essential due to the lack of directed information flow, compared to, for instance, regulatory networks (Yu *et al.* 2007), a hypothesis that could explain the observed lack of correlation.

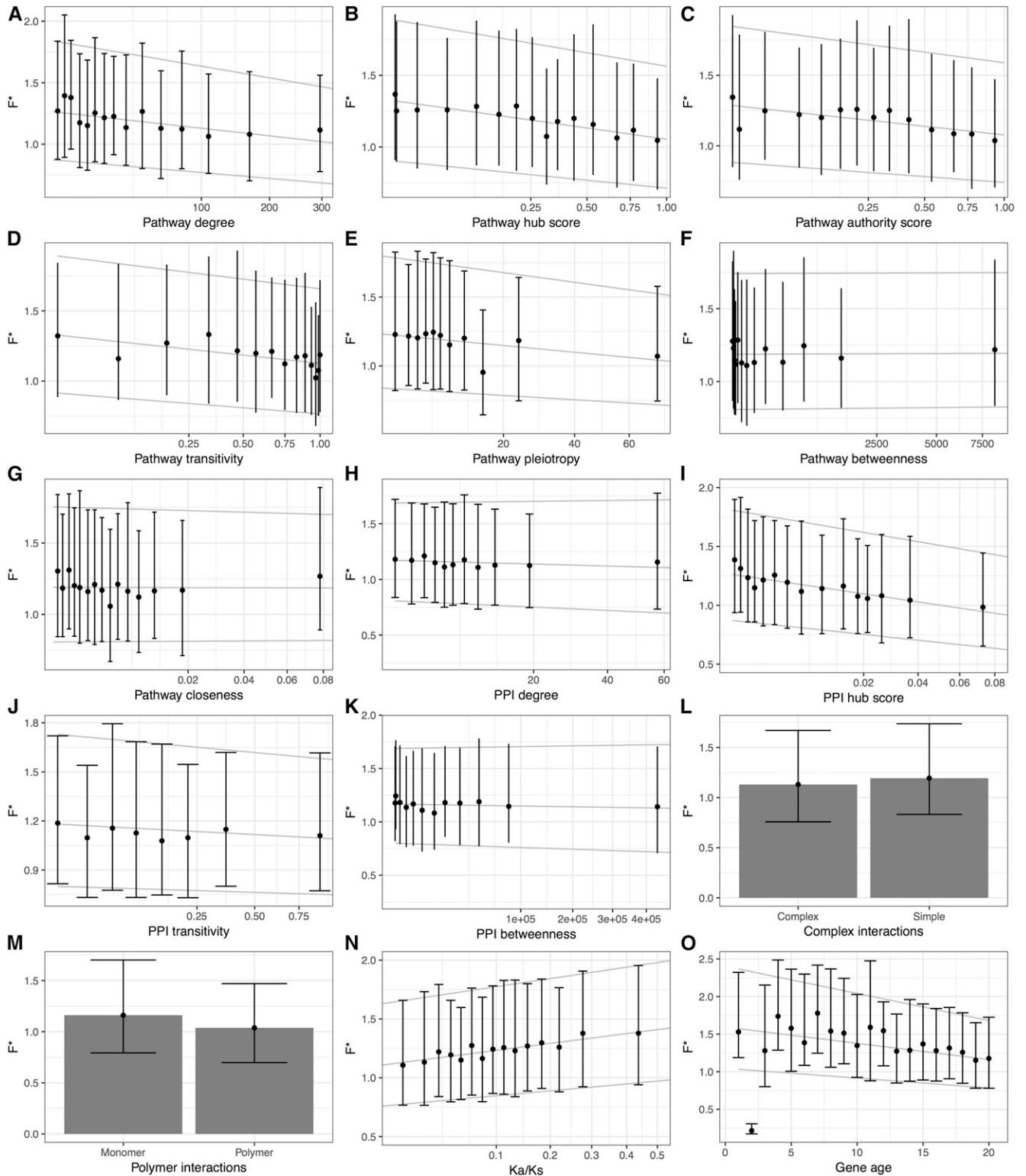
By applying similar measures on the PPI network, we report significant negative correlations between  $F^*$  and PPI centrality measures (Figure 4, H–K and Table 2). Because the PPI network is not directed, authority scores and hub scores cannot be distinguished. The results obtained with the mouse

PPI interaction network are qualitatively similar to the ones obtained by Li *et al.* (2010) on Yeast expression data (Li *et al.* 2010). In addition, we further report that genes involved in complex interactions (that is, genes that interact with more than one other protein simultaneously) have reduced noise in gene expression (Wilcoxon rank test,  $P$ -value =  $8.053 \times 10^{-05}$ , Figure 4L), corroborating previous findings in Yeast (Fraser *et al.* 2004). Conversely, genes involved in polymeric interactions, that is, where multiple copies of the encoded protein interact with each other, did not show significantly different noise than other genes (Wilcoxon rank test,  $P$ -value = 0.0821, Figure 4M).

It was previously shown that centrality measures negatively correlate with evolutionary rate (Hahn and Kern 2004). Our results suggest that central genes are selectively constrained for their transcriptional noise, and that centrality therefore also influences the regulation of gene expression. Interestingly, it has been reported that central genes tend to be more duplicated (Vitkup *et al.* 2006). The authors proposed that such duplication events would have been favored as they would confer greater robustness to deleterious mutations in proteins. Our results are compatible with another nonexclusive, possible advantage: having more gene copies could reduce transcriptional noise by averaging the number of transcripts produced by each gene copy (Raser and O'Shea 2005).

### Network structure impacts transcriptional noise of constitutive genes

Whereas estimators of node centrality highlight gene-specific properties inside a given network, measures at the whole-network level enable the comparison of networks with distinct



**Figure 4** Factors driving stochastic gene expression. Correlation of  $F^*$  and all tested network centrality measures (A-G: pathway networks, H-M: protein-protein interaction networks), as well as protein conservation (Ka/Ks ratio) and gene age (N and O). Point and bars represent median, first, and third quartiles for each category of mean expression obtained by discretization of the  $x$ -axis, together with the quantile regression lines estimated on the full data set. PPI, protein-protein interaction.

properties. We computed the size, diameter, and global transitivity for each annotated network in our data set (1,364 networks, see Supplementary Material, File S1), which we compared

with the average  $F^*$  measure of all constitutive nodes. The size of a network is defined as its total number of nodes, while diameter is the length of the shortest path between the two



**Table 2 Correlation of transcriptional noise with gene centrality measures and pleiotropy, as estimated from pathway annotations and PPI networks**

Data	Measure	Correlation with	
		F*	P-value
Pathways	Degree	-0.0745	$1.14 \times 10^{-13***}$
	Hub score	-0.0808	$6.61 \times 10^{-16***}$
	Authority score	-0.0666	$2.72 \times 10^{-11***}$
	Clustering coefficient	-0.0794	$4.55 \times 10^{-15***}$
	Closeness	-0.0254	$1.09 \times 10^{-02*}$
	Betweenness	-0.0175	$8.65 \times 10^{-02}$
	Pleiotropy	-0.0514	$8.31 \times 10^{-07***}$
	Size	-0.0514	$3.91 \times 10^{-03***}$
	Diameter	0.0061	$7.55 \times 10^{-01}$ (NS)
	Global transitivity	-0.1532	$3.06 \times 10^{-17***}$
	PPI	Degree	-0.0249
Hub score		-0.0942	$< 2.2 \times 10^{-16***}$
Transitivity		-0.0338	$6.24 \times 10^{-04***}$
Betweenness		-0.0140	$1.31 \times 10^{-01}$ (NS)

All correlations are computed using Kendall's rank correlation test, with P-value codes defined as \*\*\* < 0.001 < \*\* < 0.01 < \* < 0.05 < · < 0.1. NS, non-significant; PPI, protein-protein interaction.

most distant nodes. Transitivity is a measure of connectivity, defined as the average of all nodes' clustering coefficients. Interestingly, while network size is positively correlated with average degree and transitivity (Kendall's  $\tau = 0.5880$ ,  $P$ -value <  $2.2 \times 10^{-16}$  and Kendall's  $\tau = 0.1166$ ,  $P$ -value =  $1.08 \times 10^{-10}$ , respectively), diameter displays a positive correlation with average degree (Kendall's  $\tau = 0.2959$ ,  $P$ -value <  $2.2e-16$ ) but a negative correlation with transitivity (Kendall's  $\tau = -0.0840$ ,  $P$ -value =  $2.17 \times 10^{-05}$ ). This is because diameter increases logarithmically with size, that is, the addition of new nodes to large networks does not increase the diameter as much as addition to small networks. This suggests that larger networks are relatively more compact than smaller ones, and that their constitutive nodes are therefore more connected. We find that average transcriptional noise correlates negatively with network size (Kendall's  $\tau = -0.0514$ ,  $P$ -value = 0.0039), while being independent of the diameter (Kendall's  $\tau = 0.0061$ ,  $P$ -value = 0.7547 see Table 3). These results are in line with the node-based analyses, and show that the more connections a network has, the less stochastic the expression of the underlying genes is. This supports the view of Raser and O'Shea (2005), that the gene-extrinsic, pathway-intrinsic level is functionally pertinent and needs to be distinguished from the globally-extrinsic level.

We further asked whether genes with similar transcriptional noise tend to synthesize proteins that connect to each other (positive assortativity) in a given network or, on the contrary, tend to avoid each other (negative assortativity). We considered all Reactome pathways annotated to the mouse and estimated their respective F\* assortativity. We found the mean assortativity to be significantly negative, with a value of -0.1384 (one sample Wilcoxon rank test,  $P$ -value <  $2.2e-16$ ), meaning that proteins with different F\* values tend to connect with each other (Figure S3). Maslov and Sneppen (2002) reported a negative assortativity between hubs in PPI networks, which they hypothesized to be the result of selection for reduced vulnerability to

deleterious perturbations. However, in our data set, we find the assortativity of hub scores to be significantly positive (average of 0.1221, one sample Wilcoxon rank test,  $P$ -value =  $1.212 \times 10^{-12}$ , Figure S5), although with a large distribution of assortativity values. As we showed that hub scores correlate negatively with F\* (Table 2), we asked whether the assortativity of hub proteins can explain the assortativity of F\*. We found a significantly positive correlation between the two assortativity measures (Kendall's  $\tau = 0.2581$ ,  $P$ -value <  $2.2 \times 10^{-16}$ ). However, the relationship between the measures is not linear (Figure S5), suggesting a distinct relationship between hub score and F\* for negative and positive hub score assortativity. Negative assortativity of hub proteins contributes to a negative assortativity of SGE (Kendall's  $\tau = 0.2730$ ,  $P$ -value <  $2.2 \times 10^{-16}$ ), while the effect vanishes for pathways with positive hub score assortativity (Kendall's  $\tau = 0.0940$ ,  $P$ -value =  $3.135 \times 10^{-04}$ ). While assortativity of F\* is closer to 0 for pathways with positive assortativity of hub score, we note that it is still significantly negative (average = -0.0818, one sample Wilcoxon test with  $P$ -value <  $2.2 \times 10^{-16}$ ). These results suggest the existence of additional constraints that act on the distribution of noisy proteins in a network.

### **Transcriptional noise is positively correlated with the evolutionary rate of proteins**

In the yeast *Saccharomyces cerevisiae*, evolutionary divergence between orthologous coding sequences correlates negatively with fitness effect on knockout strains of the corresponding genes (Hirsh and Fraser 2001), demonstrating that protein functional importance is reflected in the strength of purifying selection acting on it. Fraser *et al.* (2004) studied transcription and translation rates of yeast genes and classified genes in distinct noise categories according to their expression strategies. They reported that essential genes display lower expression noise than the rest. Following these pioneering observations, we hypothesized that genes under strong purifying selection at the protein sequence level should also be highly constrained for their expression and therefore display a lower transcriptional noise. To test this hypothesis, we correlated F\* with the ratio of Ka/Ks, as measured by sequence comparison between mouse genes and their human orthologs, after discarding genes with evidence for positive selection ( $n = 5$ ). In agreement with our prediction, we report a significantly positive correlation between the Ka/Ks ratio and F\* (Figure 4N, Kendall's  $\tau = 0.0557$ ,  $P$ -value <  $1.143 \times 10^{-05}$ ), that is, highly constrained genes (low Ka/Ks ratio) display less transcriptional noise (low F\*) than fast-evolving ones. This result demonstrates that genes encoding proteins under strong purifying selection are also more constrained on their transcriptional noise.

### **Older genes are less noisy**

Evolution of new genes was long thought to occur via duplication and modification of existing genetic material ["evolutionary tinkering," (Jacob 1977)]. However, evidence for *de novo* gene emergence is becoming more and more common (Tautz and Domazet-Lošo 2011; Xie *et al.* 2012). *De novo*-created genes

**Table 3 Linear models of transcriptional noise with genomic and epigenomic factors**

	OLS			GLS		
	Coefficient	SE	P-value	Coefficient	SE	P-value
(Intercept)	0.1612	0.0781	0.0392*	0.1665	0.0663	0.0121*
PC1	0.0390	0.0065	< 0.0001***	0.0396	0.0065	< 0.0001***
PC2	-0.0048	0.0069	0.4854	-0.0048	0.0069	0.4838
PC3	-0.0526	0.0091	< 0.0001***	-0.0518	0.0092	< 0.0001***
PC4	-0.0102	0.0097	0.2905	-0.0109	0.0100	0.2773
PC5	0.0117	0.0106	0.2713	0.0123	0.0106	0.2456
PC6	-0.0152	0.0107	0.1536	-0.0152	0.0109	0.1623
PC7	0.0210	0.0102	0.0384*	0.0211	0.0110	0.0561
PC8	0.0100	0.0113	0.3778	0.0073	0.0114	0.5250
TFPC1	0.0028	0.0041	0.4912	0.0025	0.0034	0.4658
TFPC2	0.0025	0.0027	0.3664	0.0024	0.0026	0.3585
TFPC3	0.0032	0.0042	0.4513	0.0032	0.0037	0.3825
HistPC1	-0.0031	0.001	0.0015**	-0.0033	0.0010	0.0007***
HistPC2	-0.0027	0.0016	0.0846	-0.0029	0.0015	0.0566

All correlations are computed using Kendall's rank correlation test, with *P*-value codes defined as \*\*\* < 0.001 < \*\* < 0.01 < \* < 0.05 < . < 0.1. OLS, Ordinary Least Squares; GLS, Generalized Least Squares; Pathway PC1–8, principal components on centrality measures, protein conservation, and gene age; TFPC1–3, principal components of the logistic PCA on transcription factor binding evidence; HistPC1 and 2, principal components of the logistic PCA on histone modification marks.

undergo several optimization steps, including their integration into a regulatory network (Neme and Tautz 2013). We tested whether the historical process of incorporation of new genes into pathways impacts the evolution of transcriptional noise. We used the phylostratigraphic approach of Neme and Tautz (2013), which categorizes genes into 20 strata, to compute gene age and tested for a correlation with  $F^*$ . As older genes tend to be more conserved (Wolf *et al.* 2009), more central [according to the preferential attachment model of network growth (Jeong *et al.* 2000, 2001)], and more pleiotropic, we controlled for these confounding factors (Kendall's  $\tau = -0.0663$ , *P*-value =  $1.58 \times 10^{-37}$ ; partial correlation controlling for Ka/Ks ratio, centrality measures and pleiotropy level, Figure 4O). These results suggest that older genes are more deterministically expressed while younger genes are noisier. While we cannot rule out that functional constraints not fully accounted for by the Ka/Ks ratio could at least partially explain the correlation of gene age and transcriptional noise, we hypothesize that the observed correlation results from ancient genes having acquired more complex regulation schemes through time. Such schemes include, for instance, negative feedback loops, which have been shown to stabilize gene expression and reduce expression noise (Becskei and Serrano 2000; Thattai and Oudenaarden 2001).

#### **Position in the protein network is the main driver of transcriptional noise**

To jointly assess the effect of network topology, epigenomic factors, Ka/Ks ratio, and gene age, we modeled the patterns of transcriptional noise as a function of multiple predictive factors within the linear model framework. This analysis could be performed on a set of 2794 genes for which values were available jointly for all variables. To avoid collinearity issues because some of these variables are intrinsically correlated, we performed data reduction procedures prior to modeling. For continuous variables, including pathway and PPI network variables, Ka/Ks ratio, and gene age, we conducted a PCA and

used as synthetic measures the first eight PCs, explaining together > 80% of the total inertia (Figure S2A). The first PC (PC1) of the PCA analysis is associated with pathway centrality measures (degree, hub score, authority score, and transitivity, Figure S2B). The second PC (PC2) corresponds to PPI centrality measures (degree, hub score, and betweenness), while the third component (PC3) relates to gene age and Ka/Ks ratio. The fourth component (PC4) is associated with PPI complex interactions and transitivity. PC5 and PC6 are essentially associated with betweenness and closeness of the pathway network, PC7 with PPI polymeric interactions, and PC8 with pathway pleiotropy. As TFs and histone mark data are binary (presence/absence for each gene), we performed a logistic PCA for both types of variable (Landgraf and Lee 2015). For TFs, we selected the three first components (hereby denoted as TFPC), which explained 78% of deviance (Figure S3A). The loads on the first component (TFPC1) are all negative, meaning that TFPC1 captures a global correlation trend and does not discriminate between TFs. *Tcfcp2l1* appears to be the TF with the highest correlation to TFPC1. The second component TFPC2 is dominated by *TCFC* (positive loading) and *Oct4* (negative loading), while the third component TFPC3 is dominated by *Esrrb* (positive loading), *MYC*, *nMyc*, and *E2F1* (negative loadings, Figure S3B). For histone marks, the two first components (hereby noted HistPC) explained 95% of variance and were therefore retained (Figure S4A). HistPC1 is dominated by mark H3K27me3 linked to gene repression (negative loadings), and HistPC2 by marks H3K4me1 and H3K4me3 linked to gene activation (positive loadings, Figure S4A).

We fitted a linear model with  $F^*$  as a response variable and all 13 synthetic variables as explanatory variables. We find that PC1 has a significant positive effect on  $F^*$  (Table 3). As the loadings of the centrality measures on PC1 are negative (Figure S2C), this result is consistent with our finding of a negative correlation of pathway-based centrality measures



with  $F^*$ . PC3 has a highly significant negative effect on  $F^*$ , which is consistent with a negative correlation with gene age (positive loading on PC3) and a positive correlation with the Ka/Ks ratio (negative loading on PC3, Figure S2D). The last highly significant variable is the first PC of the logistic PCA on histone methylation patterns, HistPC1, which has a negative effect on  $F^*$ . Because the loadings are essentially negative on HistPC1, this suggests a positive effect of methylation, in particular the repressive H3K27me3. Altogether, the linear model with all variables explained 4.01% of the total variance (adjusted  $R^2$ ). This small value indicates either that gene idiosyncrasies largely predominate over general effects, or that our estimates of transcriptional noise have a large measurement error, or both. To compare the individual effects of each explanatory variable, we conducted a relative importance analysis. As a mean of comparison, we fitted a similar model with mean expression as a response variable. We find that pathway centrality measures (PC1 variable) account for 38% of the explained variance, while protein constraints and gene age (PC3) account for 32%. Chromatin state (HistPC1) accounts for another 15% of the variance (Figure 5). These results contrast with the model of mean expression, where HistPC1 and HistPC2 account for 51 and 9% of the explained variance, respectively, and PC1 and PC3 20 and 10% only (Figure 5). This suggests that: (1) among all factors tested, position in the protein network is the main driver of the evolution of gene-specific stochastic expression, followed by protein constraints and gene age, and (2) that different selective pressures act on the mean and cell-to-cell variability of gene expression.

We further included the effect of 3D organization of the genome to assess whether it could act as a confounding factor. We developed a correlation model that allowed for genes in contact to have correlated values of transcriptional noise. The correlation model was fitted together with the previous linear model in the GLS framework. This new model allows for one additional parameter,  $\lambda$ , which captures the strength of correlation due to 3D organization of the genome (see *Materials and Methods*). The estimate of  $\lambda$  was found to be 0.0016, which means that the spatial autocorrelation of transcriptional noise is low on average. This estimate is significantly higher than zero, and model comparison using AIC favors the linear model with 3D correlation (AIC = 4880.858 vs. AIC = 4890.396 for a linear model without 3D correlation). Despite the significant effect of 3D genome correlation, our results were qualitatively and quantitatively very similar to the model ignoring 3D correlation (Table 3).

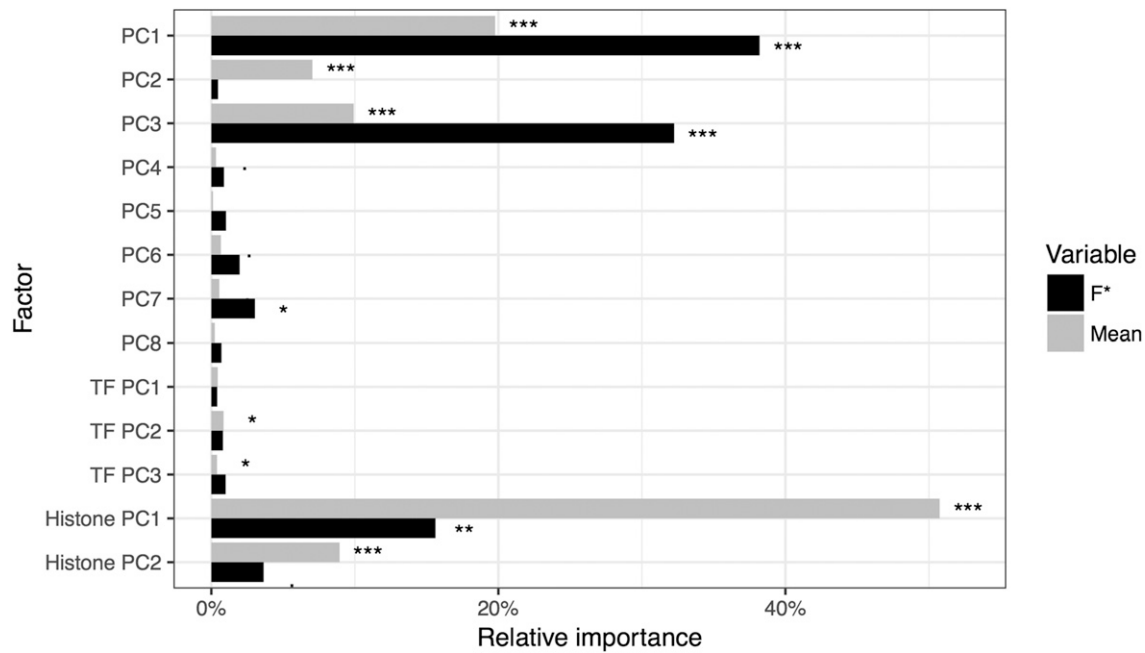
### **Analysis of BMDCs supports the generality of the results**

We assessed the reproducibility of our results by analyzing an additional single-cell transcriptomics data set of 95 unstimulated BMDCs (Shalek *et al.* 2014). After filtering (see *Materials and Methods*), the data set consisted of 11,640 genes. Using the same normalization procedure as for the ESC data set, we nonetheless report a weak but significant negative correlation between  $F^*$  and mean expression, even with a degree

5 polynomial regression ( $-0.0459$ ,  $P$ -value  $< 1.13e-13$ ). This effect is due to cell RPKM values being extremely skewed in this data set, due to the distribution per gene. To assess the impact of the residual correlation with the mean, we computed a value of  $F^*$  (noted  $F_R^*$ ) on a restricted data set where the variance was between one-eighth and eight times the mean (75% of all genes) using a quantile regression on the median instead of a linear regression. A second-degree polynomial quantile regression proved to be sufficient to remove the effect of mean expression (Kendall's  $\tau = 0.0114$ ,  $P$ -value = 0.1125) on this restricted data set. As all results were consistent when using the  $F_R^*$  and  $F^*$  measures, we only discuss here results obtained with  $F^*$  and refer to Supplementary Data 1 (available on FigShare under the DOI 10.6084/m9.figshare.4587169) for detailed results obtained with the  $F_R^*$  measure.

We report a highly significant positive correlation between  $F^*$  values measured on the 8792 genes with expression in both data sets, suggesting that cell-to-cell variance in gene expression is, to a large extent, conserved among the two cell types (Kendall's  $\tau = 0.1289$ ,  $P$ -value  $< 2.2 \times 10^{-16}$ , Figure S6A). GO terms or Reactome pathway enrichment analyses reveal less significant but consistent terms with the ESC analysis: the high- $F^*$  gene set did not show any significantly enriched GO term or Reactome pathway (FDR set to 1%) and the low- $F^*$  gene set revealed RNA binding as a significantly enriched molecular function, as well as 21 enriched pathways (Figure S7). In agreement with results from the ESC analysis, many of the most significantly enriched pathways relate to gene expression, including translation and splicing. Interestingly, the two most significant pathways are “Vesicle-mediated transport” and “Membrane trafficking,” two essential pathways for the functioning of dendritic cells. Analyses of network centrality measures also generally showed consistent results with the ESC data set, with more central genes displaying reduced gene expression noise (Figure S6, B–N and Table S1). Quantitative differences consisted of PPI betweenness, as well as pathway closeness and betweenness being highly significantly negatively correlated with  $F^*$  while they were only weakly significant or non-significant with the ESC data set. The only discrepancies that we report between the two data sets relate to pathway-level statistics. Pathway size appeared to be significantly positively correlated with mean  $F^*$ , while it was negatively correlated on the ESC data set, yet with a comparatively higher  $P$ -value. Similarly, pathway diameter was significantly positively correlated with mean  $F^*$  in the BMDC data set, while it was not significant with the ESC data. We currently have no hypothesis to explain this particular discrepancy. While these results support the generality of our observations, they also illustrate that, in detail, the fine structure of translational noise may vary in a cell type-specific manner.

We fitted linear models as for the ESC data set, with the exception that no epigenomic and 3D genome data were available for this cell type. Data reduction was performed using PCA, with the eight first PCs explaining 81% of the total deviance (Figure S8A). We report consistent results with the



**Figure 5** Relative importance of explanatory factors on mean gene expression and  $F^*$ . Significance codes refer to ANOVA test of variance: \*\*\* < 0.001 < \*\* < 0.01 < \* < 0.05 < · < 0.1. PC, principal component; TF, transcription factor.

ESC analysis, with all major effects similar in direction and intensity, highlighting the impact of network centrality measures on expression noise (Table S2). However, with the BMDC data, the PC2, which is associated with PPI centrality measures (Figure S8B), appears to have a significant negative impact on  $F^*$ , while it was not significant with the ESC data set. As the loading of the PPI centrality measures are positive on PC2, this is consistent with central genes having a lower transcriptional noise as for the pathway network metrics (Figure S8C). Relative importance analysis revealed that network centrality measures contributed most to the explained variance (48 and 21% for PC1 and PC2 respectively), while the contribution of protein constraints and gene age (PC3) was 24%.

### **Biological, not technical, noise is responsible for the observed patterns**

The noise in gene expression measured from single-cell transcriptomics is a combination of biological and technical noise. While the two sources of noise are *a priori* independent, gene-specific technical noise has been observed in microarray experiments (Pozhitkov *et al.* 2007), making a correlation of the two types of noise plausible. If similar effects also affect RNA sequencing experiments, technical noise could be correlated to gene function and therefore act as a covariate in our analyses. To assess whether this is the case, we used the data set of Shalek *et al.* (2013), which contains both single-cell transcriptomics and three replicates of 10,000 pooled-cell RNA sequencing. In traditional RNA sequencing, which is typically performed on pooled populations of several thousands of cells, biological noise is averaged out so that the resulting measured noise between replicates is essentially the result of technical noise. We computed the mean and variance in expression of each gene across

the three populations of cells. By plotting the variance vs. the mean in log-space, we were able to compute a technical  $F^*$  ( $F_t^*$ ) value for each gene (see *Materials and Methods*). We fitted linear models as for the single-cell data using  $F_t^*$  instead of  $F^*$ . We report that no variable had a significant effect on  $F_t^*$  (Table S3). In addition, there was no enrichment of the lower 10th percentile for any particular pathway or GO term. The upper 90th percentile showed no GO term enrichment, but four pathways appeared to be significant: “Chromosome maintenance” (adjusted  $P$ -value = 0.0043), “Polymerase switching on the C-strand of the telomere” (adjusted  $P$ -value = 0.0062), “Polymerase switching” (adjusted  $P$ -value = 0.0062), and “Leading strand synthesis” (adjusted  $P$ -value = 0.0062), which all relate to DNA replication. While it is unclear why genes involved in these pathways would display higher technical variance in RNA sequencing, these results differ strikingly from our analyses of single-cell RNA sequencing and therefore suggest that technical variance does not act as a confounding factor in our analyses.

Because only three replicates were available in the pooled RNA sequencing data set, we asked whether the resulting estimate of mean and variance in expression is accurate enough to allow proper inference of noise and its correlation with other variables. We conducted a jackknife procedure, where we sampled the original cells from the ESC data set and reestimated  $F^*$  for each sample. We tested combinations of 3, 5, 10, and 15 cells, with 1000 samples in each case. In each sample, we computed  $F^*$  with the same procedure as for the complete data set, and fitted a linear model with all 13 synthetic variables. For computational efficiency, we did not include 3D correlation in this analysis. We compute for each variable the number of samples where the effect is significant at the 5% level and has the same sign as in the model fitted on the full data set. We find that the model

coefficients are very robust to the number of cells used (Figure S9A) and that three cells are enough to infer the effect of the PC1 and PC3 variables, the most significant in our analyses. Two main conclusions can be drawn from this jackknife analysis: (1) that the lack of significant effect of our explanatory variables on technical noise is not due to the low number of replicates used to compute the mean and variance in expression, and (2) that our conclusions are very robust to the actual cells used in the analysis, ruling out drop-out and amplification biases as possible source of errors (Kharchenko *et al.* 2014).

## Discussion

Through this work, we provide the first genome-wide evolutionary and systemic study of transcriptional noise, using mouse cells as a model. We have shown that transcriptional noise correlates with functional constraints not only at the level of the gene itself via the protein it encodes, but also at the level of the pathway(s) the gene belongs to. We further discuss here potential confounding factors in our analyses and argue that our results are compatible with selection acting to reduce noise propagation at the network level.

In this study, we exhibited several factors explaining the variation in transcriptional noise between genes. While highly significant, the effects we report are of small size, and a complex model accounting for all tested sources of variation only explains a few percent of the total observed variance. There are several possible explanations for this reduced explanatory power. (1) Transcriptional noise is a proxy for noise in gene expression, at which selection occurs (Figure 1). As transcriptional noise is not randomly distributed across the genome, it must constitute a significant component of expression noise, in agreement with previous observations (Blake *et al.* 2003; Newman *et al.* 2006). However, translational noise might constitute an important part of the expression noise and was not assessed in this study. (2) Gene expression levels were assessed on ESCs in culture. Such an experimental system may result in gene expression that differs from that in natural conditions under which natural selection acted. (3) Functional annotations in particular pathways and gene interaction are incomplete, and network-based measures most likely have large error rates. (4) While the newly introduced  $F^*$  measure allowed us to assess the distribution of transcriptional noise independently of the average mean expression, it does not capture the full complexity of SGE. Explicit modeling, for instance based in the  $\beta$ -Poisson model (Vu *et al.* 2016), is a promising avenue for the development of more sophisticated quantitative measures.

In a pioneering study, Fraser *et al.* (2004), followed by Shalek *et al.* (2013), demonstrated that essential genes whose deletion is deleterious, and genes encoding subunits of molecular complexes as well as housekeeping genes, display reduced gene expression noise. Our findings go beyond these early observations by providing a statistical assessment of the joint effect of multiple explanatory factors. Our analyses reveal that network centrality measures are the explanatory factors that explain the most significant part of the distribution of transcriptional noise

in the genome. Network-based statistics were first tested by Li *et al.* (2010) using PPI data in Yeast. While we are able to extend these results to mouse cells, we show that more detailed annotation, as provided by the Reactome database, can lead to new insights into the selective forces acting on expression noise. Our results suggest that pathways constitute a relevant systemic level of organization, at which selection can act and drive the evolution of SGE at the gene level. This multi-level selection mechanism, we propose, can be explained by selection against noise propagation within networks. It has been experimentally demonstrated that expression noise can be transmitted from one gene to another with which it is interacting (Pedraza and van Oudenaarden 2005). Large noise at the network level is deleterious (Barkai and Leibler 1999) but each gene does not contribute equally to it, thus the strength of selective pressure against noise varies among genes in a given network. We have shown that highly connected, “central” proteins typically display reduced transcriptional noise. Such nodes are likely to constitute key players in the flow of noise in intracellular networks as they are more likely to transmit noise to other components. In accordance with this hypothesis, we find genes with the lowest amount of transcriptional noise to be enriched for top-level functions, particularly if they are involved in the regulation of other genes.

These results have several implications for the evolution of gene networks. First, this means that new connections in a network can potentially be deleterious if they link genes with highly stochastic expression. Second, distinct selective pressures at the “regulome” and “interactome” levels (Figure 1) might act in opposite directions. We expect genes encoding highly connected proteins to have more complex regulation schemes, particularly if their proteins are involved in several biological pathways. In accordance, several studies have demonstrated that expression noise of a gene positively correlates with the number of TFs controlling its regulation (Sharon *et al.* 2014), a correlation that we also find significant in the data set analyzed in this work. Central genes, while being under negative selection against stochastic behavior, are then more likely to be controlled by numerous TFs that increase transcriptional noise. As a consequence, if the number of connections at the interactome level is correlated with the number of connections at the regulome level, we predict the existence of a trade-off in the number of connections that a gene can make in a network. Alternatively, highly connected genes might evolve regulatory mechanisms allowing them to uncouple these two levels: negative feedback loops, for instance, where the product of a gene downregulates its own production, have been shown to stabilize expression and significantly reduce stochasticity (Becskei and Serrano 2000; Dublanche *et al.* 2006; Tao *et al.* 2007). We therefore predict that negative feedback loops are more likely to occur at genes that are more central in protein networks, as they will confer greater resilience against high SGE, which is advantageous for this class of genes.

Our results enabled the identification of possible selective pressures acting on the level of stochasticity in gene expression. However, the mechanisms by which the amount of stochasticity

can be controlled remain to be elucidated. We evoked the existence of negative feedback loops that reduce stochasticity and the multiplicity of upstream regulators that increase it. Recent work by Wolf *et al.* (2015) and Metzger *et al.* (2015) add further perspective to this scheme. Wolf and colleagues found that, in *Escherichia coli*, noise is higher for natural than experimentally evolved promoters selected for their mean expression level. They hypothesized that higher noise is selectively advantageous in cases of changing environments. On the other hand, Metzger and colleagues performed mutagenesis experiments and found signatures of selection for reduced noise in natural populations of *S. cerevisiae*. These seemingly opposing results, combined with our observations, provide additional evidence that the amount of stochasticity in the expression of single genes has an optimum, as high values are deleterious because of noise propagation in the network; while lower values, which result in reduced phenotypic plasticity, might be suboptimal in cases of dynamic environments.

### Conclusions

Using a new measure of transcriptional noise, our results demonstrate that the position of a protein in the interactome is a major driver of selection against SGE. As such, transcriptional noise is an essential component of the phenotype, in addition to the mean expression level and the actual sequence and structure of the encoded proteins. This is currently an underappreciated phenomenon, and gene expression studies that focus only on the mean expression of genes may be missing key information about expression diversity. The study of gene expression must consider changes in noise in addition to changes in mean expression level as a putative explanation for adaptation. However, further work that aims to unravel the exact structure of the regulome is needed to fully understand how transcriptional noise is generated or inhibited.

### Acknowledgments

The authors thank Rafiq Neme-Garrido, Frederic Bartels, and Estelle Renaud for fruitful discussions about this work; Andrew Landgraf for help with the logistic PCA analysis; and Diethard Tautz for comments on an earlier version of this manuscript. They also thank two anonymous reviewers and the editors for their constructive comments. J.Y.D. acknowledges funding from the Max Planck Society. This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft), within the priority program (SPP) 1590.

Author contributions: G.V.B. and J.Y.D. designed the experiments and wrote the manuscript. G.V.B., N.P., and J.Y.D. conducted the analyses.

### Literature Cited

Alexa, A., J. Rahnenführer, and T. Lengauer, 2006 Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22: 1600–1607.  
 Arkin, A., J. Ross, and H. H. McAdams, 1998 Stochastic kinetic analysis of developmental pathway bifurcation in phage  $\lambda$ -infected *Escherichia coli* cells. *Genetics* 149: 1633–1648.

Barabási, A.-L., and R. Albert, 1999 Emergence of scaling in random networks. *Science* 286: 509–513.  
 Barabási, A.-L., and Z. N. Oltvai, 2004 Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5: 101–113.  
 Bar-Even, A., J. Paulsson, N. Maheshri, M. Carmi, E. O. Shea *et al.*, 2006 Noise in protein expression scales with natural protein abundance. *Nat. Genet.* 38: 636–643.  
 Barkai, N., and S. Leibler, 1999 Circadian clocks limited by noise. *Nature* 403: 267–268.  
 Barski, A., S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones *et al.*, 2007 High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823–837.  
 Batada, N. N., and L. D. Hurst, 2007 Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat. Genet.* 39: 945–949.  
 Becskei, A., and L. Serrano, 2000 Engineering stability in gene networks by autoregulation. *Nature* 405: 590–593.  
 Becskei, A., B. B. Kaufmann, and A. van Oudenaarden, 2005 Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat. Genet.* 37: 937–944.  
 Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57: 289–300.  
 Blake, W. J., M. Kærn, C. R. Cantor, and J. J. Collins, 2003 Noise in eukaryotic gene expression. *Nature* 422: 633–637.  
 Chubb, J. R., T. Trcek, S. M. Shenoy, and R. H. Singer, 2006 Transcriptional pulsing of a developmental gene. *Curr. Biol.* 16: 1018–1025.  
 Csardi, G., and T. Nepusz, 2006 The igraph software package for complex network research. *InterJournal Complex Systems* 1695: 1695.  
 Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li *et al.*, 2012 Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485: 376–380.  
 Dray, S., and A.-B. Dufour, 2007 The ade4 Package: implementing the duality diagram for ecologists. *J. Stat. Softw.* 22. Available at: <https://www.jstatsoft.org/article/view/v022i04>.  
 Dublanche, Y., K. Michalodimitrakis, N. Kümmerer, M. Foglierini, and L. Serrano, 2006 Noise in transcription negative feedback loops: simulation and experimental analysis. *Mol. Syst. Biol.* 2: 41.  
 Eldar, A., and M. B. Elowitz, 2010 Functional roles for noise in genetic circuits. *Nature* 467: 167–173.  
 Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain, 2002 Stochastic gene expression in a single cell. *Science* 297: 1183–1186.  
 Fabregat, A., K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann *et al.*, 2016 The reactome pathway knowledgebase. *Nucleic Acids Res.* 44: D481–D487.  
 Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman, 2002 Evolutionary rate in the protein interaction network. *Science* 296: 750–752.  
 Fraser, H. B., A. E. Hirsh, G. Giaever, J. Kumm, and M. B. Eisen, 2004 Noise minimization in eukaryotic gene expression. *PLoS Biol.* 2: e137.  
 Gillespie, D. T., 1977 Exact simulation of coupled chemical reactions. *J. Phys. Chem.* 81: 2340–2361.  
 Grossmann, S., S. Bauer, P. N. Robinson, and M. Vingron, 2007 Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics* 23: 3024–3031.  
 Guimera, R., and L. A. N. Amaral, 2005 Functional cartography of complex metabolic networks. *Nature* 433: 895–900.  
 Hahn, M. W., and A. D. Kern, 2004 Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* 22: 7–10.



- Hahn, M. W., G. C. Conant, and A. Wagner, 2004 Molecular evolution in large genetic networks: does connectivity equal constraint? *J. Mol. Evol.* 58: 203–211.
- Harrell, F. E., 2015 *Regression Modeling Strategies*. Springer-Verlag, Heidelberg, Germany.
- Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray, 1999 From molecular to modular cell biology. *Nature* 402: C47–C52.
- Hebenstreit, D., 2013 Are gene loops the cause of transcriptional noise? *Trends Genet.* 29: 333–338.
- Hirsh, A., and H. Fraser, 2001 Protein dispensability and rate of evolution. *Nature* 411: 1046–1049.
- Hussain, S. P., and C. C. Harris, 2006 p53 biological network: at the crossroads of the cellular-stress response pathway and molecular carcinogenesis. *J. Nippon Med. Sch.* 73: 54–64.
- Jacob, F., 1977 Evolution and tinkering. *Science* 196: 1161–1166.
- Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, 2000 The large-scale organization of metabolic networks. *Nature* 407: 651–654.
- Jeong, H., S. P. Mason, A. L. Barabási, and Z. N. Oltvai, 2001 Lethality and centrality in protein networks. *Nature* 411: 41–42.
- Jovelin, R., and P. C. Phillips, 2009 Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol.* 10: R35.
- Joy, M. P., A. Brock, D. E. Ingber, and S. Huang, 2005 High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* 2005: 96–103.
- Kaufmann, B. B., and A. van Oudenaarden, 2007 Stochastic gene expression: from single molecules to the proteome. *Curr. Opin. Genet. Dev.* 17: 107–112.
- Kepler, T. B., and T. C. Elston, 2001 Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys. J.* 81: 3116–3136.
- Kharchenko, P. V., L. Silberstein, and D. T. Scadden, 2014 Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11: 740–742.
- Landgraf, A. J., and Y. Lee, 2015 Dimensionality reduction for binary data through the projection of natural parameters. *arXiv* Available at: <https://arxiv.org/abs/1510.06112>.
- Lehner, B., 2008 Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol. Syst. Biol.* 4: 170.
- Li, J., R. Min, F. J. Vizeacoumar, K. Jin, X. Xin *et al.*, 2010 Exploiting the determinants of stochastic gene expression in *Saccharomyces cerevisiae* for genome-wide prediction of expression noise. *Proc. Natl. Acad. Sci. USA* 107: 10472–10477.
- Lindeman, R. H., P. F. Merenda, and R. Z. Gold, 1979 *Introduction to Bivariate and Multivariate Analysis*. Scott Foresman & Co, Glenview, IL.
- Maslov, S., and K. Sneppen, 2002 Specificity and stability in topology of protein networks. *Science* 296: 910–913.
- McAdams, H. H., and A. Arkin, 1997 Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* 94: 814–819.
- Metzger, B. P. H., D. C. Yuan, J. D. Gruber, F. Duveau, and P. J. Wittkopp, 2015 Selection on noise constrains variation in a eukaryotic promoter. *Nature* 521: 344–347.
- Mora, A., and I. M. Donaldson, 2011 iRefR: an R package to manipulate the iRefIndex consolidated protein interaction database. *BMC Bioinformatics* 12: 455.
- Neme, R., and D. Tautz, 2013 Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. *BMC Genomics* 14: 117.
- Newman, J. R. S., S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble *et al.*, 2006 Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 441: 840–846.
- Newman, M. E. J., 2003 The structure and function of complex networks. *SIAM Rev.* 45: 167–256.
- Norman, T. M., N. D. Lord, J. Paulsson, and R. Losick, 2015 Stochastic switching of cell fate in microbes. *Annu. Rev. Microbiol.* 69: 381–403.
- Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman, and A. V. Oudenaarden, 2002 Regulation of noise in the expression of a single gene. *Nat. Genet.* 31: 69–73.
- Pál, C., B. Papp, and L. D. Hurst, 2001 Highly expressed genes in yeast evolve slowly. *Genetics* 158: 927–931.
- Pedraza, J. M., and A. van Oudenaarden, 2005 Noise propagation in gene networks. *Science* 307: 1965–1969.
- Pombo, A., and N. Dillon, 2015 Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* 16: 245–257.
- Pozhitkov, A. E., D. Tautz, and P. A. Noble, 2007 Oligonucleotide microarrays: widely applied—poorly understood. *Brief. Funct. Genomic. Proteomic.* 6: 141–148.
- Raj, A., and A. V. Oudenaarden, 2008 Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135: 216–226.
- Raser, J. M., and E. K. O’Shea, 2005 Noise in gene expression: origins, consequences, and control. *Science* 309: 2010–2013.
- Razick, S., G. Magklaras, and I. M. Donaldson, 2008 iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* 9: 405.
- Sales, G., E. Calura, D. Cavalieri, and C. Romualdi, 2012 Graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* 13: 20.
- Sánchez, A., and J. Kondev, 2008 Transcriptional control of noise in gene expression. *Proc. Natl. Acad. Sci. USA* 105: 5081–5086.
- Sasagawa, Y., I. Nikaido, T. Hayashi, H. Danno, K. D. Uno *et al.*, 2013 Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* 14: R31.
- Sauer, U., M. Heineman, and N. Zamboni, 2007 Getting closer to the whole picture. *Science* 316: 550–551.
- Shahrezaei, V., and P. S. Swain, 2008 The stochastic nature of biochemical networks. *Curr. Opin. Biotechnol.* 19: 369–374.
- Shalek, A. K., R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaubomme *et al.*, 2013 Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498: 236–240.
- Shalek, A. K., R. Satija, J. Shuga, J. J. Trombetta, D. Gennert *et al.*, 2014 Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510: 363–369.
- Sharon, E., D. Van Dijk, Y. Kalma, L. Keren, O. Manor *et al.*, 2014 Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res.* 24: 1698–1706.
- Spudich, J. L., and D. E. Koshland, Jr., 1976 Non-genetic individuality: chance in the single cell. *Nature* 262: 467–471.
- Suter, D. M., N. Molina, D. Gatfield, K. Schneider, U. Schibler *et al.*, 2011 Mammalian genes are transcribed with widely different bursting kinetics. *Science* 332: 472–474.
- Taniguchi, Y., P. J. Choi, G. Li, H. Chen, M. Babu *et al.*, 2011 Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329: 533–539.
- Tao, Y., X. Zheng, and Y. Sun, 2007 Effect of feedback regulation on stochastic gene expression. *J. Theor. Biol.* 247: 827–836.
- Tautz, D., and T. Domazet-Lošo, 2011 The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12: 692–702.
- Thattai, M., and A. V. Oudenaarden, 2001 Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. USA* 98: 8614–8619.
- Thattai, M., and A. V. Oudenaarden, 2004 Stochastic gene expression in fluctuating environments. *Genetics* 167: 523–530.
- Venables, W. N., and B. D. Ripley, 2002 *Modern Applied Statistics with S*. Springer, New York.
- Viney, M., and S. E. Reece, 2013 Adaptive noise. *Proc. Biol. Sci.* 280: 20131104.

- Vitkup, D., P. Kharchenko, and A. Wagner, 2006 Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* 7: R39.
- Vu, T. N., Q. F. Wills, K. R. Kalari, N. Niu, L. Wang *et al.*, 2016 Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* 32: 2128–2135.
- Wang, G.-Z., M. J. Lercher, and L. D. Hurst, 2011 Transcriptional coupling of neighboring genes and gene expression noise: evidence that gene orientation and noncoding transcripts are modulators of noise. *Genome Biol. Evol.* 3: 320–331.
- Wang, Z., and J. Zhang, 2011 Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Natl. Acad. Sci. USA* 108: E67–E76.
- Wolf, L., O. K. Silander, and E. J. van Nimwegen, 2015 Expression noise facilitates the evolution of gene regulation. *Elife* 4: 1–48.
- Wolf, Y. I., P. S. Novichkov, G. P. Karev, E. V. Koonin, and D. J. Lipman, 2009 The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl. Acad. Sci. USA* 106: 7273–7280.
- Xie, C., Y. E. Zhang, J. Y. Chen, C. J. Liu, W. Z. Zhou *et al.*, 2012 Hominoid-specific *de novo* protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* 8: e1002942.
- Yu, G., and Q.-Y. He, 2016 ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* 12: 477–479.
- Yu, H., P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, 2007 The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* 3: 713–720.
- Zerbino, D. R., S. P. Wilder, N. Johnson, T. Juettemann, and P. R. Flicek, 2015 The ensembl regulatory build. *Genome Biol.* 16: 56.

*Communicating editor: A. Moses*

# LIST OF FIGURES

## Introduction

1. A very simple Ancestral Recombination Graph..... 10

## Chapter 1

1. Schematic representation of iSMC ..... 21
2. Recombination map recovery under various simulated scenarios..... 24
3. Inference of recombination maps in the presence of recent population growth..... 25
4. Evolution of the recombination landscape in hominins..... 29

## Chapter 2

1. Schematic representation of iSMC ..... 43
2. Overview of the simulation study..... 46
3. Mutation map recovery under various simulated scenarios ..... 47
4. The genomic landscapes along chromosome 1 of *Zymoseptoria tritici*..... 49

## Chapter 3

1. A systemic view of gene expression..... 58
2. Transcriptional noise and mean gene expression.....61
3. Enriched pathways in the low-noise gene set..... 64
4. Factors driving stochastic gene expression..... 67
5. Relative importance of explanatory factors on mean gene expression and  $F^*$  ..... 73

## **AUTHOR CONTRIBUTIONS**

**Thesis title:** Darwin throws dice: modelling stochastic processes of molecular evolution.

**Chapter 1:** J.Y.D and G.V.B conceived and designed the study, analysed the datasets and wrote the manuscript. G.V.B developed the software. G.V.B and N.P performed the simulation study.

**Chapter 2:** J.Y.D and G.V.B conceived and designed the study, performed the simulation study and wrote the manuscript.

**Chapter 3:** G.V.B. and J.Y.D. designed the experiments and wrote the manuscript. G.V.B., N.P., and J.Y.D. conducted the analyses.

Plön, February 2nd, 2019.

---

Gustavo Valadares Barroso

---

Julien Yann Dutheil



## ACKNOWLEDGEMENTS

First and foremost, I thank my invaluable supervisor, Julien Yann Dutheil. First, for believing that an empiricist could enter the field of computational biology (and Coalescent Theory) with nothing but a strong motivation. “*Don't worry; whatever you do not know, you learn*”. Then, for keeping the door of his office open to my multiple intrusions – that often turned into hour-long discussions. Third, for his patience (I was not easy!), dedication and attention to mentor me not only about technical questions, but also on the more general and historical aspects of evolutionary biology. Lastly, for his good temper under any circumstances. I will spend the next years trying to understand – and emulate, if I can – how he was always smiling regardless of the problems we were facing. I am honoured to have been your student!

In this important moment of my life, I enjoy the exercise of reflecting back to where it all started. I thank my seventh grade biology teacher, “Pink”, for breaking the rules of the syllabus and lecturing about genetics instead of anatomy. Wherever you are, know that you had a critical impact in sending me down the path of understanding rather than memorising. And I thank my first mentor, Geraldo Moretto, who – among so many things – introduced me to population genetics and taught me the value of statistics. You are incredibly wise and I am happy to have you as a friend.

I am grateful to the members of my Thesis Advisory Committee – Asger Hobolth, Bernhard Haubold and Eva Stukenbrock – for pushing me forward to always do my best. Besides them, I learned a lot from my fellow PhD students and postdocs, specially Fabian Klötzl (always helpful when I needed guidance with C++), Alice Feurtey (who introduced me to broad sense Genomics) and Eric Hugoson (who taught me useful Bash tricks). I also had the pleasure to work with Nataša Puzović and Ana Filipa Moutinho, with whom I am looking forward to collaborating in the future. Discussing with Guy Reeves topics ranging from politics to science was always fun! Last but not least, I thank Diethard Tautz for creating a pleasant atmosphere in the department – and for countless Friday beers!

Special thanks go to my favourite girl, Vandana Revathi Venkateswaran, for sharing with me the most important moments of the past two years – both in and out of the Max Planck Institute. Your contagious smile, enthusiasm about life and care with others were bursts of motivation in the critical moments. I love you.

Moving across the Atlantic and leaving behind family and friends was a challenging moment in my life. I have kept you all close to my heart. Honourable mentions to Carlos Raphael Röhrig, Daniel Rial, David Richard da Luz and Rafael Hiroki Takano for visiting me in Europe and brightening my stay here with beer, laughter and old stories. Also, I am indebted to two friends I met at the institute: Bilal Haider (a.k.a Billy Boy Thompson) and Eric Hugoson (a.k.a Jeeves), thank you for providing an oasis of acid, politically incorrect and sarcastic humor, the importance of which to maintaining my sanity cannot be overstated. I am sure we will recapitulate those jokes – and create new ones! – whenever we meet. I got by with a little help from my friends.

Finally, coping with the pressure of an academic sense of achievement would not have been possible without the support of my family: Heloisa Barroso, Walter Wiele, Nayr Wiele, Mariana Wiele, Paula Wiele, Denise Wiele, Adelmo Coito and specially my mother, Désirée Wiele, who sacrificed a lot so that I could follow my dreams. Your unconditional love somehow gave me the freedom to always do things my way. I dedicate this thesis to my beloved grandfather, Walter Wiele, who taught me to enjoy the little things in life. He passed away on November 25, 2018, his atoms now stochastically traveling across the universe.

## **AFFIDAVIT**

I hereby declare that:

- i. Apart from my supervisor's guidance, the content and design of this thesis is the product of my own work. The co-authors contributions are listed in the dedicated section;
- ii. This thesis has not been already submitted either partially or wholly as part of a doctoral degree to another examination body, and no other materials are published or submitted for publication than indicated in the thesis;
- iii. The preparation of the thesis has been subjected to the Rules of Good Scientific Practice of the German Research Foundation.
- iv. Prior to this thesis, I have not attempted and failed to obtain a doctoral degree.


Plön, February 2nd, 2019.

---

Gustavo Valadares Barroso

# CURRICULUM VITAE

GUSTAVO VALADARES BARROSO

 gvbarroso@gmail.com



## EDUCATION AND PROFESSIONAL EXPERIENCE

---

- 2016 – 2019 **PhD, Evolutionary Genetics**  
Max Planck Institute for Evolutionary Biology / University of Kiel
- 2014 – 2015 High-school teacher
- 2012 – 2013 Professional Poker Player
- 2009 – 2011 **MSc, Genetics and Evolution**  
University of Sao Paulo
- 2005 – 2008 **BSc, Biology**  
Regional University of Blumenau

## SELECTED EVENTS DURING MY DOCTORAL STUDIES

---

- 2018 *An integrative model for population genomics inference* (**Invited Talk**)  
Host: Laurent Excoffier – University of Bern
- 2018 *The Neanderthal Recombination Map* (**Talk**)  
II Joint Congress on Evolutionary Biology – Montpellier
- 2018 *Inference of recombination maps from single pairs of genomes* (**Talk**)  
**Best Talk Prize** in the 2018 Annual Aquavit Symposium – Plön
- 2016 *Genome-wide investigation of eukaryotic transcriptional noise reveals the signature of selection at different levels of organization* (**Talk**)  
8<sup>th</sup> Theoretical Biology Workshop – Plön
- 2018 **Co-supervision of internship student Natasha Puzovic (MSc)**  
Max Planck Institute for Evolutionary Biology
- 2017 **Co-supervision of internship student Pallavi Misra (BSc)**  
Max Planck Institute for Evolutionary Biology
- 2016 **Co-supervision of internship student Natasha Puzovic (BSc)**  
Max Planck Institute for Evolutionary Biology