Genome-scale strategies controlling the impact of deleterious mutations

Iñigo Martincorena

Darwin College

A dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy

May 2012

To my family and Marta

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university.

This dissertation does not exceed the specified limit of 60,000 words as set by the Biology Degree Committee.

May, 2012

Iñigo Martincorena

Genome-scale strategies controlling the impact of deleterious mutations

Iñigo Martincorena

Every living organism faces the opposing evolutionary forces of adapting to changes while maintaining genome stability. On the one hand, mutations are necessary for adaptation. On the other however, spontaneous mutations are far more likely to be deleterious than adaptive. This poses a constant threat that has driven the evolution of a plethora of safeguarding mechanisms of genome integrity.

Though molecular biology research has led to substantial insights about the mechanisms of DNA damage and repair, our understanding keeps growing very actively and important questions remain unanswered. For example, very little is known about how mutational and repair processes act at different places along a genome. This thesis exploits genomic tools to investigate how selection for genome integrity has led to the evolution of strategies that, contrary to traditional repair, reduce the impact of deleterious mutations without compromising adaptability.

The first part of this thesis focuses on the fundamental question of how the point mutation rate varies along a genome. By combining phylogenetics and population genetics techniques to analyse 34 *Escherichia coli* genomes, I show that the neutral point mutation rate appears to vary by over an order of magnitude across 2,659 genes, with mutational hot and cold spots spanning several kilobases. Importantly this variation is not random, but instead appears to have been evolutionarily optimised to reduce the risk of deleterious mutations. An evolutionary risk-management model is introduced to explain how non-random mutation rates can evolve in a genome.

The second part of this thesis focuses on a mechanism by which integrity is maintained in the human transcriptome against the deleterious effects of intronic *Alu* elements. By examining genome-wide RNA-binding and RNA-sequencing data, the splicing regulator hnRNPC is shown to bind to hundreds of intronic *Alu* sequences, so repressing their accidental exonisation. This allows the deleterious disruption of hundreds of human transcripts to be avoided.

In summary, this thesis introduces two novel strategies that reduce the impact of deleterious spontaneous mutations to safeguard genome and transcriptome integrity. The work also exemplifies how the current boom in genome resequencing and functional genomics can increase our understanding of mechanisms controlling the impact of mutations. This will have important implications for our understanding of human diseases and evolution.

Acknowledgements

This dissertation is the culmination of a few years' journey during which I had the privilege of meeting very interesting people. It is now a pleasure to thank the many people who helped during the way.

My decision to pursue a PhD on bioinformatics after an eminently experimental degree was motivated by the beautiful discovery of theoretical biology and complexity theory. I am very grateful to the wonderful months with Joaquin and Jorge, in which we played with anything from social networks to neuroscience.

I am indebted to Nick Luscombe, my PhD supervisor, who patiently convinced me to stay in Cambridge for a PhD. I am most thankful for his friendship and his continuous support, as well as for giving me the freedom to explore my own interests. Without his broad interest in science this work would not have been possible.

I would like to thank all past and current members of the Luscombe group for their company and support during these past years and for innumerable hours chatting over instant coffee. I am particularly grateful to Aswin Seshasayee and Juanma Vaquerizas for their friendship and guidance.

I am also grateful to those whose advice and encouragement made this work possible. Special thanks to Albert Vilella and Kathi Zarnack for patiently proofreading parts of this thesis. Also thanks to Albert, Greg Jordan and Ari Loytynoja for early discussions on evolutionary theory, and to Julian König, Kathi Zarnack and Jernej Ule for inviting me to participate in their beautiful project. Moreover, I wish to thank the people that helped me during the revision of the *Nature* paper: Martin Ackermann, Adam Eyre-Walker, Nick Goldman, Boris Lenhard, John Marioni, Julian Parkhill, Olivier Tenaillon and Chris Tyler-Smith.

I am also indebted to the funding bodies that kindly supported me during the past four years: the Caja Madrid Foundation, the sadly extinct Spanish Ministry of Science and Innovation, the Marie Curie Fellowships Programme and the Cambridge University European Trust. I also want to thank Julian Parkhill, Gordon Dougan and Michael Quail for supporting my experiments and hosting me in their labs.

Most importantly, I thank my family for everything. I owe them all that I am and all that I have. I am immensely grateful to my parents Juan and Rosa; they raised me, taught me and loved me. They have unconditionally supported me throughout my life and are my permanent inspiration and ground. Finally, I am hugely grateful to Marta, for sharing this journey with me. Without her constant love and encouragement none of this would have been possible.

Table of contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix 1

Appendix 2

List of Tables

Chapter 1

Chapter 2

Chapter 4

List of Abbreviations

8-oxoG	8-oxo-7,8-dihydroguanine
A	Adenine
bp	Base pair
C	Cytosine
CAI	Codon adaptation index
cDNA	Complementary DNA
ChiP-Seq	Chromatin immunoprecipitation followed by sequencing
COG	Clusters of orthologous groups
CUB	Codon usage bias
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
EST	Expressed sequence tag
G	Guanine
HERV	Human endogenous retrovirus
HGT	Horizontal gene transfer
hnRNP	Heterogeneous nuclear ribonucleoprotein
Indels	Insertions and deletions
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
kb	kilobase
KD	Knockdown
LCR	Low-copy number repeats
LD	Linkage disequilibrium
LINE	Long interspersed elements

LOWESS	Locally weighted scatterplot smoothing
LTR	Long terminal repeat
Mb	Megabase
ML	Maximum-likelihood
MMEJ	Microhomology-mediated end joining
mRNA	Messenger RNA
NAHR	Non-allelic homologous recombination
NHEJ	Non-homologous end-joining
nt	Nucleotide
PCR	Polymerase chain reaction
PTB	Polypyrimidine tract-binding protein
RAD-Seq	Restriction site associated DNA sequencing
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
ROS	Reactive oxygen species
RT	Reverse transcription
S. enterica	*Salmonella enterica*
SNP	Single-nucleotide polymorphism
snRNP	Small nuclear ribonucleoprotein
siRNA	Small interference RNAs
T	Thymine
T-tract	Thymine tract
TCR	Transcription-coupled repair
TD	Tajima's D
TDnon	Tajima's D at non-synonymous sites
TDsyn	Tajima's D at synonymous sites
TLS	Translesion synthesis
U-tract	Uridine tract
U2AF	U2 auxiliary factor

> *Evolutionary plasticity can be purchased only at the ruthlessly dear price of continuously sacrificing some individuals to death from unfavourable mutations.*
>
> — Theodosius Dobzhansky, *Genetics and the Origin of Species* (1937)

Introduction

1.1. Genome integrity and adaptability

1.1.1. The dilemma

The continuity of life from one generation to the next rests on the stability of the information stored in the DNA. If this information was not faithfully reproduced, species could not exist (Lynch et al., 1993). However, survival in a permanently changing environment also requires change, and so organisms evolve under the opposing forces of zealously safeguarding their genetic information while allowing errors to permit adaptation.

This evolutionary dilemma between genomic stability and plasticity has had a profound impact on the evolution of life. In fact it is the evolutionary pressure for maintaining genomic stability that has led to the evolution of very accurate DNA replication machinery and a plethora of DNA-protection mechanisms and repair pathways for correcting damaged DNA.

Though research in molecular biology has led to substantial insights into the mechanisms of DNA damage and repair, our understanding keeps growing very actively and important questions remain unanswered. For example, very little is known about how mutational and repair processes act at different places along a genome or in different cell types within an individual. These are questions with fundamental implications for human genetic disorders and cancer.

This introduction starts by summarising the current knowledge of mutagenesis and repair. This is followed by a description of how selection acting on the balance between genetic stability and plasticity has driven the evolution of these safeguarding mechanisms. This provides the context for the work of this thesis, in which I introduce two novel mechanisms that increase the stability of the information stored in the genome, while maintaining or even fostering evolutionary plasticity.

1.1.2. DNA damage

The integrity of genomic DNA is constantly threatened by multiple endogenous and exogenous processes that are estimated to cause in each cell every day more than 20,000 damaging events (Lindahl and Wood, 1999, Lange et al., 2011, Loeb, 2011) and over 10,000 replication errors (Preston, 2005). Below I review some of the most important sources of DNA mutagenesis.

1.1.2.1. Spontaneous hydrolysis

Studies carried out 40 years ago using ^{14}C-labelled purine or pyrimidine bases or digestion by AP endonucleases, estimated that the DNA of a human cell loses 2,000-10,000 purine bases (adenine and guanine) by spontaneous hydrolysis every day (Lindahl, 1993). In contrast, cytosine and thymine (pyrimidines) are lost at just 5% of this rate. If unrepaired, these abasic sites can block transcription and replication, but they are usually readily repaired by ubiquitous AP endonucleases (Lindahl, 1993).

In addition to the intrinsic lability of the N-glycosyl bond, the nitrogenous bases are also susceptible to hydrolytic deamination (see Figure 1.1). Everyday, about 100 cytosines per human cell are estimated to deaminate spontaneously into uracil. This lesion is premutagenic[1] since uracil preferentially pairs with adenine, potentially leading to C-to-T transitions (Strachan and Read, 2004). In the model bacterium *Escherichia coli*, uracils are excised by the uracil-DNA glycosylase encoded by the *ung* gene, whose inactivation increases the spontaneous mutation rate by 10-fold. Less frequently, deamination occurs in adenines rendering hypoxanthine, a premutagenic lesion that can cause A-to-G transitions.

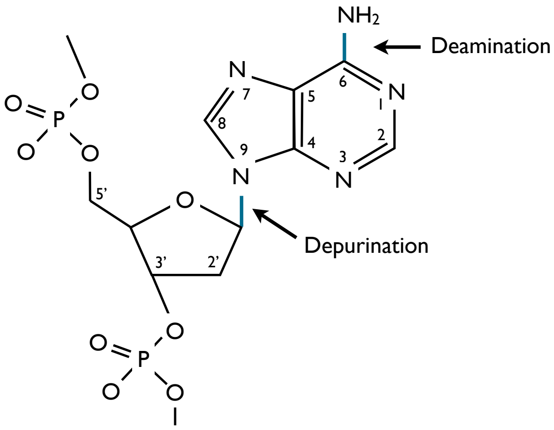

Figure 1.1*. Depurination and deamination of adenine by spontaneous hydrolysis.*

1.1.2.2. Reactive oxygen species

Reactive oxygen species (ROS) include superoxide radicals (O_2^-), hydrogen peroxide (H_2O_2) and hydroxyl radicals (·OH). ROS are an inevitable product of aerobic metabolism and can also be generated by environmental factors such as ionising radiation, UV irradiation and oxidative stress (Snyder and Champness, 2007).

ROS can cause different types of DNA damage. The major premutagenic form of oxidative DNA damage is 8-oxoG (8-oxo-7,8-dihydroguanine), which pairs preferentially with adenine and so can cause G-to-T transversions (Lindahl, 1993). This lesion is removed by a specific DNA glycosylase, present both in mammalian cells and *E. coli*. The inactivation of the *fpg* (*mutM*) gene that encodes this glycosylase in *E. coli* leads to a 10-fold increase in the global mutation rate.

[1] A *premutagenic* lesion refers to a DNA lesion that can lead to a mutation, typically by the erroneous incorporation of a base during replication.

It is believed that 8-oxoG and uracil formation are two of the main premutagenic lesions in the DNA (Lindahl, 1993). In addition to point mutations, ROS can also cause structural alterations in the DNA, such as deletions, insertions and rearrangements, by inducing breaks in the DNA backbone (Wiseman and Halliwell, 1996). Further, two nearby single-strand breaks in opposite strands can give rise to double-strand breaks. Given the importance of this threat, cells have evolved enzymes that remove ROS, including catalases, peroxide reductases and the superoxide dismutase, as well as specific repair enzymes such as exonucleases and glycosilases.

In addition to damage to the nuclear genome, ROS also threaten the integrity of the mitochondrial genome, particularly susceptible given the higher concentration of ROS in the mitochondria, the lack of protection by histones and a less efficient repair (Lindahl, 1993). As a consequence, rates of 8-oxoG formation have been estimated to be several times higher in mitochondrial DNA (Wiseman and Halliwell, 1996).

Oxidative damage to nuclear and mitochondrial DNA has been associated with cancer, ageing and age-associated degenerative disorders (Wiseman and Halliwell, 1996).

1.1.2.3. Endogenous methylation

1.1.2.3.1. Nonenzymatic DNA methylation

The best characterised and probably most important source of nonenzymatic methylation is *S*-adenosylmethionine (SAM). This molecule is an efficient methyl donor and is used as cofactor in most methylation reactions (Lindahl, 1993). In the absence of enzymatic catalysis, SAM is capable of transferring a methyl group to various nucleophiles at a slow rate, thus behaving as a weak endogenous alkylating agent (see below). Like other alkylating agents, SAM targets ring nitrogens of purine residues, mainly leading to the production of 7-methylguanine and 3-methyladenine. While 7-methylguanine is not premutagenic, 3-methyladenine is a cytotoxic lesion that causes DNA polymerase to stall, so blocking replication. A study estimated that about 600 3-methyladenine residues are endogenously generated in each human cell per day by this reaction (Rydberg and Lindahl, 1982).

1.1.2.3.2. Enzymatic DNA methylation

Enzymatic methylation of DNA is widespread and plays a very important role in the regulation of gene expression in mammalian cells. It typically occurs at the C5 position of cytosines yielding 5-methylcytosine, a reaction that virtually does not occur in the absence of enzymatic catalysis. 5-methylcytosine is premutagenic since it can spontaneously deaminate into thymine (C-to-T transition).

In fungi and animals, cytosine methylation mainly occurs at CpG dinucleotides and in vertebrates as many as ~70-80% of CpG cytosines are methylated. At CpG sites the cytosine transition rate is increased by ~30-fold in great apes (Hominidae) and ~15-fold in other mammals (Hodgkinson and Eyre-Walker, 2011). In addition, methylated CpG motifs also have a higher transversion rate of the guanine (G-to-T and G-to-C) (Ketterling et al., 1994).

1.1.2.4. Exogenous mutagens

1.1.2.4.1. Alkylating agents

Alkylating agents are molecules capable of transferring an alkyl group (*i.e.* an aliphatic hydrocarbon such as a methyl group) to DNA bases. Both the bases and the phosphates of the backbone can be alkylated, but some of the most reactive groups are the N7 of guanine and the N3 of adenine. Methylation or ethylation of these groups, for example by ethyl methanesulfonate (EMS, nitrogen mustard gas) or methyl methanesulfonate (MMS), cause major disruptions of the helix and can be premutagenic (Snyder and Champness, 2007).

Other important examples of alkylating agents are N-methyl-N′-nitro-N-nitrosoguanidine (MNNG) and methylnitrosourea, which attack other atoms of guanine and thymine producing *O*^{6}-methylguanine and *O*^{4}-methylthymine (Lindahl, 1993). Since these modifications do not cause significant distortions in the helix, they are not repaired by the main repair pathways, making them highly mutagenic (Snyder and Champness, 2007).

1.1.2.4.2. Ionising and non-ionising radiation

Ionising radiation (gamma rays and X-rays) can cause single- or double-strand breaks in the DNA as well as produce ROS.

Ultraviolet (UV) radiation from sun exposure is a major source of natural damage to DNA. The conjugated-ring structure of the bases strongly absorbs light in the UV wavelengths, making their double-bonds highly reactive. The most common lesions in the DNA by UV radiation are pyrimidine dimers, typically cys-syn cyclobutane thymine dimers (CPDs) and 6-4 photoproducts or Dewar photoproducts (which occur at 30% the frequency of cyclobutane dimers) (see Figure 1.2).

Pyrimidine dimers cause major distortions of the structure of the DNA as well as the local unpairing with the opposite strand. DNA or RNA-polymerases that encounter a pyrimidine dimer will stall at the site of the lesion, blocking replication or transcription. As described in Section 1.1.3.1, these lesions can be repaired by direct photoreactivation or by the nucleotide excision repair pathway. Alternatively, an error-prone polymerase can be used to bypass these lesions before repair, so allowing replication to resume. Since some error-prone polymerases bypassing TT dimers preferentially incorporate adenines, TT dimers are less mutagenic than TC dimers.

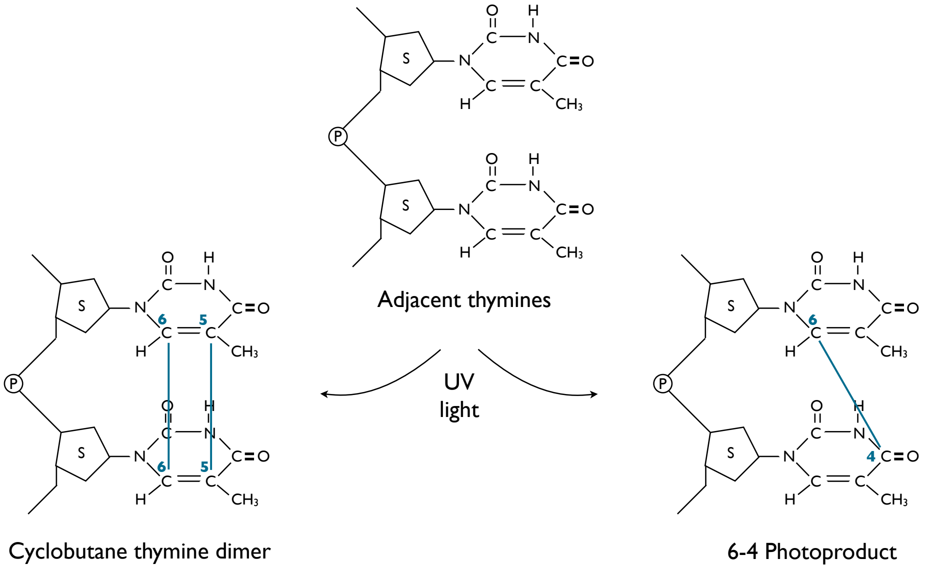

Figure 1.2*. Formation of pyrimidine dimers at two adjacent thymines.*

1.1.2.4.3. Crosslinking agents

Many chemicals such as UV-activated psolarens, mitomycin, cisplatin and ethyl methanesulfonate can cause interstrand covalent crosslinks in the DNA (Snyder and Champness, 2007). These DNA lesions block transcription and replication and can cause chromosomal breaks and rearrangements.

1.1.2.4.4. Intercalating agents

Intercalating agents such as ethidium bromide, 9-aminoacridine and proflavide, have planar structures that allow their intercalation between two consecutive bases on a DNA strand. This increases the distance between consecutive bases, preventing proper alignment with bases on the opposite strand and causing the two strands to slip with respect to each other. This typically leads to the insertion or deletion of one or a few bases during replication, particularly at short repeats, which are more prone to polymerase slippage (Snyder and Champness, 2007). These lesions are typically repaired by the mismatch repair pathway (see below).

1.1.2.4.5. Deaminating agents

In addition to deamination by spontaneous hydrolysis, some types of chemicals also induce the removal of amino groups from DNA bases. Typical deaminating agents with great mutagenic potential are bisulfite and nitrous acid. Bisulfite is specific for single-stranded cytosines, causing C-to-T transitions in unpaired DNA. Nitrous acid can deaminate cytosines, adenines and guanines, leading to C-to-T and A-to-G transitions, as well as deletions (Snyder and Champness, 2007).

1.1.2.5. Replication errors

Multiple repair pathways have evolved to remove as much DNA damage as possible before DNA replication. However, the fidelity of DNA polymerases is also limited in the absence of DNA damage. The error-rate during replication per base pair per division is on the order of ~$5x10^{-10}$ for *E. coli* and ~$5x10^{-11}$ for humans[2] (Drake, 1991, Drake et al., 1998).

[2] The estimation of $5x10^{-11}$ errors per base per replication in humans is based on assuming 400 divisions during spermatogenesis and an estimated mutation rate per base per generation of $2x10^{-8}$ (Drake et al., 1998).

In most organisms this overall accuracy of DNA replication results from the combined action of three, highly conserved, components acting sequentially: (1) use of the correct nucleotide by the DNA polymerase (error-rate ~10^{-5} in *E. coli*), (2) removal of erroneous bases by the 3′→5′ exonuclease proofreading activity of the replicative DNA polymerase (which reduces the error-rate by an additional factor of ~10^{-2} in *E. coli*), and (3) postreplicative repair by the DNA mismatch repair pathway (MMR, an additional ~10^{-3} reduction in *E. coli*). The combination of these three control steps leads to the overall ~10^{-10} rate observed (Fijalkowska et al., 2012).

In addition to mutagens directly damaging the DNA, molecules acting as base analogs can substantially raise the basal error-rate during replication. Base analogs can be generated endogenously by oxidative stress, inflammation and aberrant nucleotide synthesis, and due to their resemblance of normal DNA bases they can be incorporated in the newly synthesised DNA (Waisertreiger et al., 2010). This can be mutagenic as base analogs often pair with wrong bases, causing single-base mutations. For example 2-aminopurine (analog of adenine and guanine) and 5-bromouracile (analog of thymine) can cause C-to-T or T-to-C transitions (Snyder and Champness, 2007).

1.1.2.5.1. Error-prone polymerases

Most organisms, ranging from bacteria to mammals, have multiple DNA polymerases, each characterised by different error-rates and involved in distinct biological functions. For example, *E. coli* has five different DNA polymerases (see Table 1.1), the yeast *Saccharomyces cerevisiae* has eight and human cells at least sixteen (Fijalkowska et al., 2012).

DNA polymerases participating in the replication of DNA during S phase of the cell cycle are typically highly accurate and posses 3′→5′ exonuclease proofreading activity. However, most organisms also have error-prone polymerases that lack proofreading activities and incur in much higher error-rates. Error-prone polymerases are typically used to bypass damaged bases, such as pyrimidine dimers, where other polymerases are unable to proceed. As some damaged sites do not offer any base complementarity (*e.g.* pyrimidine dimers), these polymerases bypass a lesion by introducing unspecific bases (often an adenine), so rescuing the process of replication at the cost of more frequent mismatches.

	Pol I	***Pol II***	***Pol III***	***Pol IV***	***Pol V***
Family	A	B	C	γ	γ
Activities	5'-3' polymerase 3'-5' exonucl. 5'-3' exonucl.	5'-3' polymerase 3'-5' exonucl.	5'-3' polymerase 3'-5' exonucl.	5'-3' polymerase	5'-3' polymerase
#/cell					
- SOS	400	50-75	10-20	150-250	<15
+ SOS	400	350-1000	10-20	1200-2500	200
Biological functions	DNA replication, Okazaki fragment maturation, DNA repair	DNA replication (backup DNA polymerase), DNA repair, TLS	DNA replication (main replicative polymerase), DNA repair	TLS	TLS

Table 1.1. *Relevant properties of the five DNA polymerases in E. coli. Adapted from (Fijalkowska et al., 2012). The table shows the activities of each of the DNA polymerases, the functions in which they are involved and the number of polymerases per cell from each type, both in normal conditions (-SOS) and under induction of the SOS-response (see Section 1.1.3.2). TLS = Translesion synthesis.*

In addition to translesion synthesis, DNA polymerase exchanges can occur in the replication complex, and error-prone DNA polymerases have been shown to enter the replication complex also during the synthesis of undamaged DNA (Fijalkowska et al., 2012). The regulation of this process and its impact on genome stability remains largely unknown. Interestingly, the use of error-prone polymerases has been reported to play a role in somatic hypermutation in the vertebrate immune system (Seki et al., 2005) (see below).

1.1.2.6. Insertional mutagenesis by viruses and transposons

Certain viruses, such as retroviruses and bornaviruses in mammals and some bacteriophages, are capable of integrating in the genome of the host cell during their life cycle. If such integration events occur in the germline of multicellular organisms, viral sequences can effectively enter the genome of a species. In fact, it is estimated that ~8% of the human genome derives from endogenised retroviral sequences (Lander et al.,

2001) and sequences from non-retroviral viruses have also been detected (Horie et al., 2010).

A more frequent source of insertions are transposons. There are typically two major types of transposons: (1) Class I or retrotransposons, which are sequence elements that are transcribed into RNA, reverse transcribed into a DNA fragment and finally integrated into the genome (*i.e.* they follow a "copy and paste" mechanism), and (2) Class II of DNA transposons, which are DNA elements capable of *jumping* to different positions within a genome using a transposase (*i.e.* "cut and paste").

Given their importance their relevance in human genome evolution and their importance in the context of this thesis (Chapter 5), the section below describes in more detail the biology and main classes of human retrotransposons.

1.1.2.6.1. <u>Human retrotransposons</u>

Approximately 45% of the human genome derives from transposable elements: (a) DNA transposons (~3% of the human genome) and (b) retrotransposons (~42%). Most of these elements are currently inactive, but a small fraction of retrotransposons are still mobile (Cordaux and Batzer, 2009).

Retrotransposons can be divided into two subgroups:

1. **Long terminal repeat (LTR) retrotransposons**. They are endogenous retroviruses (Human endogenous retroviruses, HERVs). They account for ~8% of the human genome and are almost completely inactive now.

2. **Non-LTR retrotransposons**. They mostly comprise LINE-1, *Alu* and SVA elements (together accounting for ~33% of the human genome). Only a small number of copies of these elements have been shown to be active, but more than 60 *de novo* insertions have been reported to cause genetic disorders in humans (Cordaux and Batzer, 2009).

    - **LINE-1 elements**. There are more than 500,000 copies of LINE-1 elements, constituting ~17% of the human genome. They have propagated throughout the human genome during the past ~150 million years. The canonical full-length sequence is ~6 kb long and contains a 5′ UTR with an RNA polymerase II promoter, two open reading frames (ORF1 and ORF2) and a

3′ UTR with a polyadenylation signal. ORF2 encodes an enzyme with endonuclease and reverse transcriptase activities, required for their propagation (Cordaux and Batzer, 2009). Most copies of LINE-1 are incomplete, with long 5′ truncations. Less than 100 copies are believed to be active nowadays (Brouha et al., 2003).

- ***Alu* elements**. There are more than 1 million *Alu* elements in the human genome, resulting from their mobilisation over the last ~65 million years. The canonical full-length *Alu* element is ~300 bp long and is formed by two arms (resulting from the fusion of two monomers derived from the 7SL RNA gene). The arms are separated by an A-rich region and the *Alu* element ends with a polyA tail of variable length. *Alu* elements do not encode for any protein and so they use the retrotransposition machinery encoded by LINE-1 elements. *Alu* elements are transcribed by RNA polymerase III and lack a termination signal, leading to transcription of the downstream flanking sequence.

- **SVA elements**. There are ~3,000 copies of SVA elements in the human genome. Their canonical sequence is ~2 kb long, and is composed of a hexamer repeat region, an *Alu*-like region, several tandem repeats, a HERV-like region and a polyA tail of variable length. They have propagated during the last 25 million years, and like *Alu* elements, they are non-autonomous and can only propagate using the LINE-1 machinery.

Retrotransposons are normally strongly repressed in germ cells by (a) methylation, (b) cosuppression by small interference RNAs (siRNAs) and (c) chromatin compaction, minimising their accidental expression (Kazazian, 2004). The rate of *Alu* insertions has been estimated to be once every 20 generations, and the rate of LINE-1 insertions is once every 20-200 generations (Cordaux and Batzer, 2009).

Interestingly, however, some studies suggest that these elements may be much more active in certain cell types, such as neural progenitor cells (Coufal et al., 2009, Baillie et al., 2011). Further, LINE-1 retrotransposons are often hypomethylated in cancer (Belancio et al., 2010), where they could cause somatic mutations either by direct insertional mutagenesis or by the creation and unspecific repair of double-strand breaks by the

LINE-1 endonuclease (Cordaux and Batzer, 2009). Unraveling the frequency and importance of somatic retrotransposition and LINE-1 ORF2 activity in different tissues will be an active area of research over the next few years.

1.1.2.7. Instability of repeated sequences

The highly repetitive nature of a large fraction of the human genome causes substantial genomic instability far beyond insertional mutagenesis by active retrotransposons. Two major mechanisms by which repeats can mediate mutations in a genome are polymerase slippage at low complexity tracts and non-allelic homologous recombination (NHAR).

1.1.2.7.1. Polymerase slippage at low complexity tracts

Sequences composed of the tandem repetition of a single nucleotide (homopolymeric tracts) or a short motif of 2-6 nucleotides (short tandem repeats or microsatellites), are highly unstable. During DNA replication, these sequences can often slip backwards or forwards one or more motifs, thus introducing insertions or deletions in the newly synthesised strand (see Figure 1.3). Consequently, these sequences suffer a high rate of short indels, a property with important evolutionary and clinical consequences.

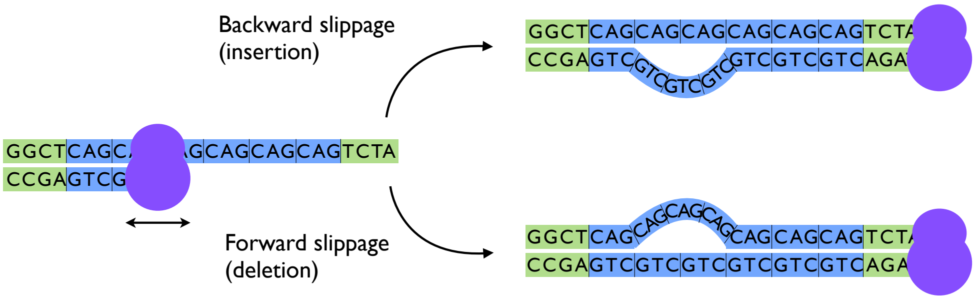
Backward slippage
(insertion)
GGCTCAGCAGCAGCAGCAGCAGTCTA
CCGAGTCGTCGTCGTCGTCGTCGTCAGA
GGCTCAGCA
AGCAGCAGCAGTCTA
CCGAGTCG
Forward slippage
(deletion)
GGCTCAGCAGCAGCAGCAGCAGCAGTCTA
CCGAGTCGTCGTCGTCGTCGTCAGA

Figure 1.3. Polymerase slippage at short repeats lead to small indels. These repeats have important evolutionary implications in bacteria (bacterial contingency loci) and important clinical implications in humans (trinucleotide repeat disorders).

Some bacterial species, such as *Campylobacter jejuni*, have taken advantage of these unstable sequences by evolving homopolymeric tracts at the start of certain genes involved in the biosynthesis of surface structures (Parkhill et al., 2000). The high rate of short indels often causes frameshifts leading to the (easily reversible) inactivation of some genes, in turn altering the antigenic nature of the bacterial surface and helping evade the immune system. Other bacterial species have been reported to use similar strategies (often called bacterial contingency loci) (Moxon et al., 2006).

Polymerase slippage at microsatellites, particularly at trinucleotide repeats within the coding sequences of specific genes, is known to cause a series of genetic disorders in humans (often known as trinucleotide repeat disorders). A particularly important repeat is the $(CAG)_n$ repeat, which encodes for a tract of polyglutamines of variable length in different individuals. Expansion of these repeats has been linked with at least nine different hereditary genetic disorders (Riley and Orr, 2006), including Huntington's disease (MacDonald et al., 1993).

1.1.2.7.2. Non-allelic homologous recombination (NAHR) mediated by repeats

NAHR refers to the pairing and recombination[3] between two sequences with very high similarity located in different (non-homologous) regions of a genome. Non-allelic crossover[3] recombination between two nearby repeats (typically no farther than 10 Mb) in the same orientation ("direct repeats") leads to duplication of the sequence in between the repeats in one chromosome and its reciprocal deletion in the other chromosome (Figure 1.4). Alternatively, NAHR can occur between nearby repeats in opposing orientations ("inverted repeats") causing inversion of the entire sequence in between the repeats (Strachan and Read, 2004).

NAHR in the human genome typically occurs between low-copy number repeats (LCRs), which are segmental duplications longer than 10 kb and over ~97% similar that mostly arose during primate evolution. LCRs are hot-spots for NAHR and are respon-

[3] *Homologous recombination* is a process by which two homologous DNA sequences pair and exchange information. In eukaryotes it typically occurs between sister chromatids or homologous chromosomes during mitosis or meiosis, following the formation of double-strand breaks. Homologous recombination in eukaryotes can lead to a *chromosomal crossover* by which the two homologous DNA sequences are broken and then reconnect to the other chromosome. Alternatively, homologous recombination in eukaryotes can lead to *gene conversion*, which involves the unidirectional transfer of a segment of DNA sequence from a donor sequence to a highly homologous acceptor sequence. Mechanistically gene conversion occurs when a single-stranded end of a broken DNA molecule (generated during the repair of a double-strand break) invades a homologous sequence and uses it as a template for repair (Chen et al., 2007b).

sible for most recurrent rearrangements in a genome (Stankiewicz and Lupski, 2010). However, NAHR can also occur between shorter interspersed repeats, such as the ubiquitous LINE-1 and *Alu* elements, which in turn favours the formation of novel LCRs (Kim et al., 2008). This process of repeat-mediated NAHR has important evolutionary and clinical implications in humans (Chen et al., 2010).

NAHR frequently occurs at gene clusters with a high level of paralogy. In addition to causing deletions, duplications and inversions of entire genes by non-allelic crossovers, non-allelic gene conversion[3] can also lead to the inactivation of genes by sequence exchange from highly similar pseudogenes (Strachan and Read, 2004).

Although the mechanism of non-allelic recombination is different in bacteria, a similar recombination process between interspersed repeats has been suggested to be responsible for large structural changes in bacterial genomes (Achaz et al., 2003, Ling and Cordaux, 2010).

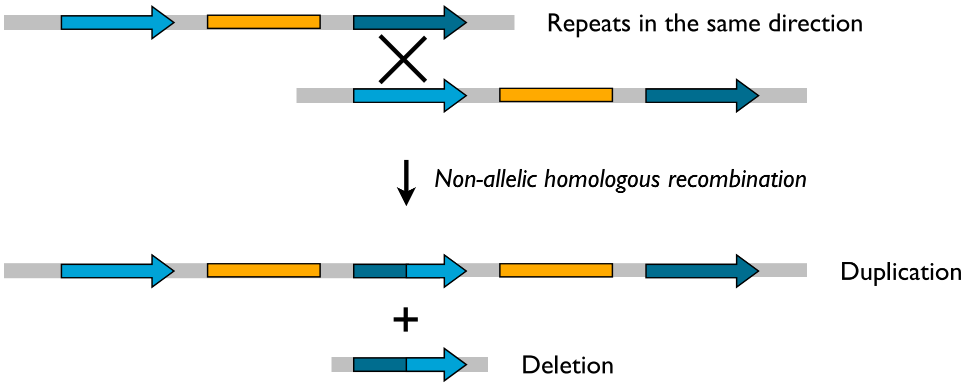
Repeats in the same direction
Non-allelic homologous recombination
Duplication
+
Deletion

Figure 1.4. *NAHR between proximal direct repeats leads to the duplication of the entire sequence contained between the repeats in one chromosome and its reciprocal deletion in the other chromosome. Figure based on (Kooy, 2010).*

1.1.2.8. Larger copy number changes

Some of the mutational processes described above can induce double-strand breaks and large-scale structural changes, leading to chromosomal translocations and large deletions, duplications or inversions. However, the increasing understanding of the landscape of somatic alterations during cancer, thanks to whole-genome resequencing techniques, is revealing the frequent occurrence of much larger chromosomal changes that require different mechanisms, such as chromothripsis (Stephens et al., 2011) and whole-chromosome aneuploidy (Gordon et al., 2012).

1.1.2.8.1. Chromothripsis

Chromothripsis is a phenomenon by which specific regions of the genome are shattered in multiple fragments in a single catastrophic event and reassembled in a disordered fashion, causing tens or hundreds of localised chromosomal rearrangements. Chromothripsis has been reported in at least 2-3% of tumours, although its frequency is much higher in certain cancer types, such as bone tumours (Stephens et al., 2011).

The mechanism for the chromosomal shattering remains largely unknown. Since the damage is often confined to a a single chromosome arm or a few loci in a reduced number of chromosomes, it has been suggested that chromothripsis may occur during mitosis, when chromosomes are condensed. In such conditions, exposure to ionising radiation might be capable of shattering a very restricted area of the genome (Maher and Wilson, 2012). Alternatively, other authors have suggested that mitotic segregation errors can lead to the formation of micronuclei, whose aberrant replication can induce their shattering. As micronuclei are naturally restricted to a region of the genome, the subsequent stitching of genomic fragments may also provide an explanation for the rearrangement patterns observed in chromothripsis (Crasta et al., 2012). In addition, survival of a chromothripsis event may require failures of checkpoints for chromosomal integrity, for example by the inactivation of p53 (Rausch et al., 2012).

1.1.2.8.2. Whole-arm and whole-chromosome aneuploidy

It has been shown that typically around 25% of the genome of a tumour is affected by whole-arm or even whole-chromosome somatic copy number alterations (Beroukhim et al., 2010).

The exact mechanisms that give rise to whole-arm or whole-chromosome aneuploidy are being actively investigated, but chromosomal segregation errors during mitosis are believed to be involved. Some *in vitro* estimates suggest that chromosomal mis-segregation may occur once every 100 divisions (Gordon et al., 2012), and the disruption of certain pathways may increase this basal rate. Segregation errors can be due to processes such as merotelic attachments, where a single kinetochore attaches to microtubules arising from both poles of the mitotic spindle, and often require the existence of defects in the spindle assembly checkpoint.

Interestingly, missegregation of chromosomes not only causes chromosomic aneuploidy, but has also been shown to increase DNA damage (Janssen et al., 2011).

1.1.3. Safeguarding mechanisms for genome integrity

The previous section described how genomic integrity is constantly jeopardised by a plethora of mutagenic processes affecting tens or hundreds of thousands of bases in each cell each day. This poses a major threat to the survival of living organisms and it has led to the evolution of a multitude of DNA repair pathways and other protection mechanisms.

This section summarises some of the main mechanisms safeguarding genomic integrity. As in the previous section on mutagenesis, here I summarise the different repair processes organised based on their mechanism. Although not all DNA repair systems are shared by all organisms, analogous forms of most of the systems can be found in organisms ranging from bacteria to mammals.

1.1.3.1. DNA repair

As described in the previous section, endogenous and exogenous factors can cause DNA damage in the form of abasic sites, altered bases (*e.g.* deamination and alkylation), crosslinks and double-strand breaks, among others. Since DNA replication can be blocked by these lesions (by the stalling of the DNA polymerase) or can introduce erroneous nucleotides, many DNA repair pathways have evolved to remove these lesions before replication. Based on their mechanism, repair pathways can be classified in three major types: (1) pre-replicative repair by direct reversal of the damage (“direct repair”), (2) pre-replicative repair by complete removal of the affected nucleotide (“excision repair”), and (3) post-replicative repair (mismatch repair).

1.1.3.1.1. Direct reversal

In humans, three enzymes have been described so far that directly correct DNA damage. The best characterised enzyme is the O^6-methylguanine-DNA methyltransferase, capable of removing the methyl group of methylated guanines (Strachan and

Read, 2004). Similar methyltransferases are capable of removing alkyl groups from damaged bases in bacteria. The main methyltransferases in *E. coli* are Ada and Ogt that repair O^6-methylguanine and O^4-methylthymine respectively. Interestingly, these methyltransferases are not real enzymes as they become inactivated upon acquiring the methyl group, and so are only capable of performing a single demethylation (Snyder and Champness, 2007).

A particularly interesting form of direct repair, common to most organisms but no longer active in placental mammals, is the repair of UV-induced cyclobutane pyrimidine dimers by the enzyme photolyase, in a process called photoreactivation (Snyder and Champness, 2007). Photolyase binds to a pyrimidine dimer in the DNA and, through a $FADH_2$ group, absorbs visible light between 350 and 500 nm obtaining the necessary energy to break the cyclobutane ring.

1.1.3.1.2. Single-strand damage

A much more common form of repair involves the excision of the affected base followed by the re-synthesis of the missing nucleotides (excision repair).

1.1.3.1.2.1. Base excision repair (BER) pathway

BER is primarily responsible for the repair of small DNA lesions that generally do not distort the helical structure of the DNA, but can cause mutations by pairing with wrong bases during replication (*e.g.* oxidised, alkylate and deaminated bases). BER is initiated by DNA glycosylases that recognise a specific modified base and hydrolyse the N-glycosil bond leaving an abasic site (apurinic/apyrimidinic). Abasic sites are then cleaved by an AP-endonuclease leaving a one-nucleotide gap and a hanging sugar (which is removed by an exonuclease). The resulting gap is then filled by a DNA polymerase, which either synthesises the single missing nucleotide (*short-patch BER*) or ~2-10 nucleotides (*long-patch BER*). Finally, a DNA ligase creates the missing phosphodiester bond (Liu et al., 2007).

Examples of DNA glycosylases are (Ide and Kotera, 2004):

1. AlkA (*E. coli*) and MPG (human): removal of 3-methyladenine and hypoxathine.

2. UDG (*E. coli*) and UNG (human): removal of uracil.

3. Fpg (*E. coli*) and hOGG1 (human): removal of 8-oxoG and FapyG.

Some glycosylases, such as Fpg, are bifunctional and have AP-endonuclease activity.

1.1.3.1.2.2. Nucleotide excision repair (NER) pathway

NER is one of the most important repair systems in most organisms as it repairs a wide diversity of DNA damage that causes bulky distortions in the helix, including most types of alkylations, pyrimidine dimers and base-sugar crosslinks. NER is also capable of repairing crosslinks between two strands, by acting in combination with recombinational repair (Snyder and Champness, 2007).

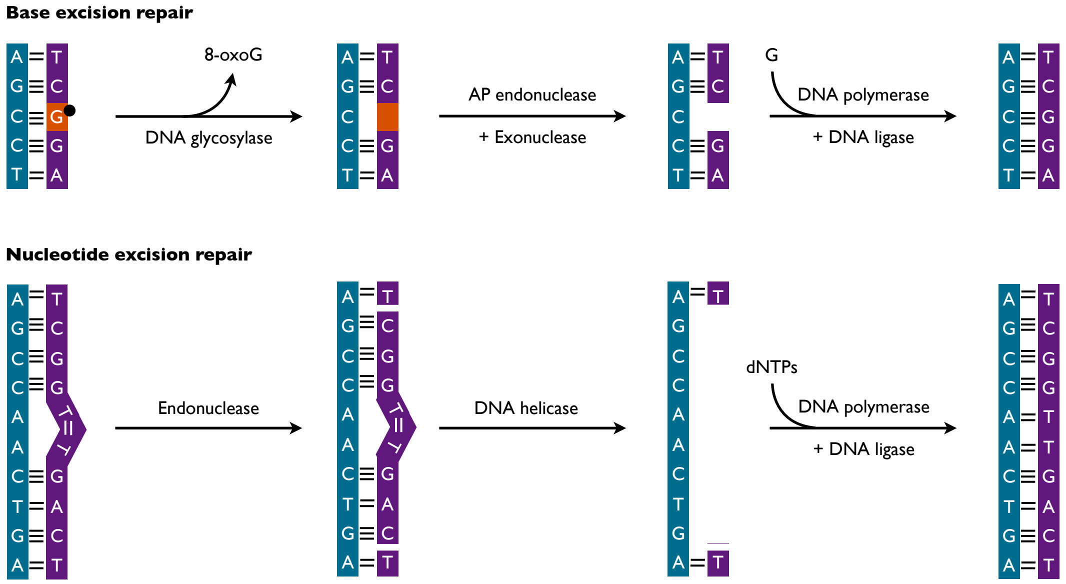

Figure 1.5. Schematic representation of the BER and NER pathways. *The sugar-phosphate backbone is represented as a coloured box. Hydrogen bonds are depicted as horizontal lines between the bases. Figure based on Strachan and Read, 2004.*

NER characteristically removes short pieces of DNA surrounding the damaged bases, and then re-synthesises the missing nucleotides. In *E. coli*, this is performed the UvrABC complex, conserved across all eubacterial species and some Archaea. A complex composed of two UvrA proteins and one UvrB is initially bound unspecifically to the genome scanning the DNA until it finds a distortion. UvrB then binds the lesion, UvrA is released and UvrC is recruited. UvrB cuts ~4 nucleotides downstream of the damage and UvrC ~7 nucleotides upstream of the damage. Finally, the UvrD helicase removes the nucleotides in between (including the damaged nucleotides) and DNA polymerase I re-synthesises the removed strand (Snyder and Champness, 2007).

The NER pathway is more complicated in mammals, involving nine proteins; however, the general mechanism is similar. Inactivation of some NER proteins is the cause of two hereditary genetic disorders in humans: Xeroderma pigmentosum and Cockayne syndrome (Subba Rao, 2007).

1.1.3.1.2.3. Mismatch repair

DNA mismatch repair is a post-replicative repair pathway that recognises and repairs erroneous insertion, deletion and misincorporation of bases arising during replication or recombination. It recognises mismatches in recently synthesised DNA, and specifically corrects the base in the newly synthesised strand.

The exact mechanism by which the newly synthesised strand is recognised is not known for many organisms. In *E. coli*, endogenous DNA is methylated by the Dam methylase at the adenine of the sequence 5′-GATC-3′. As the newly synthesised strand is temporarily unmethylated, the mismatch repair distinguishes the parental from the new strand and so repairs the latter. This pathway is known as methyl-directed mismatch repair.

Methyl-directed mismatch repair starts by the recognition of the mismatch by a MutS dimer. This recruits two units of MutL and one of MutH. MutH then cuts at the nearest unmethylated GATC and exonucleases degrade the newly synthesised strand including the site of the mismatch. Finally, DNA polymerase III re-synthesises the missing strand and a DNA ligase closes the gap (Snyder and Champness, 2007).

Mismatch repair in humans involves at least five MutS analogs and four MutL homologs (Kang et al., 2005). Mutations in genes involved in mismatch repair predispose the carrier to several types of cancers, including colon, ovary, uterus and kidney cancer. For example, hereditary nonpolyposis colon cancer (HNPCC) results from mutations in the human *mutS homolog 2 gene* (*hMSH2*).

1.1.3.1.3. Double-strand breaks

Double-strand breaks are particularly dangerous as they can lead to chromosomal rearrangements. In mammals they are repaired by three major mechanisms: non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ) and homologous recombination (Watson et al., 2004).

In NHEJ, the human DNA ligase IV directly joins two DNA ends (Wilson et al., 1997), independently of their terminal sequences, a desperate measure that often causes rearrangements (Strachan and Read, 2004). If the double-strand break contains single-stranded ends, repair can exploit microhomologies of 5-25 base pairs to anneal compatible ends before joining (MMEJ). MMEJ is error-prone as it typically deletes overhanging oligonucleotides that do not anneal, and it often mediates the formation of rearrangements (McVey and Lee, 2008).

Homologous recombination repair requires the use of an identical sequence to be used as a template (a sister chromatid, available in G2 phase of the cells cycle, or a homologous chromosome). It typically involves the machinery used in chromosomal crossover during meiosis, although this pathway is not as well characterised as the excision repair pathways. Human genes involved in this pathway include NBS (mutated in Nijmegen syndrome), BLM (mutated in Bloom syndrome) and the breast cancer predisposition genes BRCA1 and BRCA2 (Strachan and Read, 2004).

1.1.3.1.4. Translesion synthesis

Translesion synthesis is a DNA damage tolerance mechanism that allows DNA replication to bypass DNA lesions such as pyrimidine dimers and abasic sites (Waters et al., 2009). As described in Section 1.1.2.5.1, it involves the use of alternative DNA polymerases. These polymerases can be error-prone or relatively error-free. For example, the human DNA polymerase η incorporates adenines opposite to thymine dimers, resulting in the accurate repair of the lesion (Cruet-Hennequart et al., 2010). In contrast, DNA polymerase ι bypasses 6-4 photoproducts and abasic sites, typically introducing wrong bases (McDonald et al., 2001).

1.1.3.1.5. Transcription-coupled repair (TCR)

TCR is a DNA repair mechanism associated with the RNA polymerase II complex, which detects transcription elongation complexes stalled at DNA lesions, releases them and recruits the NER pathway to repair the damage. As in the case of translesion synthesis, TCR is a DNA damage tolerance mechanism as its main function is to avoid the blocking of transcription, which could lead to cell death (Svejstrup, 2002).

TCR is highly conserved in prokaryotes and eukaryotes. The enzyme responsible for TCR in *E. coli* is Mfd (Deaconescu et al., 2006). In humans, it comprises a series of factors, starting with the recognition of the lesion by the xeroderma pigmentosum C XPC/HHR23B complex. Inactivation of TCR in humans by mutations in the proteins CSA or CSB are known to cause Cockayne syndrome (Fousteri and Mullenders, 2008).

1.1.3.1.6. Recombination repair of damaged replication forks

Recombination repair of replication forks is another type of DNA damage tolerance mechanism. It uses the recombination machinery to allow replication to bypass damaged bases, avoiding the stalling of the DNA polymerase (McGlynn and Lloyd, 2002, Snyder and Champness, 2007).

1.1.3.2. Other protection mechanism

In addition to DNA repair, organisms have evolved additional mechanisms to ensure the maintenance of genomic integrity. Some of these mechanisms are normally repressed and are induced upon recognition of extensive DNA damage in the cell. Examples of these mechanisms include:

1. **Enzymes capable of removing reactive oxygen species** (Section 1.1.2.2). These include superoxide dismutases, catalases and peroxidases (Snyder and Champness, 2007).

2. **Inducible protection programs**. A particularly well known example is the prokaryotic SOS response that is activated by the detection of single-stranded DNA, stalled replication forks or double-strand breaks. In *E. coli* the RecA protein binds to single-stranded DNA, in turn activating the autoprotease activity of LexA and causing its degradation. The loss of LexA, a transcriptional repressor, induces the expression of most DNA repair pathways, and later error-prone polymerases as a last resort (Schlacher et al., 2006).

3. **DNA damage checkpoints**. Several checkpoints along the cell cycle ensure the integrity of the DNA and the completeness of the DNA replication before proceeding to the next phase. Proteins involved in DNA damage checkpoints in humans include ATM, ATR, BRCA1, MDC1 and 53BP1, some of which are affected in hereditary genetic diseases and cancer.

4. **Cellular senescence and apoptosis in multicellular organisms**. In response to extensive DNA damage that cannot be repaired, cells in multicellular organisms can either irreversibly stop dividing (senescence) or self-destruct (apoptosis). Mutations inactivating p53 in humans impair apoptosis and have been linked to a wide variety of diseases including cancer, neurodegeneration, ischemia, cholestasis and atherosclerosis (Amaral et al., 2010).

1.1.4. Natural selection

Despite the repair mechanisms described in the previous section, a small fraction of mutations escape unrepaired. If they affect the DNA in the germline, they can be inherited by the next generation so entering the pool of genetic diversity of the population and fuelling the process of evolution.

Natural selection, or more generally selection, is the process by which a given biological trait becomes more or less common in a population as a function of the differential reproductive success of the individuals bearing it (Darwin, 1859). Natural selection operates on the phenotypes, the observable characteristics of an organism resulting from its genotype. It is by this indirect action on the genotype that a given genetic variant or *allele* becomes progressively more or less common in a population. In this context, at a population scale, natural selection can be considered the ultimate safeguarding mechanism of functional genomic integrity, capable of removing mutations with a deleterious effect on the organism.

1.1.4.1. Functional classification of mutations

Regarding their functional impact, mutations are often classified into a limited number of discrete classes based on their position with respect to genes. A typical functional classification for point mutations is:

1. **Coding sequence mutations**: mutations occurring in exons.

   a. <u>Synonymous mutations</u>: mutations that change one codon for another encoding the same amino acid, and so not affecting the protein sequence.

   b. Non-synonymous mutations: mutations that affect the amino acid sequence.

      i. Missense mutations: mutations that change one codon for another encoding a different amino acid, causing a change in the protein sequence.

      ii. Nonsense mutations: mutations that change an amino acid-coding codon for a stop codon, causing the early truncation of the protein.

2. **Intronic mutations**: mutations occurring in introns. Some intronic (as well as some exonic) mutations can affect the splicing process by creating or destroying splice sites or splicing regulatory signals (see Chapter 5).

3. **Intergenic mutations**: mutations occurring outside genes. Some can affect sequences with regulatory importance.

Additionally, the term *silent mutations* is often used interchangeably with "synonymous mutations" or alternatively used to refer to any point mutation that does not affect the protein sequences (*i.e.* including synonymous, intronic and intergenic mutations).

This and similar classifications are widely used in the evolutionary literature. A major reason for their popularity is that these subtypes of mutations typically have very different effects on the organism. For example, non-synonymous mutations are most often deleterious to the organism while synonymous mutations are often consider to evolve neutrally. These properties make the classification of coding mutations into synonymous and non-synonymous a simple and powerful tool to study the action of selection. For example, when comparing homologous gene sequences of two related species, the ratio between the substitution rate[4] at non-synonymous (dN or Kn) and synonymous sites (dS or Ks) reflects the direction of selection at non-synonymous sites. $dN/dS < 1$ reflects negative (also known as *purifying*) selection against mutations changing the protein sequence. In contrast, $dN/dS > 1$ reflects positive selection fa-

[4] *Substitution rate* refers to the rate of fixation of nucleotide changes in a species (or a population). Driven by selection and drift, alleles progressively increase or decrease their frequency in a population until reaching fixation or extinction. Since fixation and extinction of alleles typically occur relatively quickly in evolutionary terms, most genetic differences between the sequences of two genomes from different species are assumed to correspond to fixed differences common to all individuals of each species. Thus, substitution rates are generally estimated comparing homologous sequences from different species.

vouring changes in the protein sequence (Miyata et al., 1980, Nei and Gojobori, 1986). Similarly, the classification of coding mutations into synonymous and non-synonymous is also widely used in population genetics[5], and is the basis of popular tests of selection such as the McDonald-Kreitman test (McDonald and Kreitman, 1991).

While these classifications have important practical applications, they oversimplify the functional impact of mutations. For example, certain synonymous mutations can be deleterious by affecting the speed or accuracy of translation or by affecting splicing (Chamary et al., 2006, Drummond and Wilke, 2008). Thus, while these classifications allow for a simple approximation to the functional impact of a mutation, strictly speaking this is more accurately reflected by its fitness[6] effect.

1.1.4.2. The distribution of fitness effects

The functional impact of a genetic variant from the point of view of selection can be described by its fitness effect, often noted as 1+*s* (where *s* represents the relative increase or decrease in fitness with respect to the ancestral allele). This description applies to any mutation independently of whether it is a point mutation or a large rearrangement, and whether it occurs in a coding or a non-coding region. The fitness effect of an allele can be quantified as the relative increase or decrease in the probability that the allele contributes to the next generation. Thus, $1+s=0$ indicates that a mutation is lethal and will not contribute to the next generation, $-1>s<0$ means that a mutation is deleterious (*i.e.* less fit than the wildtype), $s=0$ means that a mutation is strictly neutral (*i.e.* it does not affect the probability of the carrier to survive and reproduce), and $s>0$ indicates that a mutation increases the fitness.

By the action of natural selection the frequency of an allele is multiplied by 1+*s* each generation. This way deleterious mutations will tend to be progressively removed from the population and advantageous mutations will tend to increase their frequency, potentially reaching fixation (*i.e.* an allelic frequency of 100% in the population). Thus, the

[5] *Population genetics* refers to the study of the dynamics of genetic differences between individuals of a population. A major focus of population genetics is the study of changes in the allele frequency distribution caused by natural selection, genetic drift, mutation, demographic changes and genetic flow between populations.

[6] The *fitness* of a genotype refers to the ability of its phenotype to both survive and reproduce, being equal to the average contribution of the genotype to the gene pool of the next generation.

fitness effect of a mutation provides a more formal and accurate definition of natural selection and of the functional impact of mutations.

A useful representation of the impact of newly occurring mutations in a given species is the distribution of fitness effects (DFE) (Eyre-Walker and Keightley, 2007), which is the frequency distribution of s values of new spontaneous mutations. The shape of this distribution varies in each organism, depending on several factors such as the population size, the gene density or the frequency of environmental changes. Similarly, different classes of mutations can have very different distributions of fitness effects, with non-synonymous mutations typically showing negative s values and synonymous mutations typically showing values close to 0.

1.1.4.2.1. Neutral mutations and genetic drift

As defined above, owing to natural selection, the frequency of an allele in a generation (f_t) approximates f_{t-1}*(1+s)/(1+f_{t-1}*s) (where f_{t-1} is the frequency of the allele in the previous generation). However, this is an idealisation that is only valid for independently segregating mutations in infinitely large populations. In real populations with finite numbers of individuals, f_t will tend to f_{t-1}*(1+s)/(1+f_{t-1}*s) with a certain amount of sampling variability in each generation. This variability will be larger the smaller the population. This means that if the fitness impact of a mutation is sufficiently small (s is close to 0) and if the population is sufficiently small, the change in the frequency of an allele in each generation can be more strongly dominated by random sampling than by selection. The effect of random sampling on the dynamics of allele frequencies is called genetic drift.

Due to the effect of genetic drift, the strength of selection on an allele is better represented by the product of s and the effective population size[7] (N_e): N_es. Using this definition, we can then predict that selection will dominate the dynamics of an allele if $|N_es|>>1$. On the contrary, if $|N_es|<<1$, then genetic drift will dominate over selection.

[7] *Effective population size* refers to the size of an idealised population of randomly mating individuals that will show the same population genetic properties as the population being studied. The effective population size is typically smaller than the *census population size* (the real number of individuals in the population), due to factors such as non-random mating, population substructure, sex-ratio bias or non mating individuals, among others.

This can be easily shown by stochastic simulations. To illustrate this we can simulated a population of size *N*, composed of two subpopulations of size *N*/2, a wildtype subpopulation with fitness=1 and a mutant subpopulation with fitness=1+*s*. In each generation, a new population of size *N* is created by randomly sampling individuals from the previous generation with a probability proportionate to their fitness (Wright-Fisher model), and we follow the evolution of the population until one of the subpopulations becomes extinct. In an infinitely large population the wildtype population always wins for s<0 and always loses for s>0. In contrast, in a finite population, genetic drift makes uncertain the fate of weakly selected alleles. This is shown in Figure 1.6, which represents the fraction of populations simulated in which the mutant subpopulation won as a function of the fitness of the mutation.

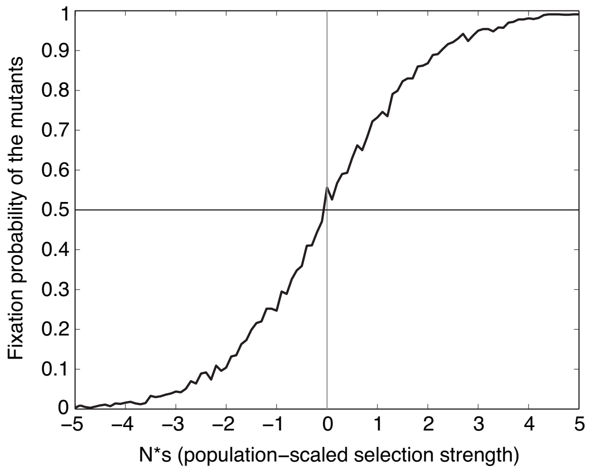

Figure 1.6. *Probability of the mutant subpopulation dominating over the wildtype population as a function of the population-scaled selection strength (Ns). 1,000 simulations were performed per data point using N=50. As predicted by theory, selection dominates the fate of an allele if* |*Ns*|*>>1 and random genetic drift dominates if* |*Ns*|*<<1. For example, if the mutant subpopulation has a weakly deleterious mutation with Ns=-1, it will still win over the wildtype in ~25% of the simulated populations.*

This formal description of selection and genetic drift has practical implications. First, it shows that there is a limit to the power of natural selection, as only mutations with |$N_e s$|>>1 will be effectively selected. Second, it provides an operational definition for neutrally evolving sites (Eyre-Walker and Keightley, 2007). Although synonymous mutations are often considered neutral in many methods in evolutionary genetics, strictly speaking it is unlikely that any mutation is purely neutral. All mutations must have a certain effect, even if this is very small. However, according to the description above, fixation of mutations with $|N_e s| << 1$ is dominated by random genetic drift, and so substitution rates at these sites can be considered to be effectively neutral.

1.2. Evolution of mutation rates

Section 1.1.3 described the large diversity of DNA repair and protection mechanisms. From an evolutionary point of view, the mere existence of these highly sophisticated mechanisms demonstrates that selection has acted on the mutation rates throughout evolution.

1.2.1. Existing models for the evolution of mutation rates

For decades, the question of whether mutation rates can be evolutionarily optimised by the action of selection has attracted considerable interest from theoreticians, starting with the seminal work by Kimura (Kimura, 1960, Kimura, 1967), Levins (Levins, 1967) and Leigh (Leigh, 1970, Leigh, 1973). Much of these early studies focused on the question of whether mutation rates in variable environments can be optimised to provide an optimal balance between genome stability and adaptability.

1.2.1.1. Optimal mutation rates

Leigh (1970) used a game theoretic approach to predict that, in an infinitely large asexual population in which the environment changes once every n generations, the optimal mutation rate is $1/n$. In a constant environment, adaptation is not needed and so the optimal mutation rate should be zero. In fact, it can be shown that in a simple model, a mutation rate of $1/n$ maximises the geometric mean of the population growth. So, this strategy can be considered to be a form of bet-hedging[8] (Nevoux et al., 2010) in

[8] *Bet-hedging* is an evolutionary strategy that maximises the geometric mean (instead of the arithmetic mean) of the fitness of a genotype in successive generations in a variable environment: $G=(x_1x_2...x_j)^{1/j}$. It is inspired in strategies used during gambling. Using a gambling analogy, maximisation of the arithmetic mean of returns will provide on average the highest return in a single bet. However, in a situation where the entire bankroll is reinvested in repeated bets, the final amount of money is a function of the product of the gains and losses in each bet, and so one should maximise the geometric mean of returns. As evolution resembles a situation of reinvesting all resources every generation, the optimal strategy against unknown environmental changes is maximising the geometric mean of the fitness along multiple generations.

which the $1/n$ frequency of mutants ensures an optimal balance between wildtype individuals (perfectly fitted for the current environment) and mutants (some of which will be ready to face an environmental change successfully).

Having derived the optimal mutation rate in a variable environment, most theoretical studies focused on whether selection can actually drive mutation rates to their optimal value. Different authors concluded that this could occur only in asexual (*i.e.* non-recombining) organisms. The reason for this is that selection on an allele affecting the mutation rate (a mutation-modifier allele) is indirect, occurring through selection on deleterious or adaptive mutations that occur elsewhere in the genome as a consequence of the change in mutation rate. For example, the loss of a repair enzyme leading to an increased mutation rate will be beneficial in very variable environments, in which adaptive mutations are constantly required. In such a case, a hypermutator allele may increase its frequency due to the association with adaptive mutations in other parts of the genome.

However, in sexual organisms with frequent crossover recombination, linkage[9] between distant parts of a genome is typically lost in an average of two generations (Lynch, 2011), and so the mutator allele does not benefit from an association with adaptive mutations. Instead, recombined genotypes incorporating both low mutation rates and adaptive mutations will be more successful as they do not suffer the mutational burden of hypermutators. Thus, genomic mutation rates are thought to be evolvable under changing environments in asexual organisms, but not in sexual organisms where the mutation rate would evolve to be as low as possible (Sniegowski et al., 2000).

However, the balance between positive and purifying selection is not the only relevant force driving mutation rates, and undoubtedly the cost of fidelity should be included in any realistic model. In fact, many authors consider that the main force behind

[9] *Genetic linkage* is the tendency of certain loci to be inherited together. In sexual organisms two mutations located in different chromosomes will only remain together for an average of two generations, as they have a 50% chance of coinciding in the same gamete. Two mutations located very close within the same chromosome will be inherited together unless a crossover recombination occurs in between them. *Linkage disequilibrium* refers to the increased probability of two mutations being inherited together due to genetic linkage. In sexual organisms linkage disequilibrium decays asymptotically to zero with increasing chromosomal distance (see Figure 3.7). In bacteria, where recombination occurs without crossover by copying a segment of a donor genome into a recipient genome (thus resembling gene conversion), a certain linkage disequilibrium exists among any pair of loci independent to their genomic distance.

the observed mutation rates in both sexual and asexual organisms is the balance between the advantage of reducing the frequency deleterious mutations and the metabolic cost of this reduction (Sniegowski et al., 2000).

In support of this idea, a more recent theoretical study (Andre and Godelle, 2006) demonstrated that using more realistic evolutionary models with a finite population of asexual organisms, recurrent positive selection (*i.e.* the need for adaptive mutations in a changing environment) cannot stabilise the mutation rate near its optimal value. Their model suggest that occasionally, strong positive selection can cause the invasion of the population by a hypermutator strain, but cannot fine-tune the evolution of mutation rates. This occasional invasion of populations by hypermutable strains has been observed experimentally in bacteria in response to the dramatic need for adaptive mutations (Bjedov et al., 2003); for example due to the presence of antibiotics (Blazquez, 2003) or during coevolution with phages (Pal et al., 2007). Similarly, genomic instability is often selected during the development of cancers[10] (Loeb, 2001, Hanahan and Weinberg, 2011). Nevertheless, the invasion of hypermutators does not appear to be a major force driving the evolution of genomic mutation rates in the long-term.

In summary, genomic mutation rates in asexual organisms (or genomes with high genetic linkage) are believed to evolve as a result of three forces: the loss of fitness by deleterious mutations, the metabolic cost of fidelity and probably only occasionally the need for adaptive mutations. In contrast, genomic mutation rates in sexual organisms are believed to be driven only by the balance between the cost of fidelity and the burden of deleterious mutations (Sniegowski et al., 2000).

1.2.1.2. The drift hypothesis: limits to the evolution of mutation rates

Recently, Michael Lynch proposed that genetic drift is an important factor affecting the evolution of mutation rates that has been neglected by previous theoretical studies (Lynch, 2010, Lynch, 2011). As the mutation rate evolves towards lower values, at a certain point the selective advantage of further reducing it will be dominated by drift ($|N_e s| << 1$). Thus, genetic drift imposes a lower bound to the evolution of the mutation rate.

[10] Note that cancer cells evolve asexually and so hypermutator alleles can increase their frequency in the population by hitchhiking with beneficial mutations.

Taking this factor into account, it then becomes a question of which lower bound -genetic drift or the cost of fidelity- is higher and so dominates the evolution of mutation rates in different organisms. Interestingly, if drift was the dominating lower bound, we should expect the genomic mutation rate to be lower in species with larger effective population sizes. Using a collection of estimated mutation rates in organisms ranging from prokaryotes to mammals, Lynch found an anti-correlation between the mutation rate and the effective population size (Lynch, 2010), providing some support for the drift hypothesis.

However, close inspection of the actual mutation rates from different organisms reveals values that appear to be several orders of magnitude higher than the limit imposed by drift. This is particularly obvious for bacteria. For asexual organisms, the selective advantage of a reduction in the genomic mutation rate can be approximated by the equation (Lynch, 2011):

$$s_d \approx \Delta U_h \,,$$

where s_d represents the fitness effect of a mutation-modifier allele that reduces the rate of deleterious mutations (U_h) by a factor (ΔU_h). In other words, the selective advantage of reducing the mutation rate by a given factor is equal to the reduction in the fraction of the offspring affected by a deleterious mutation. Since bacterial genomes are highly compact and both non-synonymous sites and intergenic sequences are strongly negatively selected, the fraction of spontaneous mutations that are significantly deleterious is very high. So the rate of deleterious mutations (U_h) should approach the rate of spontaneous mutations (μL, where μ represents the mutation rate per base per generation and L the length of the genome) and s_d would approach $\Delta\mu L$.

Using the equation above, we can predict that a mutation reducing the genomic mutation rate by 10% would provide a fitness advantage of $s_d \sim 0.1\mu L$. Since the rate of spontaneous mutations per genome per generation in most prokaryotes is approximately 0.003 *(Drake, 1991)*, the fitness advantage (s_d) of a mere 10% reduction of the genomic mutation rate in bacteria is higher than 10^{-4}. This is roughly four orders of magnitude higher than the limit imposed by drift to the evolution of mutation rates ($1/N_e$) for *E. coli (Hartl et al., 1994, Charlesworth, 2009).*

The situation in eukaryotes is slightly more complicated. Lynch (2011) shows that for a genome with free recombination, the fitness advantage of a reduction in the deleterious mutation rate can be approximated by the equation:

$$s_d \approx 2 s_{mean} \Delta U_h \ ,$$

where s_{mean} is the mean fitness cost of deleterious mutations in the genome. Thus, in eukaryotes, the situation is more complicated as we need good estimates of the rate of deleterious mutations and of the average fitness cost of mutations. Nevertheless, the limit predicted by drift in humans ($1/N_e \approx 10^{-4}$; Charlesworth, 2009) is so low in comparison to the genomic mutation rate ($\mu L \approx 100$; Eyre-Walker and Keightley, 2007) that very conservative estimates of s_d and U_h suffice to prove the point. Even if we simply consider that ~10% of non-synonymous mutations in humans have s>0.1 (Eyre-Walker et al., 2006), we can already see that a mere 10% reduction in the genomic mutation rate will provide a fitness advantage one or two orders of magnitude above the drift limit. And this analysis considers only ~0.1% of the human genome.

Thus, it appears that mutation rates observed in real organisms are several orders of magnitude above the drift limit, suggesting that the equilibrium between the cost of deleterious mutations and the cost of fidelity dominates the evolution of genomic mutation rates. Appendix 1 provides a more detailed consideration of the forces determining the genomic mutation rate in bacteria and discusses a higher bound to the evolution of mutation rates.

1.2.2. Variation of the mutation rate along a genome

Much less attention has been devoted to the evolution of local mutation rates within a genome (Cox, 1972, Moxon et al., 1994, McVean and Hurst, 1997, Altenberg, 2011). Nevertheless, to some extent the theory for global mutation rates also applies to local mutation rates. Since the strength of purifying selection and the frequency of positive selection vary for each gene, the optimal mutation rate and the lowest evolvable mutation rate in the drift hypothesis should also vary along the genome. Thus, assuming that the mutation rate can vary along the genome and that this variation is heritable, then the same forces that act on the evolution of global mutation rates should also act on the evolution of local mutation rates.

There are however at least three important differences in the evolution of local mutation rates in comparison with genomic mutation rates. First, local hypermutation near a mutation-modifier allele could also evolve in sexual organisms, owing to the strong linkage between the modifier allele and the derived advantageous mutations. Second, local hypermutation has the additional advantage of not imposing a high mutation burden in the entire genome, thus providing a stable solution to the need of regular genetic diversity at a given locus. Third, genetic drift is likely to be a major limiting factor of the evolution of local mutation rates given that the selective advantage of a modifier allele (s_d) is proportional to the size of the locus affected.

1.2.2.1. Known sources of variation of the mutation rate within a genome

Unfortunately, our understanding of how the mutation rate varies along a genome has been hindered by the lack of reliable approaches to measure local mutation rates on a large scale. Experimentally, absolute mutation rates can be determined using gene reporters in fluctuation tests, but these are unsuitable for measurements in native genes. Alternatively, in theory, relative mutation rates can be estimated from the accumulation of mutations at selectively neutral positions.

Owing to the relatively small effective population sizes in mammals, synonymous sites and certain non-coding sites are often considered to evolve largely neutrally. Taking advantage of this, multiple comparative genomics studies in mammals have used synonymous or intergenic substitution rates to explore the variation in the mutation rate within genomes. These studies revealed that the mutation rate varies substantially within mammalian genomes at different scales, with larger variations occurring at smaller scales (reviewed by Hodgkinson and Eyre-Walker, 2011):

1. **Single nucleotide variation**

    a. Heterogeneous substitution matrices. Each base has different mutation probabilities to each of the other bases. In most organisms, from bacteria to eukaryotes, G and C mutate more frequently than A and T (Hodgkinson and Eyre-Walker, 2011). Also, transitions[11] are typically more frequent than

[11] *Transition* refers to a purine-to-purine or pyrimidine-to-pyrimidine base change (A-G or T-C).

transversions[12] (Strachan and Read, 2004). The frequency of change between each pair on nucleotides is represented in a substitution matrix.

b. Context-dependent effects. Some of the best understood causes of mutation rate variation are due to the presence of certain nearby nucleotides. For example, CpG dinucleotides are hypermutable in mammals due to their frequent cytosine-methylation (Section 1.1.2.3.2). Also, in skin cancers somatic C-to-T transitions are more frequent at pyrimidine-dimers as a result of UV damage (Pleasance et al., 2010).

c. Context-independent effects. Some studies have found that single-nucleotide polymorphisms[13] (SNPs) in humans occur more frequently at sites that are also polymorphic in chimpanzees or macaques, even when accounting for the effect of neighbouring nucleotides (Hodgkinson et al., 2009). This observation does not seems to be a consequence of methodological artefacts or selection (Hodgkinson and Eyre-Walker, 2011).

2. **Small-scale variation**

a. CpG islands. These are ~1 kb CG-rich regions characterised by a high frequency of CpGs that consequently show increased mutation rates.

b. Association with transcription levels. Experimental studies have revealed that transcription is mutagenic (Beletskii and Bhagwat, 1996, Klapacz and Bhagwat, 2002, Ochman, 2003). Interestingly, however, comparative genomic studies in humans have revealed no clear association between the mutation rate and expression. Cancer genome sequencing studies have reported a lower frequency of somatic mutations at more highly expressed loci (Hodgkinson and Eyre-Walker, 2011), although the roles of selection and other biases remains unclear.

[12] *Transversion* refers to a purine-to-pyrimidine or pyrimidine-to-purine base changes.

[13] A *polymorphism* refers to a locus that shows genetic diversity (*i.e.* two or more alleles) within a population. A *single-nucleotide polymorphism (SNP)* refers to a nucleotide in a genome at which individuals of a population show two or more different bases.

3. **Large-scale variation**. According to an autocorrelation analysis of substitution rates between mouse and rat at ancestral repeats, the mutation rate in the murid genome varies in blocks of 0.1-1 Mb (Gaffney and Keightley, 2005).

4. **Variation between chromosomes**

    a. Sex chromosomes. The mutation rate is typically lower in the X chromosome and higher in the Y chromosome, than in autosomes (Hodgkinson and Eyre-Walker, 2011). This difference is largely due to the fact that males typically have higher germline mutation rates per generation than females (due to the higher number of divisions during spermatogenesis than during oogenesis) (Haldane, 1947, McVean and Hurst, 1997, Ellegren, 2007).

    b. Autosomes. Significant variations in the mutation rate have also been reported between autosomes, although this variation is much smaller than that observed in sex chromosomes (Hodgkinson and Eyre-Walker, 2011).

Importantly, the variation of the mutation rate observed in these studies is likely to be an underestimation, given that changes in local mutation rates during evolution will tend to reduce the differences between loci. Resequencing studies of human individuals and tumours are likely to provide an unprecedented picture of the variation of germline and somatic mutation rates, although methodological challenges such as regional biases in coverage and mappability will need to be carefully considered.

Unfortunately, very little is known about the variation of the mutation rate within bacterial genomes, since estimation of neutral mutation rates in bacteria is complicated by extensive gene transfer and substantial selection bias even at synonymous sites (for a review see Ochman, 2003).

1.2.2.2. Little evidence for the evolutionary optimisation of local mutation rates

Despite the availability of information about the variation of the mutation rate within mammalian genomes, there is virtually no evidence that any of this variation is evolutionary advantageous and that it has been driven by second-order selection on mutation rates. Intriguing correlations between local synonymous substitution rates and gene function or fitness cost have been reported (McVean and Hurst, 1997, Chuang and Li, 2004). However, since selection can affect synonymous substitution rates through its action on codon-usage, messenger-RNA (mRNA) folding stability and cis-regulatory elements (Chamary et al., 2006), interpretation of these observations in support of optimised mutation rates has remained contentious (Baer et al., 2007, Hodgkinson and Eyre-Walker, 2011).

Thus, to date there has been no convincing evidence that the local point mutation rate has been optimised by selection in any organism. Moreover, recently, Hodgkinson and Eyre-Walker (2011) predicted that based on our current understanding of the extent of variation of the mutation rate within mammalian genomes and their low effective population sizes, genetic drift will dominate and second-order selection is unlikely to optimise local mutation rates in mammals.

Nevertheless, currently there are at least two well-known examples of selective optimisation of local mutation rates:

1. **Bacterial contingency loci**. As described in Section 1.1.2.7.1, some bacterial genomes show hypervariable homopolymeric or short-tandem repeats in genes associated with immune evasion (Moxon et al., 1994, Moxon et al., 2006).

2. **Somatic recombination and somatic hypermutation** in the vertebrate immune system. Somatic recombination is a programmed targeted recombination that generates great diversity in the early stages of B and T cell receptor formation (Schatz and Swanson, 2011). In addition, once a circulating B cell recognises an antigen, it proliferates and undergoes somatic hypermutation to improve the affinity to the antigen. This process of programmed hypermutation produces a very high somatic mutation rate (~10^5-10^6 higher than usual) in the variable regions of immunoglobulin genes of B-cells. This involves the deamination of cytosine to uracil followed by their error-prone repair by a low-fidelity DNA polymerase (Odegard and Schatz, 2006).

Although these are unusual examples that account for a very small minority of mutations, they nevertheless provide a proof-of-principle, demonstrating that local mutation rates are evolvable.

1.3. Questions, strategies and outline of the thesis

This introduction has given an overview of our current understanding of mutagenesis and repair, and how evolution optimises mutation rates. In this context, the thesis describes the work and findings of two research projects.

1.3.1. Non-random variation of the mutation rate in the *E. coli* genome

As described in the previous sections, our understanding of the variation of the mutation rate within bacterial genomes is very limited due to the traditional lack of reliable neutral estimates in bacteria. This is unfortunate since bacteria might be more likely to evolve local mutation rates, given their large population sizes. Thus, the main goal of this thesis was to determine if bacterial genomes show evidence for the evolutionary optimisation of local mutation rates. The question is important for our understanding of evolution (as evolved local mutation rates would mean that mutations occur non-randomly with respect to fitness), and for our understanding of how mutagenesis and repair act within a genome.

Answering this question accounts for three out of the four research chapters of the thesis (Chapters 2-4). The work exploited phylogenetic and population genetic techniques to overcome traditional limitations, by taking advantage of the availability of the genome sequences of 34 strains of the model bacterium *E. coli*. Traditional studies comparing two genomes from different species suffered the fundamental limitation that the effects of selection and mutation rate on sequence divergence cannot be distinguished. However, as the two processes leave distinct patterns of polymorphisms, population genetic techniques could be applied to disentangle their relative contributions (Braverman et al., 1995, O'Fallon, 2010, Bustamante et al., 2003).

The description of this work is divided in three chapters:

1. **Chapter 2**. This chapter describes the methodology used to calculate synonymous diversity for each gene in *E. coli*. This includes a series of strict phylogenetic filters and detailed quality controls developed to avoid multiple potential artefacts, such as horizontal gene transfer, ortholog misidentification and alignment or sequencing errors. The chapter ends by describing the large variation of the synonymous diversity observed along the *E. coli* genome.

2. **Chapter 3**. Although the variation of the synonymous diversity is suggestive of the variation of the mutation rate, selection and non-selective biases could confound any interpretation of this variation. Thus, to avoid the ambiguity previously seen in the interpretation of these results, Chapter 3 carefully quantifies the impact of all major potential biases on the variation observed.

3. **Chapter 4**. This chapter evaluates whether there is any evidence of evolutionary optimisation of the local mutation rate in *E. coli*. Importantly, the observations suggest that functionally more important genes tend to experience lower neutral mutation rates. An evolutionary risk management model is then presented to explain how these non-random mutation rates can effectively evolve in bacteria. The chapter ends with a discussion and some preliminary evidence on the evolution of local mutation rates in humans.

This work, including the majority of the analyses described in these three chapters, was published in *Nature* on the 3rd May 2012:

> **Martincorena I***, Seshasayee ASN & Luscombe NM*. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature*, 485:95-8.
>
> (*) denotes corresponding author. (Martincorena et al., 2012)

1.3.2. A safeguarding mechanism against the accidental exonisation of intronic *Alu* elements in humans

In the context of my interest on control mechanisms for genome integrity, I also participated in a collaborative project between our lab at the EMBL-European Bioinformatic Institute and Jernej Ule's lab at the MRC-Laboratory of Molecular Biology (Cambridge). This project, led by Julian König on the experimental side and Kathi Zarnack on the computational side, describes a novel safeguarding mechanism that protects transcriptome integrity in humans against the accidental exonisation of intronic *Alu* elements.

This finding started with the observation that depletion of the RNA-binding protein hnRNP C leads to the destructive exonisation of thousands of intronic *Alu* elements, disrupting thousands of transcripts in humans. Combining RNA-sequencing and iCLIP (a new protocol to identify protein-RNA interactions), we were able to unravel the scale and the mechanism of the repression of intronic *Alu* elements by hnRNP C.

My exact role in this project is described in Chapter 5. It included performing an evolutionary analysis of over 800,000 *Alu* elements in the human genome and relating this with the RNA-sequencing and iCLIP data. This revealed a very detailed picture of this novel safeguarding mechanism:

1. The exonisation of intronic *Alu* elements is globally repressed by hnRNP C.
2. Accidental *Alu* exonisation uses a reduced set of cryptic splice sites and splicing signals present in the ancestral *Alu* sequence.
3. hnRNP C represses *Alu* exonisation by strongly binding at the ancestral U-tracts, hiding these splicing signals from the splicing machinery.
4. Finally, we reveal the existence of strong negative and positive selection during primate evolution to maintain and strengthen the repression by hnRNP C.

A manuscript describing this work has been submitted for peer-review publication:

> Zarnack K†, König J†, Tajnik M, **Martincorena I**, Stévant I, Reyes A, Anders S, Luscombe NM*, Ule J*. 2012. Direct competition between hnRNP C and U2AF65 controls the exonization of hundreds of *Alu* elements. *Submitted.*
>
> (†) denotes equal contribution.

1.3.3. Two novel mechanisms protecting genomic functional integrity while maintaining adaptability

In summary, this thesis introduces two novel strategies that reduce the impact of deleterious spontaneous mutations to safeguard genome and transcriptome functional integrity. Traditional repair pathways and DNA protection mechanisms are typically assumed to reduce the occurrence mutations indiscriminately, protecting the genome but in turn reducing adaptability. In contrast, the two mechanisms introduced here appear to protect genomic integrity while maintaining or even fostering adaptability.

This thesis also exemplifies how the current boom in genome resequencing and functional genomics is increasing our understanding of mechanisms controlling the impact of mutations. In the closing chapter (Chapter 6), I discuss future avenues of research exploiting these technologies to improve our understand of mutation and repair at a genome-wide scale.

2

Genome-wide variation of the synonymous diversity in *E. coli*

2.1. Summary of the chapter

This chapter describes the exact methodology used to estimate synonymous diversity robustly for 2,930 genes and describes its variation along the *E. coli* genome. These estimates are then used in Chapter 3 to study the role of selection and non-selective factors in the variation of synonymous diversity, and finally to obtain a largely unbiased estimate of the local neutral mutation rate at gene resolution.

2.2. Estimation of synonymous diversity at gene level

The pipeline below was carefully designed to ensure the reliability of the estimates of synonymous diversity by exploiting the availability of multiple sequences to detect typical artefacts such as horizontal gene transfer (HGT), ortholog misidentification, sequencing errors and misalignments.

2.2.1. Ortholog detection

The list of the 34 strains used in the study, including their NCBI IDs and names is provided in Table 2.1. The genome sequences and the gene position annotations were downloaded from NCBI in September 2009 (*ftp://ftp.ncbi.nih.gov/genomes/Bacteria/*).

GI number	*NCBI ID*	*#ORFs*	*Strain name*
49175990	NC_000913.2	4149	*Escherichia coli str. K-12 substr. MG1655*
89106884	AC_000091.1	4226	*Escherichia coli str. K-12 substr. W3110*
16445223	NC_002655.2	5312	*Escherichia coli O157:H7 EDL933*
15829254	NC_002695.1	5230	*Escherichia coli O157:H7 str. Sakai*
24111450	NC_004337.1	4177	*Shigella flexneri 2a str. 301*
26245917	NC_004431.1	5339	*Escherichia coli CFT073*
30061571	NC_004741.1	4061	*Shigella flexneri 2a str. 2457T*
74310614	NC_007384.1	4219	*Shigella sonnei Ss046*
82775382	NC_007606.1	4271	*Shigella dysenteriae Sd197*
82542618	NC_007613.1	4134	*Shigella boydii Sb227*
91209055	NC_007946.1	5021	*Escherichia coli UTI89*
110640213	NC_008253.1	4620	*Escherichia coli 536*
110804074	NC_008258.1	4115	*Shigella flexneri 5 str. 8401*
117622295	NC_008563.1	4428	*Escherichia coli APEC O1*
157159467	NC_009800.1	4378	*Escherichia coli HS*
157154711	NC_009801.1	4749	*Escherichia coli E24377A*
170018061	NC_010468.1	4200	*Escherichia coli ATCC 8739*
170079663	NC_010473.1	4126	*Escherichia coli str. K-12 substr. DH10B*
170679574	NC_010498.1	4743	*Escherichia coli SMS-3-5*
187730020	NC_010658.1	4246	*Shigella boydii CDC 3083-94*
209395693	NC_011353.1	5315	*Escherichia coli O157:H7 str. EC4115*
209917191	NC_011415.1	4679	*Escherichia coli SE11*
215485161	NC_011601.1	4554	*Escherichia coli O127:H6 str. E2348/69*
218552585	NC_011741.1	4353	*Escherichia coli IAI1*
218556939	NC_011742.1	4696	*Escherichia coli S88*
218687878	NC_011745.1	4915	*Escherichia coli ED1a*
218693476	NC_011748.1	4763	*Escherichia coli 55989*
218698419	NC_011750.1	4732	*Escherichia coli IAI39*
218703261	NC_011751.1	4826	*Escherichia coli UMN026*
238899406	NC_012759.1	4084	*Escherichia coli BW2952*
251783564	NC_012892.1	4087	*Escherichia coli BL21*
253771435	NC_012947.1	4228	*Escherichia coli BL21(DE3)*
254160123	NC_012967.1	4205	*Escherichia coli B str. REL606*
254791136	NC_013008.1	5262	*Escherichia coli O157:H7 str. TW14359*

Table 2.1. *List of E. coli strains used in the study.*

Escherichia coli str. K-12 substr. MG1655 was used as the reference genome for gene positions (synteny is generally well conserved among these strains) and functional annotations. It is the best annotated strain and most of the functional genomics data for *E. coli* have been generated using this strain.

Seven *Shigella* strains were included for completeness, since evolutionary studies have established that they belong to the *E. coli* clade (Pupo et al., 2000, Lan et al., 2004). This is also supported by the reference tree obtained in this study, which suggests that at least three groups of *Shigella* strains evolved independently within the *E. coli* clade (Figure 2.2).

We identified one-to-one orthologs using all-against-all reciprocal best-hits by BLASTP (E-value cut-off 10^{-6}). The approach has been shown to be highly sensitive, but not as specific as phylogenetic ortholog-detection approaches (Chen et al., 2007a). To maximise specificity we used a very stringent condition requiring all reciprocal best hits among all genomes to form a near-perfect clique (*i.e.* clusters with >75% of all possible pairs being BLAST best-hits). The combination of BLASTP and the filter for cliques ensured a high quality dataset of one-to-one orthologs. Further, the phylogenetic filters detailed below ensure that there are no instances of ortholog misidentification. Note that such accuracy is made possible only by the use of multiple very closely related genomes.

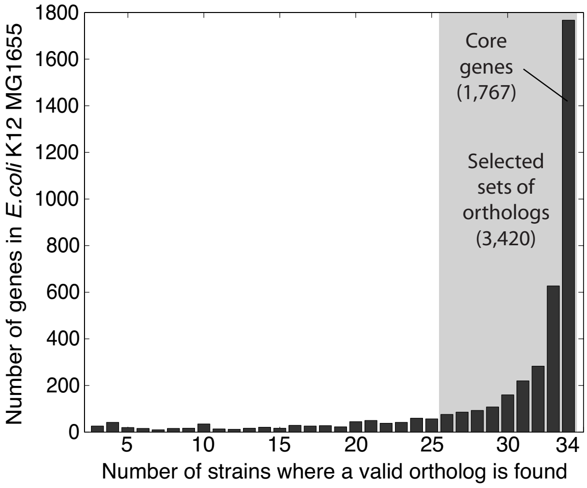

Figure 2.1. *Bar-plot showing the number of E. coli K12 MG1655 genes that find valid orthologs in different numbers of E. coli strains.*

For this study, we focused on 3,420 ortholog sets present in >75% of strains (*i.e.* at least 26 out of the 34), including the reference genome (*E. coli* K12 MG1655). We took great care to ensure that no biases arise from comparing differently sized ortholog sets; for instance, in the estimates of synonymous diversities or in the calculation of Tajima's D.

The reason we did not restrict the analysis to core genes (present in all 34 genomes) was to ensure good representation of the majority of gene functions and pathways. Figure 2.1 shows the number of *E. coli* K12 MG1655 genes with valid orthologs in all or some of the strains. Focusing only on core genes would have reduced the dataset to 1,767 ortholog sets (1,495 after all filters), which is about half the size of our current starting dataset. Note however, that analogous results to those shown in Chapters 2-4 are obtained restricting the analysis to core genes.

2.2.2. Multiple sequence alignment

Multiple sequence alignments were obtained for each orthologs set with PRANK (Loytynoja and Goldman, 2008) using codon sequences and the –F option, as recommended. Given the similarities of strains, analogous results were obtained when nucleotide sequences were aligned without considering their codon structure. Contrary to other aligners, PRANK uses phylogeny-aware gap placement and so reduces the impact of misalignments on the estimation of polymorphisms rates around insertions or deletions (indels).

To increase the reliability of our alignments and to remove gaps and low quality columns we used GBLOCKs (Castresana, 2000) (codon model, default settings). Given the high similarity of the sequences, in most cases GBLOCKs only removed columns with gaps. Removed columns without gaps never accounted for more than 4% of the length of the alignment.

2.2.3. Gene trees and reference tree reconstruction

For each orthologs set, we inferred the maximum-likelihood (ML) gene tree using PhyML (Guindon and Gascuel, 2003) (settings: DNA sequences, starting tree = BIONJ, substitution model = HKY, proportion of invariable sites = 0, number of categories = 4, *gamma* = 1). Given the high similarity between sequences, we used DNA instead of protein sequences to increase the resolution of the tree.

To infer a single reference phylogenetic tree for all the genomes, we concatenated the filtered alignment of 100 randomly chosen core genes (a total of over 80kb of DNA sequence per strain) to build a so-called super-alignment. Again, the tree was inferred with PhyML as described above and using 100 bootstrap replicates to assign confidence to internal nodes.

Figure 2.2 shows a circular phylogram representing the unrooted reference tree of the 34 *E. coli* and *Shigella* sp. strains with non-parametric bootstrap values. The tree shows that the strains are very closely related and the clade has been deeply sampled with the longest branch length being as short as 0.0085. The *Shigella* strains form three clades, which are all located inside the *E. coli* clade. The fact that well-supported internal branches separate the *Shigella* clades from other *E. coli* strains suggest that they emerged independently from *E. coli* at least three times.

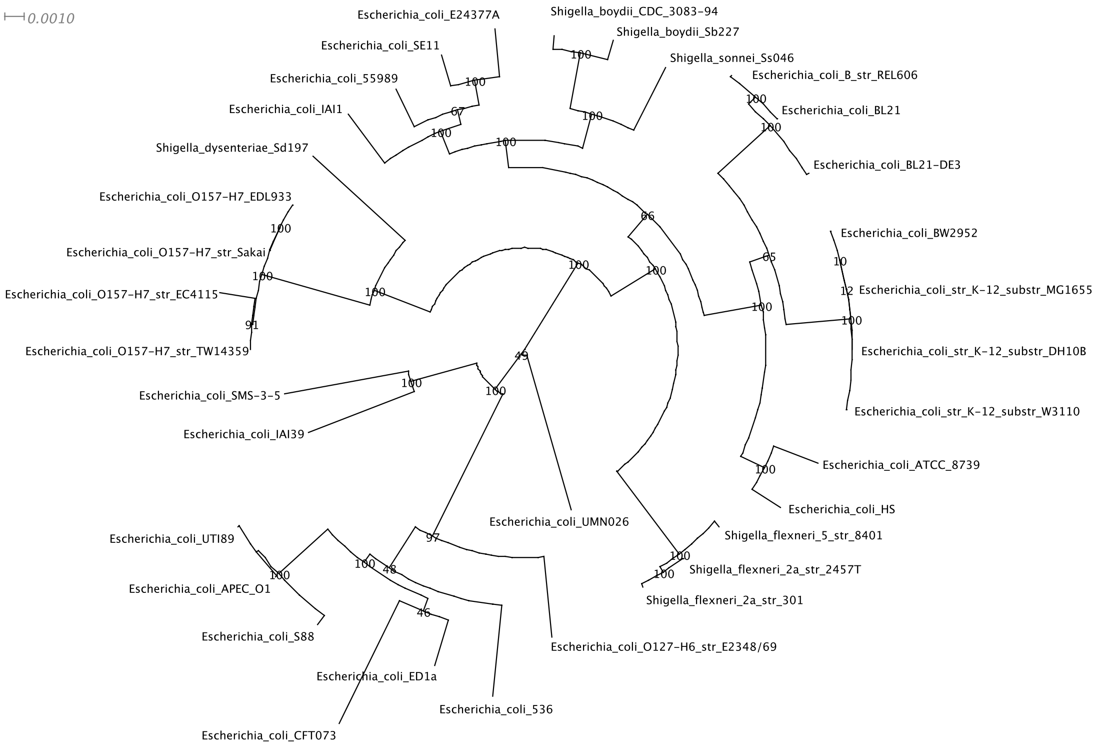
0.0010
Escherichia_coli_E24377A
Escherichia_coli_SE11
Escherichia_coli_55989
Escherichia_coli_IAI1
Shigella_dysenteriae_Sd197
Shigella_boydii_CDC_3083-94
Shigella_boydii_Sb227
Shigella_sonnei_Ss046
Escherichia_coli_B_str_REL606
Escherichia_coli_BL21
Escherichia_coli_BL21-DE3
Escherichia_coli_O157-H7_EDL933
Escherichia_coli_O157-H7_str_Sakai
Escherichia_coli_O157-H7_str_EC4115
Escherichia_coli_O157-H7_str_TW14359
Escherichia_coli_BW2952
Escherichia_coli_str_K-12_substr_MG1655
Escherichia_coli_str_K-12_substr_DH10B
Escherichia_coli_str_K-12_substr_W3110
Escherichia_coli_SMS-3-5
Escherichia_coli_IAI39
Escherichia_coli_ATCC_8739
Escherichia_coli_HS
Escherichia_coli_UMN026
Shigella_flexneri_5_str_8401
Shigella_flexneri_2a_str_2457T
Shigella_flexneri_2a_str_301
Escherichia_coli_UTI89
Escherichia_coli_APEC_O1
Escherichia_coli_S88
Escherichia_coli_ED1a
Escherichia_coli_536
Escherichia_coli_CFT073
Escherichia_coli_O127-H6_str_E2348/69
100
100
67
100
100
100
100
100
100
100
66
100
100
100
100
65
91
100
10
12
100
100
100
49
100
100
100
100
100
97
100
48
46

Figure 2.2. Unrooted maximum-likelihood reference tree of 34 E. coli and Shigella sp. strains.

To minimise the impact of within-species recombination in tree reconstruction, we also used 100 randomly chosen core genes with an intra-gene linkage disequilibrium higher than 0.95 (see Section 3.5.1.1). The resulting tree shows only a few small inconsistencies at internal short branches with the one shown in Figure 2.2. However, importantly, repeating the analyses in Chapters 2-4 with this alternative tree yields identical results.

2.2.4. Tree length and synonymous diversity

There are multiple alternative measures of synonymous divergence (between species) or diversity (within species). Early approaches typically used simplistic approaches based on counting the number of variable sites and dividing it by the total number of sites (Nei and Gojobori, 1986, Ina, 1995). Although some of these early approaches are still commonly used, they do not account for the heterogeneity of the substitution matrix and so they are more vulnerable to sequence composition biases. Currently, the best approaches to estimate synonymous divergence and diversity are based on codon models (Goldman and Yang, 1994), typically using Maximum Likelihood or Bayesian inference.

Here, in a phylogenetic context, we estimate the synonymous and non-synonymous length of each branch of each gene tree, using CODEML from the PAML package (version 4.2) (Yang, 2007). The settings used were: *tree* = ML gene tree from PhyML, *CodonFreq* = F3X4, *clock*, *kappa* = estimated by ML, *omega* = estimated by ML, *alpha* = 0, *rho* = 0. Our results were unaffected by changes in these parameters. We define pS and pN as the total synonymous and non-synonymous tree-length, respectively.

We then used the population genetic parameter θ as a measure of the nucleotide diversity at synonymous sites. Again, we employed a codon model to account for the codon structure of the sequences and for the heterogeneity of substitution matrices. In particular, we used the Bayesian framework implemented in the omegaMap software with parameters: *norders* = 10; *niter* = 150000; *thinning* = 150; *muPrior* = improper_inverse; *kappaPrior* = improper_inverse; *indelPrior* = improper_inverse; *omegaPrior* = inverse; *omegaParam* = 0.01, 100; *rhoPrior* = inverse; *rhoParam* = 0.01, 100; *muStart* = 0.1; *kappaStart* = 3.0; *indelStart* = 0.1; *omega_model* = variable; *oBlock* = 15; *rho_model* = variable; *rBlock* = 15.

Changes in the parameters used did not affect any of the results. In fact, analogous results to all those reported in Chapters 2-4 were obtained using fundamentally different estimates of synonymous diversity such as the mean of the pairwise synonymous differences (π) or the synonymous tree-length (pS), demonstrating the robustness of the observations.

We removed 203 ortholog sets that presented problems in the estimation of pS owing to the presence of ambiguous nucleotides (bases different than A, C, G or T) or a non-codon structure in at least one of the sequences.

2.3. Phylogenetic filters against multiple artefacts

Multiple factors, both methodological and biological, can interfere with the ability to identify genuine point mutations. Among the most important of these factors are ortholog misidentification (*e.g.* due to gene loss in the presence of paralogs), misalignments (often near indels), sequencing errors and horizontal gene transfer (HGT) from different species.

Although the vast majority of previous studies estimating substitution rates relied on pairwise sequence comparisons, it is important to note that these artefacts are very hard to detect when only two genomes are used. Fortunately however, they are easily identifiable using multiple sequence alignments, since they generally cause alignments with one or more unusually divergent sequences (*e.g.* in the case of ortholog misidentification or HGT) or clusters of mutations (*e.g.* due to misalignments in a gappy region -Loytynoja and Goldman, 2008-). In contrast, genuine neutral (or nearly neutral) point mutations should appear scattered along the entire alignment. From a phylogenetic perspective, genes showing a higher (or lower) mutation rate in all strains should simply display a linear expansion (or contraction) of the tree branches.

To avoid potential artefacts, we designed a set of phylogenetic filters to remove any alignments displaying unusual synonymous polymorphism densities among strains or along the gene length. In doing so, we keep only alignments for which the global pS value represents the synonymous polymorphism density in the entire gene and along the genealogy of *E. coli* strains. Later, we manually verified the quality of the alignments and applied several quality controls to confirm that the detected polymorphisms show the patterns expected for genuine point mutations (see Section 2.3.3).

2.3.1. Filters

2.3.1.1. Filter 1: Robustness against within-genes variation of pS

The goal of this filter was to remove those alignments where pS suffers from factors affecting a small segment of the gene length, including potential alignment artefacts causing apparent clusters of mutations.

Each alignment (after GBLOCKs trimming) was divided into bins of 30 codons, and pS was estimated for each bin as described in Section 2.2.4. Clusters of mutations, such as those surrounding gaps, and other local artefacts (including strong local selection) would affect only one or a few bins. Since the median is largely insensitive to outliers, we used the median of pS values from all bins (pS_bins) as a robust estimate of the density of synonymous polymorphisms. We then compared the global pS (potentially sensitive to local artefacts) and the median pS_bins values (robust). As expected, most genes show very high consistency of the two values (Figure 2.3a, Spearman's rho=0.97), whereas a small minority of genes fall clearly outside their expected value.

We applied a linear regression of pS and median pS_bins to identify alignments with inconsistent pS values, using only the central 95% values of core genes. This yielded a very reliable linear fit. To estimate useful confidence intervals, we used the standard deviation of the residuals and the critical value of the normal distribution corresponding to two-sided 99.9% thresholds as the cut-off (3.2905). Since affected alignments generally deviate dramatically from the expected regression line, variations in these cutoffs do not affect the results. The filter identified 159 suspicious alignments.

2.3.1.2. Filter 2: Consistency of pS with the median of normalised pairwise pS values

The goal of filters 2 and 3 was to remove alignments for which pS suffers from factors affecting one or a few branches of the tree (including horizontal gene transfers, ortholog misidentification and sequencing errors in one or a few strains). Given that pS measures the total tree length, factors causing artefactually very long branches would produce extreme deviations of pS. To remove these alignments we used two different filters (2 and 3). Filter 2 is very sensitive to artefacts affecting terminal or peripheral branches, whereas filter 3 is sensitive to factors affecting internal or external branches.

Filter 2 is based on performing all-against-all pairwise comparisons for each alignment and comparing the median of the normalised pairwise pS values with the global pS. Pairwise pS values were calculated by maximum-likelihood using CODEML (as described in Section 2.2.4 but using *runmode*=-2 instead).

Since not all genes have orthologs in all 34 strains, the median of pairwise pS values is not completely comparable among non-core genes. To account for this we collected a matrix of all-against-all pairwise values for all the alignments. Each observed pairwise pS was then divided by the median value of this pair from all genes (*i.e.* observed divergence of a gene in two strains normalised by the genomic expectation). In this way, each alignment provides a vector of normalised pairwise pS values. We used their median (pS_pairs) as a measure of the density of synonymous polymorphisms that is robust to deviations in pS in a few strains but sensitive to changes in the mutation rate affecting all the strains. For example, ideally a gene mutating twice as fast as the genomic average should yield a pS_pairs value around 2.

We then evaluated the consistency of pS and pS_pairs by linear regression, as described for the filter 1. Again, a majority of genes show a high consistency between the two values (see Figure 2.3b, Spearman's rho=0.84). This filter identified 178 suspicious alignments and manual evaluation confirmed that a fraction of them suffered from serious artefacts, most likely due to ortholog loss in the presence of distant paralogs or due to horizontal gene transfer.

2.3.1.3. Filter 3: Consistency of pS with the median of normalised synonymous branch lengths

If a HGT event occurred in a very deep branch of the *E. coli* tree, the newly incorporated sequence might appear in a significant fraction of the strains. This may affect the median of pairwise pS values, in which case the previous filter may fail. Thus, filter 3 was designed to complement the action of filter 2 to remove alignments with artefacts affecting one or even multiple branches, external or internal, including complex scenarios such as those described by (Hughes and Friedman, 2005).

For each alignment, we used CODEML (Section 2.2.4) to calculate the synonymous branch lengths using the topology of the reference tree. Again, we normalised these values to account for differences in numbers of genes in the ortholog sets. For core genes, the reference branch lengths were obtained using CODEML on the super-alignment described in Section 2.2.3. For each non-core gene set, we first subsampled the original super-alignment to exclude the sequences missing in the set and then we reconstructed a new reference tree and estimated the reference branch lengths using it. Dividing the observed branch lengths for each alignment by the reference branch lengths yields a vector of normalised synonymous branch lengths. For each alignment, the median of normalised branch lengths represents a robust measure of the density of synonymous polymorphisms (pS_branches). Ideally, a vertically evolving gene mutating twice as fast as the genomic average should have a median pS_branches value of around 2.

We then evaluated the consistency between the global pS and pS_branches by linear regression, as described for the previous filters. Again most genes show reasonable consistency (Spearman's rho= 0.70, Figure 2.3c), although the strength of the correlation is lower than in the previous filters. A major reason for this is that gene trees do not fully agree with the reference tree, most likely due to within-species homologous recombination (Section 3.5.1). Thus, using the reference tree for the estimation of synonymous branch lengths introduces a considerable amount of noise, as observed here. Increased noise in turn leads to wider confidence intervals and lower sensitivity, but outlier alignments are still detected since serious artefacts cause large deviations in pS. This filter detected 157 suspicious alignments.

2.3.1.4. Filter 4: Robustness of pS against using the ML gene tree or the reference tree

Given that ML gene trees can differ from the reference genomic tree, it could be argued that pS values estimated from the ML gene trees are not fully comparable among genes. Alternatively, estimating pS for each alignment using the same reference genomic tree might raise criticisms since it will generally lead to a slight overestimation of pS. Consequently, this filter was designed to remove alignments whose estimation of pS is overly sensitive to the tree used, and to keep alignments for which inconsistencies between the trees do not affect the estimation of pS.

For each alignment, we estimated alternative pS values using the reference genomic tree (Section 2.2.3) as described in Section 2.2.4 (pS_ref). We then evaluated the consistency of pS and pS_ref using linear regression as described for the previous filters, obtaining a very high correlation (Spearman's rho=0.97, see Figure 2.3d). This filter identified 170 suspicious alignments.

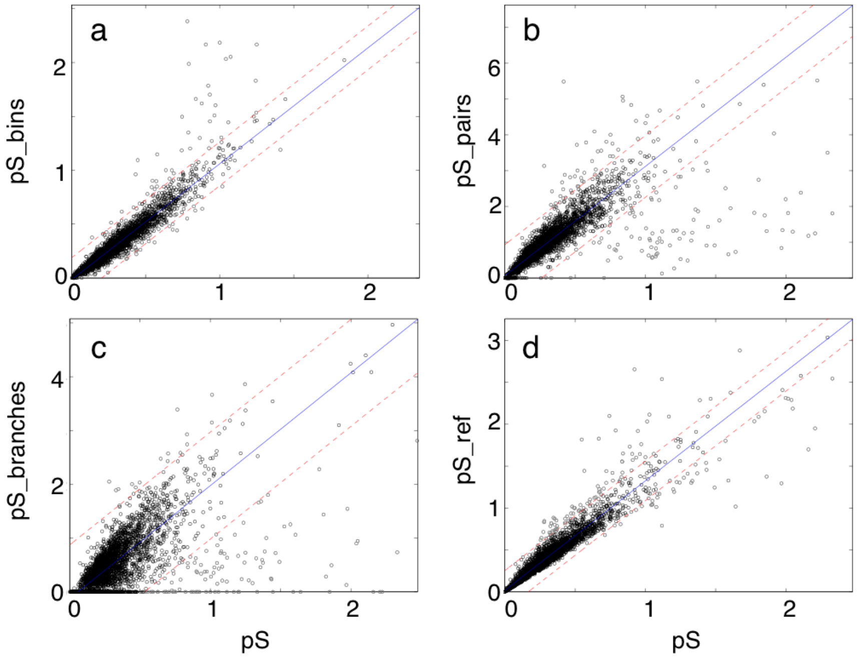

Figure 2.3. *Scatter plots displaying the relationships between global pS values and (**a**) pS_bins, (**b**) pS_pairs, (**c**) pS_branches and (**d**) pS_ref. The solid blue line marks the linear regression in each plot, and the red dotted lines denote the thresholds used to identify suspicious alignments.*

2.3.2. Evaluation of the selection from the filters

The four filters detected similar numbers of potentially artefactual alignments. Although the filters were designed to detect different types of artefacts, in practice alignments with serious problems can be detected by more than one filter and there is a large overlap between the alignments flagged by the four filters (Figure 2.4). Around half of the alignments identified as suspicious by one filter was detected by all four filters, and around 80% by at least two filters.

The redundancy between the filters makes the filtering process very robust. In fact, manual evaluation of alignments flagged by only one filter indicated that they were probably false positives and could be used in later analyses. Nevertheless, to maximise data quality, we removed all alignments flagged by at least one of the filters. This removed a total of 287 alignments and so left 2,930 alignments for further analyses.

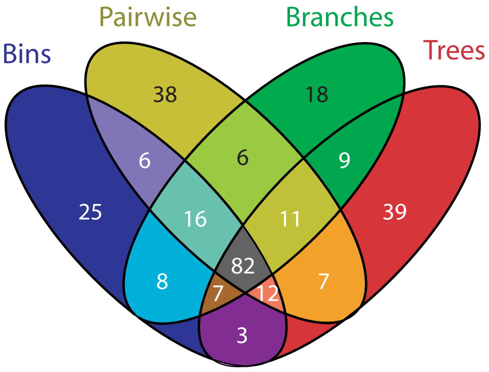

Figure 2.4. *Venn diagram representing the overlaps in numbers of alignments that were flagged by each filter (Filter 1 = Bins, Filter 2 = Pairwise, Filter 3 = Branches, Filter 4 = Trees).*

2.3.3. Quality controls

The four filters by themselves provide strict quality controls. In applying them, we removed possible variations of synonymous diversity due to factors such as horizontal gene transfer from distant genomes, ortholog misidentification and alignment errors, among others. The remaining 2,930 alignments display variations consistent with spontaneous point mutations scattered along the entire gene and among the strains.

To test the quality of the remaining alignments further, we performed the following analyses:

1. Evaluation of the consistency of the substitution matrices.
2. Evaluation of the impact of indels on the polymorphism rate.

2.3.3.1. Substitution matrices

Point mutations are not equally likely for all possible pairs of nucleotides. This is captured in the substitution matrix, which represents the relative rates at which each type of nucleotide (A, C, G and T) is replaced by another. While point mutations are much more frequent between certain nucleotide pairs (depending on the organism), artefacts such as alignment errors around indels are not expected to display a similar pattern.

To assess whether the synonymous polymorphisms observed in each alignment show consistent preference towards certain nucleotide pairs, we estimated the substitution matrix of each alignment using the synonymous SNPs. We used BASEML, from the PAML package, with the following parameters: *tree* = ML gene tree from PhyML, *model* = 7, *clock* = 0, *kappa* = 2.5 (starting value), *fix_kappa* = 0 (ML estimation of kappa), *alpha* = 0, *fix_alpha* = 1 (fixed value), *rho* = 0, *fix_rho* = 1 (fixed value), *npark* = 0, *nhome* = 0, *Mgene* = 0. We used a Generalised Time Reversible (GTR) model for consistency with the lack of an outgroup. We then compared the relative rates with the genomic expectation, evaluating the consistency of the substitution matrices of each gene with the median of all gene matrices. We used Pearson's correlation coefficient as a measure for consistency or similarity.

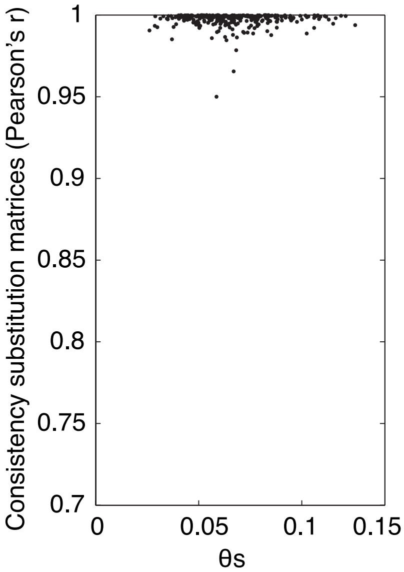

Figure 2.5. *Scatter plot displaying the relationship between the substitution matrices and θs. For each alignment we estimated the GTR substitution matrix and used Pearson's correlation to quantify the similarity of each matrix with the genomic median matrix. The plot is restricted to alignments with more than 100 synonymous SNPs to minimise the noise in the estimation of substitution matrices.*

We observed that individual gene matrices are very consistent with the genomic median. More importantly, we found that this consistency was high regardless of the θs value of the gene (Figure 2.5). Although some local variability in the substitution matrices is expected even for genuine point mutations (*e.g.* caused by transcription-coupled repair, as shown in Figure 4.4), the high consistency observed and the independence from θs fully supports the quality of the SNPs used in this study.

2.3.3.2. Impact of indels

Alignment artefacts have been found to occur more frequently around indels, particularly when using traditional multiple sequence aligners and more divergent sequences. Moreover, recent computational studies have found an increase of substitution rates around insertions and deletions in eukaryotes (Tian et al., 2008) and in bacteria (Zhu et al., 2009). These have been interpreted as evidence of "indel-induced mutations", perhaps caused by error-prone repair associated with the indel. Since both alignment artefacts and indel-induced mutations could in theory affect our estimations of point mutation rates, we performed a detailed examination of the impact of indels on our analyses.

To be strict, we defined indels as gaps embedded inside gene alignments. We excluded gaps extending to the ends of the alignments since they are likely to correspond to alternative start or stop codons rather than to real insertions or deletions.

Importantly, only ~11% of the alignments contain one or more indels. Given the very high similarity of the sequences and the lack of a single indel in most of the alignments, it seems very unlikely that indels have had any significant effect in the present study.

Nevertheless, we further checked whether there is any evidence of "indel-induced mutations" in our alignments and whether they have any impact on our estimations of synonymous diversity. To do this, we calculated the SNP frequency in the sequences flanking an indel, at codon resolution. To account partially for the effect of selection, we distinguished between synonymous and non-synonymous SNPs. The results are presented in Figure 2.6. Despite the noise in the estimation of SNP frequencies (owing to the limited number of indels), we found a higher frequency of non-synonymous SNPs in the proximity of an indel than expected (approximately up to 5 codons away). However, we did not observe any increase in the frequency of synonymous SNPs.

The observation that the rate of non-synonymous SNPs increases in the neighbourhood of indels cannot be explained by an increased mutation rate or by an obvious artefact, since those factors should also affect synonymous SNPs. The most likely explanation is that selection acts differently (transient positive selection or relaxed purifying selection) near an indel. Given that we are looking at indels occurring inside protein-coding genes, an insertion or a deletion is likely to have an important effect on the protein function, thus favouring the fixation of compensatory amino acid changes. However, synonymous sites remain unaffected.

Taken together, these observations suggest that insertions and deletions do not have any detectable effect on our estimations of synonymous diversity.

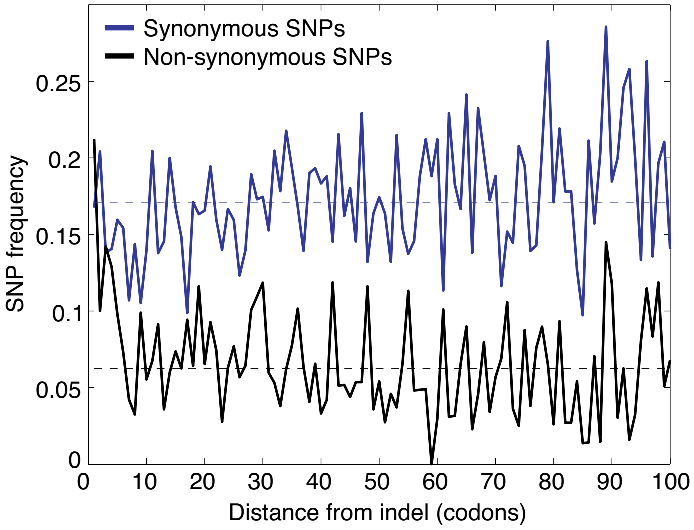

Figure 2.6. *Plot showing the variation in the frequency of synonymous and non-synonymous SNPs with respect to the distance (in codons) from an indel.*

2.4 Synonymous diversity is remarkably variable in *E. coli*

2,930 alignments of near-core genes passed all filters and quality controls, thus showing variation consistent with genuine point mutations. All together, the dataset comprises more than 140,000 single-nucleotide polymorphisms (SNPs) with an average of 49.9 SNPs per gene.

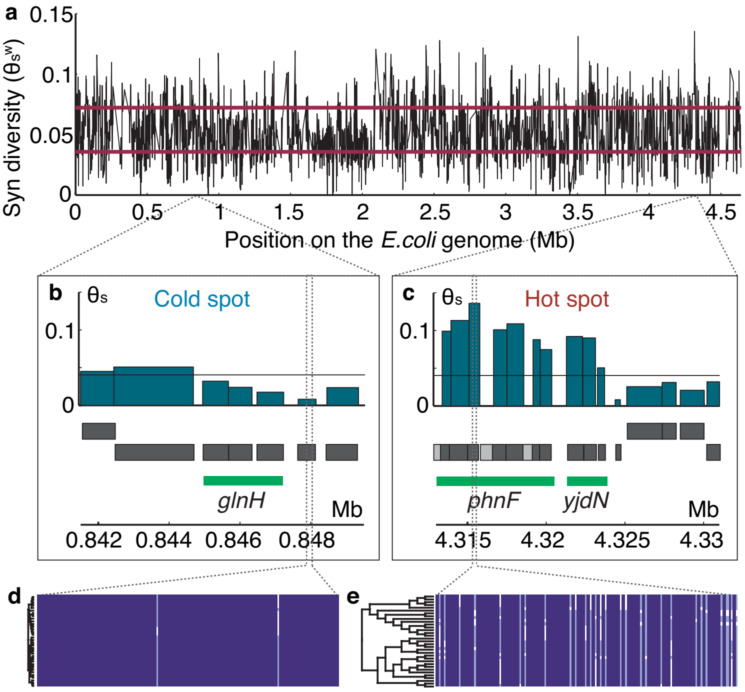

Figure 2.7. *Synonymous diversity along the E. coli genome is heterogeneous. (**a**) Distribution of the 2,930 θs values (using θs^w) plotted along the E. coli K12 MG1655 genome showing that synonymous diversity is highly variable. The red lines indicate the upper and lower 2.5% limits of the expected distribution under a regime of uniform mutation rate. 41% of θs^w values fall outside of the expected range. (**b,c**) Details of two hot and cold regions displaying the genomic coordinates, annotated genes and their respective θs values. Each gene is shown as a black box on the sense or anti-sense strands (grey for genes with no θs available). Green horizontal lines indicate operons, labelled by the name of their lead gene. (**d,e**) Multiple-sequence alignments of 180bp segments of two genes (**d**) dps (θs=0.008; TDsyn=0.35; RNA-Seq level=139.9), and (**e**) phnL (θs=0.136; TDsyn=0.33; RNA-Seq level=1.4).*

2.4.1. Genome-wide variation

Figure 2.7a immediately highlights the striking variability in the density of synonymous polymorphisms along the *E. coli* genome. Although the mean synonymous diversity (θs) is 0.04, these values vary by more than an order of magnitude from <0.002 to >0.10 among genes. This substantial variation is also evident in the alignments shown in Figure 2.7d-e.

In a simple model where mutations occur at a uniform rate along the *E. coli* genome, the density of neutral mutations per gene should also be uniform with a certain sampling variation determined by the number of polymorphisms per gene. To simulate the expected variation of θs due to the limited size of genes and the limited number of mutations observed, we used a Monte Carlo simulation.

For simplicity, we used the Watterson's estimator of θs (θs^w) calculated using:

$$\theta s^w = \frac{S_i}{a_n L_i p}$$

Where S_i is the number of synonymous segregating sites in gene *i*, a_n is the $(n\text{-}1)^{th}$ harmonic number (*n* being the number of sequences), L_i is the length of the alignment after removing codons with non-synonymous changes and *p* (=0.290) is the proportion of synonymous sites in *E. coli* gene sequences.

To be strict we restricted the analysis to the 1,615 core genes (present in all 34 strains) of our set of 2,930 filtered alignments. The total number of synonymous mutations observed in the 1,615 genes was 81,888. We then simulated the random occurrence of each of these 81,888 mutations independently. Under the model of uniform mutation rate and in the absence of any additional biases, the probability of a gene to receive each mutation is $L_i/\Sigma(L_i)$.

Figure 2.8 shows the distribution of the observed θs^w values for the 1,615 core genes (black line), together with the observed distribution of θs^w for 100 Monte Carlo simulations (grey lines) of the 1,615 genes under uniform mutation rates and no additional biases. The 100 simulations yielded very similar distributions, with an average standard deviation of 0.0078 (ranging from a minimum of 0.0074 and a maximum of 0.0083). In contrast, the distribution of the observed θs^w values has a standard deviation of 0.0193 (~2.47 times larger than expected). Similarly, the distribution of observed θs values estimated with the codon model and used as the main measure of synonymous diversity in this study has a standard deviation of 0.0208 (~2.67 times larger than expected).

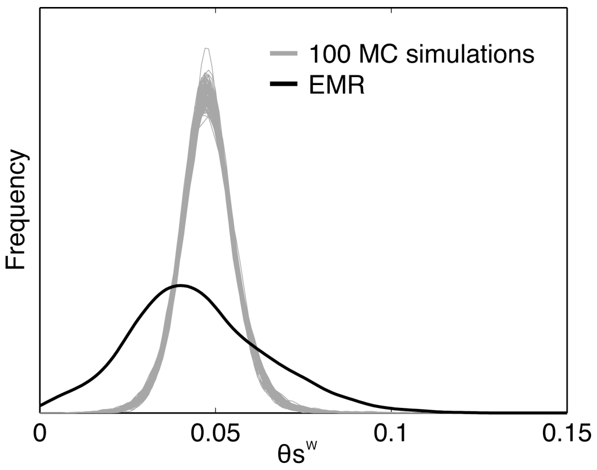

Figure 2.8. Distribution of the observed θs^w *values for all core genes (black) and distributions of* θs^w *observed in 100 Monte Carlo simulations in a regime of uniform mutation rate (grey).*

Chapter 3 carefully evaluates the role of multiple factors on the observed variation of the synonymous diversity along the *E. coli* genome. The main factors studied in Chapter 3 include selection (from known and unknown sources), sequence composition biases and within-species homologous recombination, all of which will be discussed in detail. In this context, however, it is important to note that, after removal of the effect of these biases (see Section 3.5.3), the standard deviation of the corrected θs (θs') is 0.0181 (~2.32 times larger than expected under a uniform mutation rate). This shows that, as it is explained in Chapter 3, the potential biases mentioned above only account for a small fraction of the large variation of θs among *E. coli* genes.

Since within-species homologous recombination will not only cause a directional bias in θs, but it will also increase its variance, correction by regression will not completely remove the effect of recombination on the genome-wide variation of θs. Thus, to avoid the possible overestimation of the θs variance due to within-species homologous recombination, we can restrict this analysis to very low recombining genes (linkage disequilibrium > 0.95 or a proportion of alleles consistent with the reference genealogy > 70%, see Section 3.5.1). Although as expected, the width of the θs' distributions is smaller for very low recombining genes, there is still much larger variation than expected under a uniform mutation rate (std = 0.0156 and std = 0.0177, respectively; ~2 and 2.27 times larger than expected, respectively).

These analyses suggest that the local neutral mutation rate varies substantially among genes in the *E. coli* genome.

2.4.2. Regional variation of the synonymous diversity

As shown in Figure 2.7b-c, the synonymous diversity also presents signs of regional organisation, with neighbouring genes showing more similar θs values. A closer examination of the regional variation of θs and θs′ (corrected by potential biases), demonstrates that mutational hot and cold regions often span entire operons suggesting that gene function has an effect on θs. This is quantitatively demonstrated in Figure 2.9.

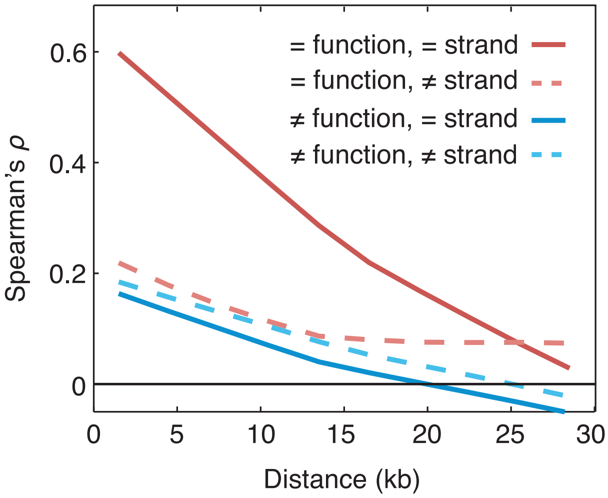

Figure 2.9. *Fitted lines showing the decay of the correlation between pairs of θs values with their genomic distance. Neighbouring genes tend to have similar θs, particularly if they share similar functions and they are encoded on the same strand (red line). Weaker but significant correlations are apparent even for functionally unrelated genes (blue) and those on opposing strands (dotted lines).*

Figure 2.9 shows the decay in the autocorrelation of θs with increasing genomic distance, using Spearman's correlation coefficients. Neighbouring genes tend to have similar θs values, particularly when they are encoded on the same strand and share common functions. In this analysis, genes were considered to share a common function if they had one or more functional annotations in common using the second level of the COG hierarchical functional annotation (Tatusov et al., 2003). A similar decay was observed when restricting the analysis to pairs of genes within the same operon, using the annotation from *ProOpDB* (Taboada et al., 2010).

In addition, a weaker correlation is also apparent for genes with distinct functions and on opposing strands, which suggests that synonymous diversity is affected by regional factors at a resolution of few kilobases.

Analogous results were obtained for θs', after removal of selection and sequence composition biases, among others.

The weight of evidence for an extraordinary claim must be proportioned to its strangeness.

—Pierre-Simon Laplace (1749–1827)
Attributed

3

Evaluation of selective and non-selective biases affecting synonymous diversity

3.1. Summary of the chapter

Chapter 2 demonstrated that within-species synonymous diversity varies largely among genes along the *E. coli* genome. While this may reflect an underlying variation of the neutral mutation rate, there are also several other factors that can affect θs.

Having removed major artefacts in Chapter 2, conceptually the most important confounding factors are selection acting on synonymous sites and homologous recombination. Since these factors can show clear functional biases, they can confound the interpretation of the genome-wide variation of θs. In addition, non-selective factors like sequence composition biases and distance to the origin of replication may also affect the local mutation rate (Ochman, 2003).

Chapter 3 carefully evaluates the role of these factors on the observed variation of the synonymous diversity along the *E. coli* genome. Exploiting the availability of a high quality set of over 140,000 synonymous SNPs, we apply extensive population genetic tests to quantify the relative contributions of selection and non-selective biases to the regional variation of synonymous diversity. Finally, this allow us to obtain an unbiased conservative estimate of the variation of the neutral mutation rate along the *E. coli* genome.

3.2. Impact of known selection factors on synonymous diversity

Selection on synonymous sites is traditionally assumed to be weak. Nevertheless, these sites are not purely neutral, as suggested by the strong negative correlation between the synonymous fixation rate (dS or Ks) and codon usage bias observed in many organisms (Drummond and Wilke, 2008). Further, selection on mRNA-folding stability is also known to act at synonymous sites, particularly in the 5′-end of genes, where a certain secondary structure is required (Eyre-Walker and Bulmer, 1993, Kudla et al., 2009)

The impact of selective forces on the variation of (intra-species) synonymous diversity along the genome is, however, largely unknown. We quantified the impact of selection on θs using multiple alternative approaches. All our analyses indicate that while selection at synonymous sites exists, its impact on the variation of θs is too small to affect our findings. Further, we introduce a regression approach able to remove selection bias to near completion.

3.2.1. Codon Usage Bias

Codon Usage Bias (CUB) refers to differences in the usage frequency of certain synonymous codons in coding DNA, a phenomenon observed in many organisms. This is believed to reflect the action of selection favouring the usage of those codons whose tRNAs are more highly expressed, allowing more efficient and probably more accurate translation (Drummond and Wilke, 2008, Plotkin and Kudla, 2011). CUB is arguably the strongest and most studied source of selection known to act at synonymous sites.

3.2.1.1. Lack of correlation between θs and CUB

The best known evidence for selection at synonymous sites in bacteria is a strong negative correlation between interspecies synonymous divergence (dS) and CUB (Ikemura, 1981, Ikemura, 1985, Drummond and Wilke, 2008). Thus, as a first approach to study selection bias on synonymous diversity (θs), we analysed its correlation with CUB.

We calculated the codon adaptation index (CAI) (Sharp and Li, 1987), using the CUSP and CAI functions from EMBOSS (Rice et al., 2000). The optimal set of codons was calculated with CUSP using all genes annotated as ribosomal proteins in *E. coli* K12 MG1655. As an alternative, we also used all the genes of the reference genome, obtaining analogous results. CUB values were calculated for each orthologs set using the equation: $CUB_i = (CAI_i – CAI_{mean}) / CAI_{std}$.

To check the validity of our CUB estimate we performed several alternative estimations of codon usage bias in *E. coli*. To account for variation among orthologous sequences, CAI was alternatively calculated as the median CAI of all genes of each alignment. To control for known biases in codon usage at both 5′ and 3′-ends of sequences, we trimmed the first 50 and last 20 codons (Eyre-Walker, 1996). Both controls gave analogous results to the original calculation. An alternative (and older) measure of codon optimality, recently used in several papers about selection on codon usage (Drummond and Wilke, 2008, Warnecke and Hurst, 2010), is the fraction of optimal codons in a gene (Fop) (Ikemura, 1981). We computed Fop using CodonW (*http://codonw.sourceforge.net/*) with default parameter settings for *E. coli*, again finding analogous results to those reported in the main text.

Having estimated CUB for all genes, we studied its correlation with θs. Very strikingly, we did not observe any relevant correlation between them for most genes (Spearman's rho=-0.058, Figure 3.1a). Only few genes of very high CUB appear to show significantly lower θs values. Indeed, exclusion of genes with CUB>2.5 removes any residual association with θs (Spearman's rho=0.012, *P*=0.38; Figure 3.1).

In clear contrast, inter-species pairwise values of synonymous divergence (dS) between *E. coli* (strain K12 MG1655) and *Salmonella enterica* (serovar Typhi Ty2) show the strong negative correlation that has previously been reported (Spearman's rho=-0.57, Figure 3.1b). Similarly, contrary to θs, within-species non-synonymous diversity (estimated as the mean of all-against-all pairwise dN values per alignment, π_n) also shows a strong negative correlation with CUB (Spearman's rho=-0.46).

The lack of correlation between θs and CUB implies that selection on codon usage is not strong enough to affect the gene-to-gene variation of within-species synonymous diversity in *E. coli* (θs). While this might appear surprising for a reader only familiar with traditional analyses between species, this is a relatively well-known phenomenon that was originally reported in *Drosophila* and formally explained using a population genetics approach (McVean and Charlesworth, 1999). This study demonstrated that while weak selection on codon usage is strong enough to cause deviations of inter-species dS, it did not significantly affect within-species θs in *Drosophila*. Next section applies this model to *E. coli*.

This analysis shows that, contrary to dS, the gene-to-gene variation of θs is largely neutral with respect to codon usage.

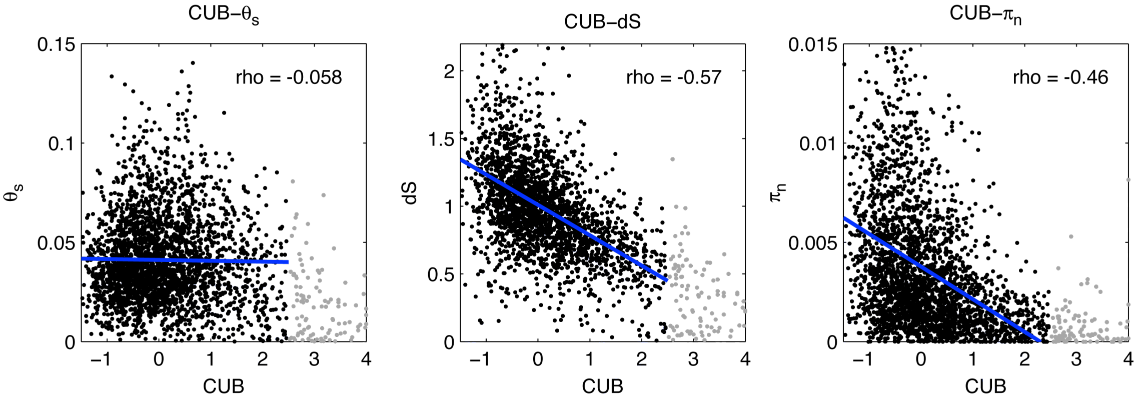

Figure 3.1. *Correlation of (**a**) within-species synonymous diversity (θs), (**b**) inter-species synonymous divergence (dS) and (**c**) within-species non-synonymous diversity (π_n) with CUB in E. coli. Spearman's rho values were calculated for all point as they are more robust to outliers. For the linear fitting (sensitive to outliers), CUB values higher than 2.5 were excluded (3.8% of all points). Exclusion of extreme CUB genes removes any trace of correlation between θs and CUB (Spearman's rho = 0.012, P=0.52; Pearson's r = -0.017, P=0.38).*

3.2.1.2. A population genetic approach to selection on codon usage

To explain the apparent contradiction between the inter-species correlation and the lack of correlation for within-species comparisons, we used a long-standing population genetic model of selection on codon usage (Li, 1987, Bulmer, 1991, Kondrashov, 1995, McVean and Charlesworth, 1999). This is a mutation-selection model in which selection favours the use of preferred codons while spontaneous mutations tend to reduce the optimality of a sequence. Through alternative approaches, different studies (Bulmer, 1991, McVean and Charlesworth, 1999) concluded that using 2-fold degenerate sites, the proportion of optimal codons at mutation-selection-drift equilibrium is defined by:

$$\gamma = 2Ns \approx \frac{ln(P\frac{\mu_{10}}{\mu_{01}})}{ln(1-P)}$$

Where γ is the scaled selection coefficient, *N* is the effective population size, *s* is the selection coefficient, *P* is the fraction of each 2-fold degenerate site with the optimal codon, μ_{10} is the relative mutation rate from the preferred to the unpreferred codon and μ_{01} is the relative mutation rate from the preferred to the unpreferred codon.

Using this equation we estimated γ for each of the twelve 2-fold degenerate sites for all genes in *E. coli* (K12 MG1655) and calculated a mean γ per gene (weighted by the relative frequency of each site). Although assuming equal forward and reverse mutation rates (*i.e.* μ_{10}/μ_{01}=1) is a common practice (Hartl et al., 1994), to obtain a better estimation of γ, we calculated an empirical μ_{10}/μ_{01} ratio for each 2-fold degenerate site from our set of synonymous polymorphisms using the genome of *Escherichia fergusonii* ATCC 35469 (NC_011740) as the outgroup. In particular, we used T>C/C>T = 0.657; A>G/G>A = 0.568 (note that all 2-fold degenerate sites correspond to transitions so there is not need to calculate this ratio for any transversion).

Figure 3.2a shows the observed distribution of population-scaled selection coefficients. In agreement with previous reports (Hartl et al., 1994), most *E. coli* genes showed a relatively small average scaled selection coefficient (γ_{mean}=-0.41, γ_{std}=0.40, and γ>-1 for 93.3% of genes). These selection coefficients are expected to have very small impact on within-species synonymous diversity (θ) but a substantial effect on inter-species synonymous divergence (dS).

According to (Nielsen and Yang, 2003), assuming independence of sites, the loss in divergence (*D*), relative to a neutral reference ($D_{neutral}$), due to selection can be approximated by:

$$\frac{D}{D_{neutral}} = \frac{\gamma}{1 - e^{-\gamma}}$$

And simplifying McVean and Charlesworth's (1999) equation, the loss in diversity (π) relative to a neutral reference ($\pi_{neutral}$) can be approximated by:

$$\frac{\pi}{\pi_{neutral}} = \frac{2(e^{\gamma} - 1)}{\gamma(1 + e^{\gamma})}$$

Thus, a gene under purifying selection with γ=-0.41 will only suffer a reduction in diversity of around 1.4%, but a reduction in divergence of around 19.1%. Similarly, for a gene under purifying selection with γ=-1, the relative loss in diversity and divergence will roughly be around 7.6% and 41.8%, respectively. This explains the apparent contradiction between the lack of selection bias by CUB when using θs and the strong negative correlation of CUB with dS or even π_n (which is expected to be under much stronger selection pressures). Importantly, this also predicts that, contrary to dS, the variation of θs will be largely neutral against weak selection forces (γ>-1).

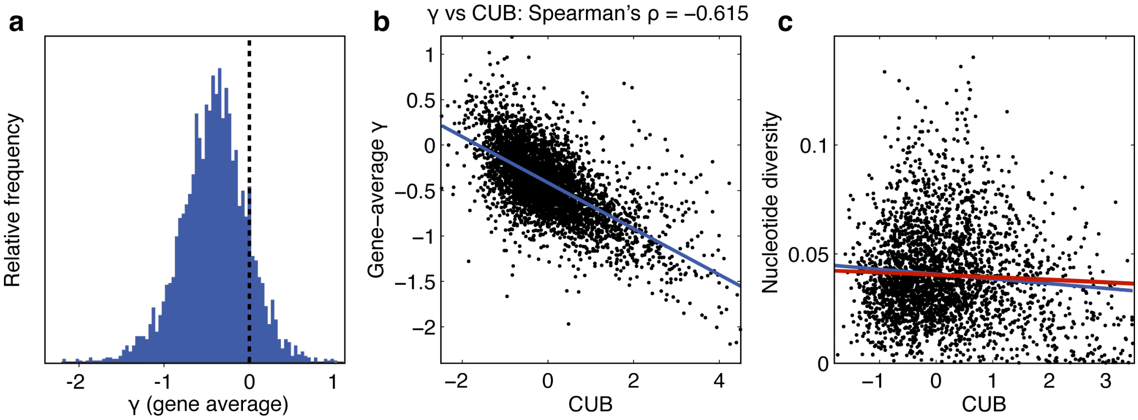

Figure 3.2. *Estimation of γ using a population genetic approach to codon usage. (**a**) Histogram of the population-scaled γ selection coefficients for selection on CUB in E. coli. (**b**) Scatter plot showing the relationship between CUB and the estimated gene-average scaled selection coefficient. (**c**) Scatter plot showing the impact of CUB on nucleotide diversity. The population genetic approach predicts very small impact of selection on codon usage on nucleotide diversity (red line represents the linear regression between CUB and the predicted nucleotide diversity). The prediction agrees notably well with the observed decay in nucleotide diversity with CUB (blue line)*

It should be noted, however, that the mutation-selection-drift model may underestimate the selection (γ) on CUB to a certain extent as it does not account for cases where an unpreferred codon is actually preferred at a particular position. To challenge the validity of the γ estimates, we can compare the predictions of this model to reality. Based on McVean and Charlesworth's model (1999, equation 15 in their paper), we can use the estimated selection coefficients to estimate the impact of codon usage bias on nucleotide diversity.

Figure 3.2c shows the predicted decay in nucleotide diversity with increasing codon usage bias (using a weighted average of the estimated nucleotide diversity). As observed in the actual data, the estimated selection coefficients predict that selection on codon usage would have only very small impact on synonymous diversity (the slope of the linear regression using real data is only slightly more negative than predicted, suggesting that the γ coefficients are only slightly underestimated). This analysis provides a quantitative explanation for the lack of selection bias by codon usage in our dataset.

Altogether, this shows that the large level of variation in θs reported here and its functional associations (described in Chapter 4) are independent of the traditional selective explanations based on codon usage. Further, this implies that global selective forces (*i.e.* acting in most sites) much stronger than codon usage would be needed to substantially affect the gene-to-gene variation of synonymous diversity. Yet codon usage is so far the strongest and most global selection force known to act at synonymous sites in *E. coli*.

3.2.2. mRNA folding and ribosome binding

An additional factor known to act on synonymous sites in *E. coli* is selection to ensure the correct folding and stability of the mRNAs (Kudla et al., 2009), possibly to facilitate the correct binding of the ribosome. This force is thought to act preferentially at the 5′ end of genes, as suggested by evolutionary comparisons showing a decay of CAI and dS in the 5′ end of genes (Eyre-Walker and Bulmer, 1993) and by experiments (Kudla et al., 2009).

Since the effects of this selection force are mainly observed in the first 100 nucleotides after the start codon, it is unlikely to have any impact on our results (note that because of the filters used, every θs value in our dataset is unaffected by changes in θs in a small fraction of the gene length). Nevertheless, to avoid the potential confounding impact of this factor, we repeated all our analyses after trimming the first 50 and last 30 codons of all alignments (Eyre-Walker, 1994), obtaining analogous results.

Thus, the strongest known selective factors acting on synonymous sites in *E. coli* -namely codon usage bias and mRNA-folding stability- do not explain the genome-wide variation in synonymous diversity observed along the *E. coli* genome.

3.3. Impact of selection from unknown forces

The analyses in the previous section indicate that known selective forces acting at synonymous sites are not strong enough to substantially affect the genome-wide variation of θs, at least for the vast majority of genes in *E. coli*. The analyses also imply that any unknown selective force would need to be stronger and at least as global as CUB,to have a significant role in the variation of θs.

This section uses neutrality tests based on the allele frequency spectrum to study the impact of unknown selective factors acting on synonymous sites. Since these tests are not affected by the local polymorphism density, they provide an independent relative estimate of selection which can be used to disentangle the role of mutation and selection in explaining the variation of θs along the genome.

3.3.1. Neutrality tests based on the allele frequency spectrum

The allele frequency spectrum is the distribution of allele frequencies in a set of polymorphisms, where the frequency of a given allele refers to the fraction of the population that carries it. Interestingly, the shape of the spectrum depends on several factors, including selection and demographic processes. For example, purifying selection tends to favour lower allele frequencies since weakly deleterious alleles are progressively removed from the population. Similarly, growing populations, for example due to a demographic bottleneck or a selective sweep, also lead to left-skewed allele spectra. In contrast, balancing selection[14] and population contractions lead to a higher

[14] *Balancing selection* refers to a selective pressure to maintain different alleles in the population.

occurrence of high frequency alleles than expected under neutral evolution and constant population sizes.

A series of summary statistics and neutrality tests have been developed to capture changes in the shape of the allele frequency spectrum. Neutrality tests are commonly used to study selection and demographic processes using polymorphisms datasets. These tests are particularly sensitive to weak selection, and thus are well suited for the study of selection on synonymous sites. Here we used some of the best known neutrality tests, including Tajima's D (Tajima, 1989) and Fu&Li's D* and F* (Fu and Li, 1993).

Tajima's test is based on the fact that under neutrality the number of polymorphic sites (S_n) and the average number of polymorphisms in pairwise comparisons (k) are both simple functions of $\theta=2N\mu$. Tajima's D (TD) measures the difference between the two estimates of θ. In a population of constant size evolving neutrally in which individuals have been sampled randomly, the two estimates should be identical and so TD should be zero. In contrast, TD will be negative when there is an excess of rare (*i.e.* more recent) polymorphisms and positive when there is a excess of common polymorphisms. Thus, negative values reflect purifying selection or population expansion, while positive values reflect population size contraction or balancing selection.

Because of these complex dependencies, a single value of TD cannot distinguish between selection and demographic alternatives such as bottlenecks or population subdivisions. However, this is not a problem when comparing multiple loci since demographic processes affect all sites whereas selection is locus- or site-specific (Tajima, 1989). Furthermore, the conservation of long-distance linkage due to the lack of crossover recombination in bacteria ensures that most loci within a genome share a very similar demographic history and that the effective population size is largely uniform (see Section 3.4 on the effect of hitchhiking and background selection). This contrasts with sexual organisms that have crossover recombination, leading to linkage disequilibrium asymptotically decaying to zero with distance and so to different regions of the genome potentially having different demographic histories and heterogeneous effective population sizes due to hitchhiking and background selection (see Figure 3.7).

As an alternative to TD we used Fu&Li's D* and F*, two measures of neutrality that also exploit the allele frequency spectrum. Using these measures instead of TD yielded similar estimations of selection and analogous functional results.

3.3.1.1. Calculation of Tajima's D and Fu&Li's D and F**

For each alignment, we classified all bi-allelic polymorphisms into synonymous or non-synonymous, and restricted the analysis to codons having a single polymorphic site. For both groups of polymorphisms, Tajima's D was calculated using the *tajima89d_test.m* function of the Matlab package PGEtoolbox (Cai, 2008): TDsyn (synonymous polymorphisms) and TDnon (non-synonymous polymorphisms). Fu&Li's D* and F* were calculated using the *fuli93dsfs_test.m* function.

Schaeffer noted that TD is weakly dependent on the number of sequences used for its calculation and proposed a normalised metric by dividing the standard TD by its theoretical minimum value (Schaeffer, 2002). Since the alignments used here vary from 26 to 34 sequences, we used Schaeffer's normalisation in all the analyses to avoid any potential bias due to sample size. Nevertheless, analyses with and without the normalisation yielded identical results.

As a first test, we confirmed the ability of TD to detect purifying selection in our dataset by comparing the distribution of TDsyn and TDnon, as shown in Figure 3.3. The clear separation of the two distributions, already reported by Touchon *et al.* (Touchon et al., 2009), unambiguously demonstrates the sensitivity of TD in the conditions of this study.

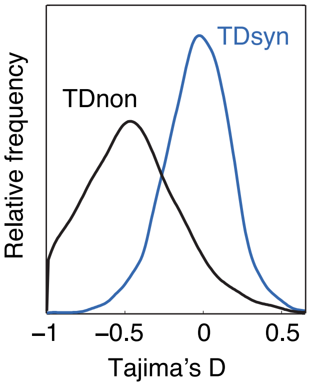

Figure 3.3. *Probability density distributions of normalised Tajima's D at synonymous (TDsyn; blue) and at non-synonymous sites (TDnon; black) across all filtered near-core genes in E. coli. The distributions indicate strong purifying selection at non-synonymous sites relative to synonymous sites. Further, this analysis confirms that the diversity among E. coli strains corresponds to polymorphisms segregating in a population, rather than fixed mutations in relatively isolated subpopulations*

3.3.1.2. Quantification and removal of the selection bias on θs

To quantify the contribution of selection on synonymous sites to the gene-to-gene variation of θs along the *E. coli* genome, we used linear regression. Interestingly, this analysis revealed that the proportion of the variance of θs explained by TDsyn is smaller than 5% ($R^2<0.05$) (Figure 3.4a). This indicates that purifying selection by known or unknown factors does not explain the large variation of θs along the *E. coli* genome. This is consistent with the independently obtained observations above that CUB and mRNA-folding stability have minimal impact on the observed variation of θs. Similarly, the possibility that balancing selection may be a relevant force in the variation of θs is also ruled out by the very weak association of Tajima's D and θs.

A possible source of concern regarding the above analysis is that the estimation of TD may be inaccurate when only few polymorphisms per gene are available, which could lead to an underestimation of the actual fraction of the variance explained by selection. According to a cross-validation test, described in Section 3.3.1.4, TD values can be estimated with high accuracy when around 50 polymorphisms per gene or more are used. Thus, to avoid a possible underestimation of the role of selection in the variation of θs, due to uncertainty in the calculation of TDsyn, Figure 3.4b shows the fraction of the variance of θs explained by TDsyn when the analysis is restricted to genes with a minimum number of synonymous polymorphisms. This analysis shows that even when uncertainty in TDsyn calculations is avoided, selection on synonymous sites does not explain more than ~5% of the variance of θs.

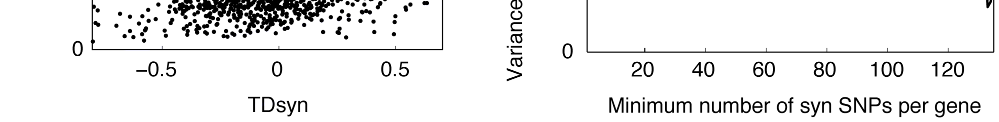

***Figure 3.4**. (**a**) Scatter plot and linear regression of TDsyn and θs. The Pearson's correlation is $r=0.18$, $P<10^{-19}$ ($R^2=0.033$). (**b**) Plot showing that the variance in θs explained by TDsyn remains around 5% even when minimising uncertainty in the calculation of Tajima's D.*

To remove the weak bias on synonymous diversity caused by purifying selection, we used LOWESS (Locally Weighted Scatterplot Smoothing) regression (as described in Section 3.5.3). We challenged the validity of this approach using: (1) a cross-validation test, (2) the allele frequency spectra and (3) extensive forward simulations, as detailed in Sections 3.3.1.4-6. All these analyses consistently support the power of the regression approach to minimise the weak effect of selection on θs.

3.3.1.3. Application of the Poisson Random Field theory

As an alternative to the above method, approaches such as the Poisson Random Field theory (Hartl et al., 1994, Bustamante et al., 2001, Bustamante et al., 2003, Zhu and Bustamante, 2005) can be used to model the effects of selection and demography in the allele frequency spectra. This offers a theoretical framework to estimate absolute values of different population genetic parameters, and to remove the effects of selection on θ (Bustamante et al., 2003). Unfortunately, the existing theory relies on strict assumptions such as independence of sites, random sampling of individuals for sequencing and simple demographic models, all of which are violated in datasets of bacterial polymorphisms.

Therefore for the purpose of this paper, we have used a non-linear regression approach that treats selection (TDsyn) as a confounding co-variable, not making strict assumptions about the population. The tests below demonstrate that this method effectively minimises the impact of selection on the genome-wide variation of θs.

3.3.1.4. Validation 1: Cross-validation test

The removal of the effects of confounding co-variables through regression is a standard statistical strategy. If the predictor variable is measured with precision and accuracy, the removal of the bias should be complete. However, inaccuracy and imprecision in the calculation of TD could limit its predictive power leaving residual bias after the correction. Given that we use a relatively small number of synonymous SNPs to estimate each TD value, imprecision is a reasonable concern.

To test the predictive power of TD we applied a cross-validation test using repeated random sub-sampling. For each alignment, the set of bi-allelic synonymous SNPs was split into a training (learning) set and a validation test (with a fixed size of 20 SNPs, although alternative sizes were explored with analogous results). We estimated TD for both sets of SNPs obtaining two independent estimations: TDsyn$_{\text{training}}$ and TDsyn$_{\text{validation}}$. θs values were then corrected by the effect of selection using TDsyn$_{\text{training}}$ by regression. To quantify the proportion of variance in θs explained by TDsyn that we are able to remove, we correlated θs and TDsyn$_{\text{validation}}$ before and after the correction by TDsyn$_{\text{training}}$. To account for Monte Carlo variation we performed the test 100 times and averaged the results. The ratio $R^2_{\text{after}}/R^2_{\text{before}}$ represents the proportion of the selection bias detected by TDsyn removed by our regression approach (Figure 3.5).

Naturally the imprecision of TD is larger when smaller numbers of SNPs are used to calculate it. In line with this, we observed that the correlation between TDsyn$_{\text{training}}$ and TDsyn$_{\text{validation}}$ increases as more SNPs are used in the calculations. To test the predictive power and the validity of our approach with regard to the numbers of SNPs used, we repeated the cross-validation test using only those alignments with a minimum number of synonymous SNPs.

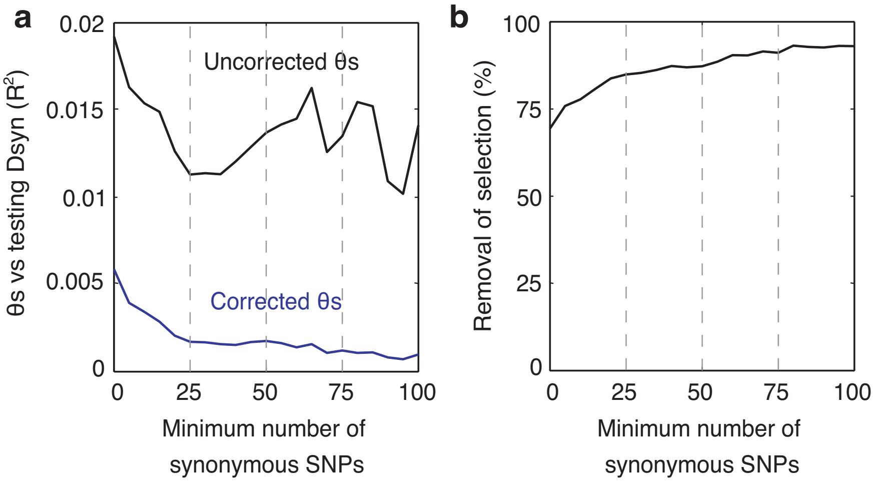

Figure 3.5. Cross-validation test. *Fitted lines showing the result of the cross-validation test challenging the correction for selection bias. (**a**) The correlation of θs and TDsyn$_{validation}$ before and after adjustment by an independent set of TDsyn values (TDsyn$_{training}$) shows that selection bias is largely removed. (**b**) Percentage of selection bias eliminated (R^2_{after}/R^2_{before}), given different numbers of synonymous SNPs in the training set.*

Figure 3.5 shows how the ratio $R^2_{\text{after}}/R^2_{\text{before}}$ increases from around 70% (when all alignments are used) up to over 90% (when only alignments with more than 75 synonymous SNPs are used). Given the already low bias ($R^2<0.05$), a removal of 90% or even lower effectively avoids any significant bias from selection on θs. These analyses demonstrate that we can measure TD with sufficient precision and largely remove its bias on θs.

3.3.1.5. Validation 2: Allele frequency spectra

While the cross-validation test demonstrates that the precision in the calculation on TDsyn is sufficient in this dataset to avoid most of the selection bias, it remains unclear how accurately TDsyn measures the strength of purifying selection acting in the alignment and so how powerful the regression approach is in the quantification and removal of selection bias on θs. The two following controls (Sections 3.3.1.5 and 3.3.1.6) address this question by two alternative approaches.

As described above, Tajima's and Fu&Li's tests detect deviations of the allele frequency spectrum from the neutral expectation. A possible concern with these tests is that they summarise a discrete distribution as a single value, and so distributions with different shapes might produce similar TD values. Though it is unlikely that balancing selection has a significant role in synonymous sites, it is conceivable that a gene experiencing a combination of purifying and balancing selection might lead to a similar TD as a neutral gene.

To test the ability of a summary statistic such as TD to account for the complexity of the whole allele spectrum, in Figure 3.6 we represent the normalised spectra for genes of different synonymous diversities before and after correction by TD. We classified all bi-allelic synonymous SNPs of core genes into four categories according to θs (Figure 3.6-left) or θs' (Figure 3.6-right) quantiles, and all non-synonymous SNPs into a single category. We then calculated the minor allele frequency of each SNP (*i.e.* the number of strains in which the least frequent allele is present), and obtained the folded allele frequency spectra for each group of genes. To facilitate the comparison between the different categories, we normalised each frequency spectrum with the frequency spectrum of all synonymous SNPs.

Figure 3.6-left shows that there is weak selection bias on synonymous sites, consistent with the weak correlation of TDsyn with θs, since the lowest quantile of θs is very slightly enriched in rare alleles. This selection bias is however very small compared with the bias observed at non-synonymous sites. Importantly, correction of θs by all major biases (θs′, see Section 3.5.3), removes most of the weak selection bias apparent in the allele spectra (Figure 3.6-right, note the lack of any clear difference between the genes with the lowest mutating genes -Q1 for θs′- and the horizontal dashed line of all synonymous SNPs). The level of bias correction is remarkable given that all alignments (including those with few synonymous SNPs) were used in this analysis. This, together with the cross-validation test, supports the power of regression on TDsyn to minimise the selection bias on θs.

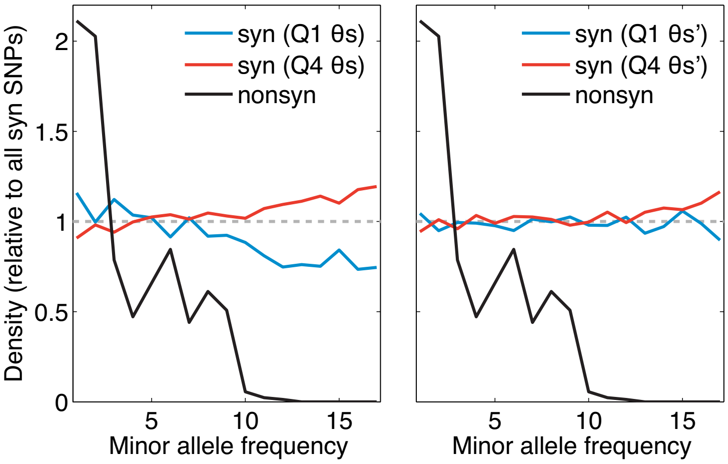

Figure 3.6. *Normalised allele frequency spectra (AFS) for synonymous (blue and red lines) and non-synonymous (black lines) polymorphisms for core genes. The AFS for non-synonymous polymorphisms shows the effects of strong purifying selection at non-synonymous sites. At synonymous sites, genes in the upper and lower quantiles of θs diverge slightly from a neutral expectation (horizontal dotted line), but the deviation is largely absent for θs′ after correction. The number of SNPs used for each line ranges from 8,026 to 33,701.*

The only type of selection that will not leave any signal in the allele frequency spectrum is immediate purifying selection (*i.e.* lethal mutations). However, this is very unlikely to have any serious effect in our estimations for several reasons. First, the frequency of lethal mutations is most likely to be strongly associated with the strength of purifying selection in non-lethal deleterious SNPs, and this is supported by our ability to detect strong purifying selection in non-synonymous SNPs (despite many non-synonymous mutations are nearly immediately lethal). Second, synonymous mutations are very unlikely to be lethal, and an unrealistically large number of lethal synonymous sites would be required to explain the large differences of θs. Third, the lack of correlation of θs with CUB is inconsistent with a strong purifying selection at synonymous sites.

3.3.1.6. Validation 3: Forward simulations

Forward simulations allow us to study the stochastic evolution of a population subject to mutation, recombination, selection, drift and demographic events. We used the forward simulator software SFS_CODE (Hernandez, 2008) to test the ability of our regression approach to remove selection bias under realistically challenging conditions, including strong linkage disequilibrium, recombination without crossover, background selection and non-random sampling of bacterial strains.

The simulations detailed below demonstrate that LOWESS regression on TDsyn successfully removes most of the selection bias under all the conditions tested:

1. The method removes the vast majority of selection bias under the conditions of the *E. coli* dataset (strong linkage disequilibrium and gene conversion).
2. The method is even more sensitive in the presence of recombination with crossing-over (data not shown).
3. Non-random sampling of bacterial strains for sequencing and variable selection coefficients appear to have no-detectable effect on the ability of the approach to remove selection.

Additionally, the observed weak correlation between synonymous diversity and Tajima's D (R^2<0.05, Figure 3.4) is far weaker than observed under simulations assuming a uniform mutation rate across loci (R^2≈0.59, CI95%: 0.55-0.63, from simulations with realistic levels of gene conversion). This further supports the evidence for large neutral variations of the local mutation rate along the *E. coli* genome.

3.3.1.6.i. Description of the simulations

SFS_CODE is a Wright-Fisher style forward population genetic simulation program for finite-site mutation models with selection, recombination and demography (Hernandez, 2008). In each run, an entire population is followed from the beginning of the simulation until the time of sampling. New generations are populated by sampling the previous generation in proportion to the relative fitness of each individual. Mutations are simulated using the Jukes-Cantor model, which occur independently at each site with a probability μ per generation. Full details of the simulation software can be found in (Hernandez, 2008).

Population size

Due to the computational intensity of forward simulations and the need to simulate very long sequences, we restricted the effective population size, and scaled the rest of population genetic parameters accordingly. This is a common and well-accepted practice in forward simulations and has been proven to have no effect in previous studies (Williamson and Orive, 2002, Zhai et al., 2009). As expected, analogous results to those reported here (N = 1,000) were obtained with different values of the effective population size (N = 100, 500, 2000).

Genome structure

We were interested in testing the ability of LOWESS regression on TD to remove selection bias among genes linked in the same genome and subject to mutation, recombination without crossing-over, selection and drift. Thus, we simulated a *model genome* comprising n loci of length L, fully linked and separated by short *intergenic* sequences. Each locus (representing a gene or a group of related genes) was subject to a different selection intensity.

Since all loci are linked into a genome, in the absence of recombination the fate of a single mutation determines the fate of the entire genome. Thus, it is important to keep a realistically low mutation rate per genome and per generation (around 0.003 in *E. coli* and always smaller than 1 in our simulations) (see Appendix 1 for a detailed analysis of the impact of the mutation rate per genome per generation on the effectiveness of selection).

Selection

To simulate selection we used the standard multiplicative model of selection, in which the fitness of an individual is the product of the fitness of each mutation it carries. A new mutation has fitness 1+s, where s is the selective coefficient of the mutation (s>0 indicates positive selection, s<0 indicates purifying selection and s=0 indicates neutrality). The selection coefficient is related to the population scaled selection coefficient $\gamma = 2Ns$. For simplicity, we used a constant selection coefficient for all sites within each simulated *gene* (or locus). Simulations with normally distributed selection coefficients produced analogous results (see Figure 3.10).

Bacterial recombination

Bacterial homologous recombination is characterised by the replacement of native short sequence segments with a homologous sequence from a donor bacteria. Thus, we modelled bacterial recombination as gene conversion events (Chen et al., 2007b) in a diploid population with random mating, using recombination/mutation ratios between 1 and 10, and geometric distributions of fragment sizes with averages varying from 50bp to 200bp (Touchon et al., 2009).

For the purpose of our simulation, the most important effect of bacterial recombination is maintaining a relatively high linkage disequilibrium across the entire genome due to the absence of crossovers.

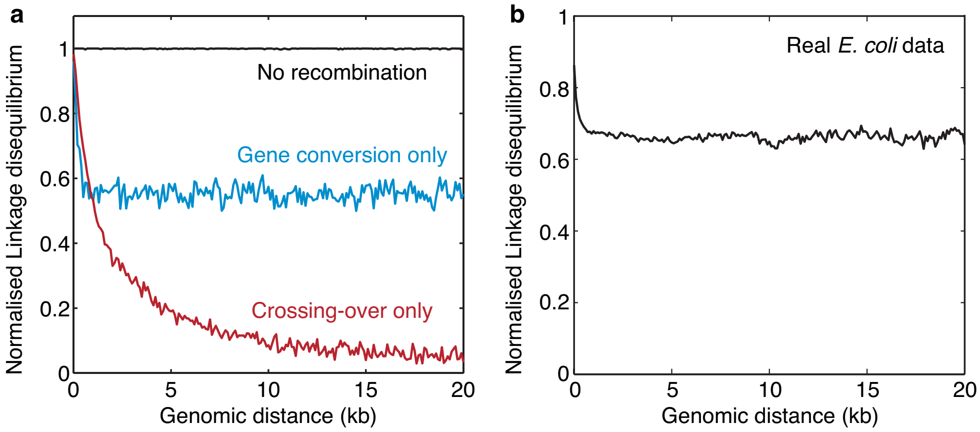

***Figure 3.7**. Plot showing the decay of linkage disequilibrium (LD) with genomic distance. (**a**) Simulated data using different recombination models. (**b**) Real decay in E. coli using over 2.5 million pairs of SNPs. As expected in the absence of crossover recombination and in the presence of gene conversion recombination, an intermediate level of linkage disequilibrium is present along the entire genome, independently of the distance between two loci. LD is calculated using the average |D′| (Section 3.5.1.1) at different distances between two alleles. Since the value of |D′| at linkage equilibrium depends on the allele frequency, to provide a normalised value of LD ranging from 1 (maximum disequilibrium) to zero (linkage equilibrium) we used the equation $LD_{normalised} = (|D'|-|D'_{eq}|)/(1-|D'_{eq}|)$. The value of |D′| expected at linkage equilibrium ($|D'_{eq}|$) was calculated by measuring |D′| after randomly permuting the distribution of alleles for all pairs of SNPs.*

Figure 3.7a shows the decay of linkage disequilibrium with distance in simulations with (a) no recombination, (b) recombination with only gene conversion (as a model for bacterial recombination) and (c) recombination with crossovers (eukaryotes). The linkage disequilibrium decay observed in the simulations of gene conversion resembles the real decay observed in *E. coli* (Figure 3.7b).

Simulation parameters

The results were consistent for different combinations of parameters and so, for the sake of brevity, here we present the results of the following parameters:

- N = 1,000
- $\theta = 0.002$
- n = 11 (11 loci per simulation, separated by 100bp of intergenic sequence)
- L = 40,000 (each locus is 40kb long, the entire simulated genome is thus 0.44Mb long)
- Gene conversion rate (when applicable) = 0.02
- Gene conversion average fragment size (when applicable) = 100
- Crossing-over rate (when applicable) = 0.01
- Scaled selection coefficients (γ) for the 11 loci = (-4, -3, -2, -1.5, -1, -0.75, -0.5, -0.25, 0, 0.25, 0.5)
- Generations: 40N (and a burn-in period of 10N)

The scaled mutation rate was set low (0.002) in order to maintain a mutation rate per genome and per generation as low as possible and always lower than 1 (0.22 using the parameters above). This, together with the low population size, requires to simulate very long genes in order to obtain a realistic number of polymorphisms per locus (around 200 with the parameters above in the absence of selection).

For each simulation, we sampled 34 sequences randomly from the last generation and used them to estimate TD and polymorphism density (using the mean of the Tajima's and Watterson's θ estimators to avoid spurious associations with TD). We then used LOWESS regression on TD (smoothing coefficient = 0.2) to remove selection bias on θ.

3.3.1.6.ii. Removal of selection

The simulations demonstrate that TD is very sensitive to selection, correlating very strongly with the gain or loss of nucleotide diversity due to selection. Within each simulated genome, the average Pearson's correlation between TD and θ was 0.76 (CI95%: 0.73-0.79) for simulations with gene conversion. This contrasts with the much lower correlation found in our dataset for *E. coli* (Figure 3.4), which is consistent with the existence of large variations in the neutral mutation rate among genes.

Figure 3.8 shows the power of the regression approach to remove selection in conditions of strong linkage and gene conversion. Overall, the approach removes a large majority of selection bias, even with selection forces much stronger than those observed on codon usage (note that the strength of selection acting on codon usage is on average γ=-0.41, with 93.3% of genes having γ>-1). Misestimation of the relative θ (with respect to the neutral locus) is lower than 5% for the entire range of selection coefficients tested.

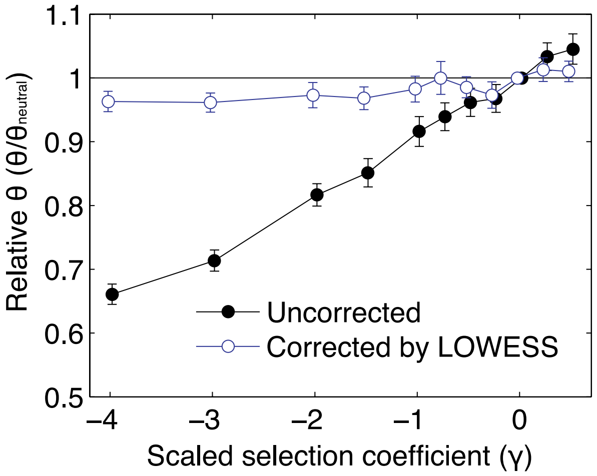

Figure 3.8*. Plot demonstrating the removal of selection bias by LOWESS regression of TD on θ under strong linkage and gene conversion. Relative θ values ($\theta/\theta_{neutral}$) before and after LOWESS correction are shown for 100 independent simulations (circles represent the median and error bars the confidence intervals of the medians).*

3.3.1.6.iii. The small or no-effect of clone-biased sampling

The historical sampling of the 34 *E. coli* strains used in this study was clearly not random. In particular, certain *E. coli* clones are significantly overrepresented in these strains, as it can be observed in the reference tree (Figure 2.2). This violates a common assumption of many population genetic methods, and together with the strong linkage (lack of independence of sites), prevents the use of absolute values of TD or frameworks such as the Poisson Random Field. However, the regression approach uses TD exclusively as a relative estimate of selection and so it is expected to be more robust to deviations from random sampling.

To evaluate the impact of clone-biased sampling, we simulated non-random sampling of strains in our forward simulations. The approach was: (1) select *C* individuals randomly from the last generation, (2) then select the *S* closest individuals of each of the *C* starting individuals (*i.e.* oversampling *C* clones from the population), and (3) select the remaining individuals randomly. For the results below (Figure 3.9) we set *C*=4 and *S*=4. This leads to a non-random selection of individuals in which the final set of genomes has an oversampling of 4 clones (with 4 strains each). This is similar to the number and size of clones in our dataset of *E. coli* strains.

Figure 3.9 shows that the performance of the regression approach is not substantially affected by clone-biased sampling.

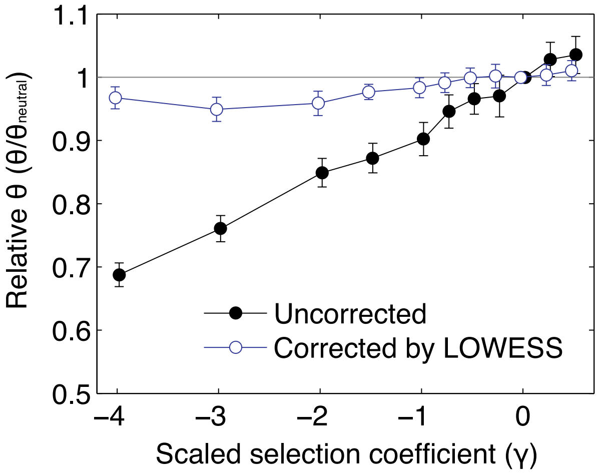

Figure 3.9. *Plot showing the removal of selection bias using LOWESS regression of TD on θ under strong linkage, gene conversion and non-random sampling.*

<u>*3.3.1.6.iv. The small or no-effect of variable selection coefficients*</u>

For simplicity, all simulations above were performed with constant selection coefficient for all sites within each simulated locus. To study the effect of variable selection coefficients on the power of our approach to remove selection bias, we repeated the simulations replacing the constant selection coefficient acting at all sites within each locus (γ_i) with a normal distribution of mean=γ_i and standard-deviation=1. Figure 3.10 shows that this significant level of within-locus variability of γ, did not affect the power of our approach.

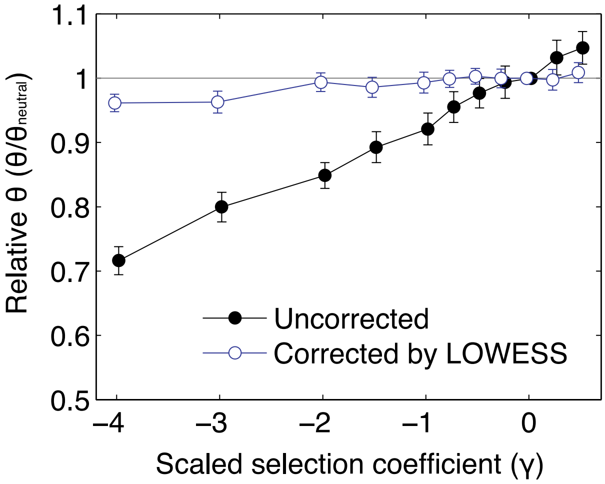

Figure 3.10. *Plot showing the removal of selection bias under strong linkage, gene conversion and normally distributed selection coefficients within each locus.*

3.3.1.7. Summary of the effect of selection on θs

Using polymorphism data and various population genetic approaches here we have shown that:

1. The impact of selection on codon usage or mRNA-folding stability (the two strongest selective forces known to act at synonymous sites) on the genome-wide variation of θs at gene level is negligible. This contrasts with previous observations on synonymous substitution rates from between-species comparisons. However, the observation is consistent with previous observations using within-species synonymous diversity in *Drosophila* and it was formally explained by McVean & Charlesworth (1999). Thus, the variation of θs that we report here is independent of traditional selective models on codon usage, such as Drummond & Wilke's (2008).

2. To account comprehensively for any source of selection, we used population genetic measures of selection based on the allele frequency spectra. This allowed us to calculate the relative strengths of selection acting on genes, and therefore enabled us to disentangle the roles of selection and mutation rate in causing the genome-wide variation of θs. In line with the lack of bias by codon usage or mRNA folding, this analysis revealed that selection from any source explains less than 5% of the observed variation of θs. We demonstrated that this is not due to the lack of power (due to imprecision or inaccuracy) of the allele frequency spectra to detect weak selection. Finally, we showed by forward simulations that the weak selection bias can be removed to near-completion by non-linear regression, without the requirement for strict assumptions about the data.

Both lines of evidence independently and consistently show that the selection bias on the genome-wide variation of θs is very weak and does not account for any of the functional associations described in Chapter 4.

3.4. Impact of selection at linked sites

In addition to direct selection, positive and purifying selection at linked sites (known as hitchhiking and background selection respectively) affect local neutral diversity in eukaryotes (Andolfatto, 2001). Thus, despite the analyses above clearly show that selection at synonymous is not responsible for the large gene-to-gene variation of θs, indirect selection also needs to be accounted for before discarding selection as a confounding factor in our analysis.

Here we show that neither background selection nor hitchhiking, highly relevant in the (eukaryotic) population genetic literature, have any considerable effect on the gene-to-gene variation of synonymous diversity in *E. coli*.

3.4.1. Background selection

In eukaryotes, stronger purifying selection at non-neutral sites has been shown to cause a reduction in the effective population size at linked sites, leading to a local decrease in neutral diversity (Andolfatto, 2001).

Interestingly, however, it can be shown that background selection can only preferentially affect local diversity in the presence of crossover recombination. In the absence of crossover recombination, even at moderately high rates of non-crossover recombination, purifying selection at a non-neutral site extends through the entire genome and cannot cause a local relative decrease of nucleotide diversity. Given that in *E. coli* the rate of recombination and the rate of mutation are roughly in the same order of magnitude (Touchon et al., 2009), and that a bacterial recombination event affects only a very small fraction of the genome (Touchon et al., 2009), background selection cannot cause any significant variation in the effective population size along a bacterial genome.

This can be easily shown by simulation. Using the parameters described above (Section 3.3.1.6.i), here we simulate 7 linked coding loci in which non-synonymous sites are under selection (γ_i = -6, -5, -4, -3, -2, -1, 0) and synonymous sites are purely neutral. Again, crossover recombination and non-crossover recombination (gene conversion) are simulated as described above. Figure 3.11a shows that in the presence of realistic levels of non-crossover recombination, selection on non-neutral sites cannot preferentially affect local neutral diversity.

3.4.2. Hitchhiking

Hitchhiking is the phenomenon by which positive selection on a given mutation extends to all linked sites. In the absence of recombination, positive selection will increase the frequency of the positively selected allele together with all other mutations carried by the individual. In the presence of bacterial recombination, the segment of DNA carrying the positively selected mutation can spread horizontally to other individuals, which could potentially cause a local reduction of the neutral diversity around the site undergoing the selective sweep (Tenaillon et al., 2010, Schubert et al., 2009).

To study the impact of hitchhiking at realistic or even moderately higher levels of non-crossover recombination, we repeated the simulations above but using positive selection coefficients (γi = 0, 1, 2, 3, 4, 5, 6). According to these simulations (Figure 3.11b), even at relatively high recombination/mutation ratios, linkage remains the dominant force and hitchhiking appears to have very limited power to cause substantial deviations of the relative neutral diversity. Much stronger positive selection may, however, have a larger impact on local neutral diversity, although at realistic recombination rates linkage is likely dominate.

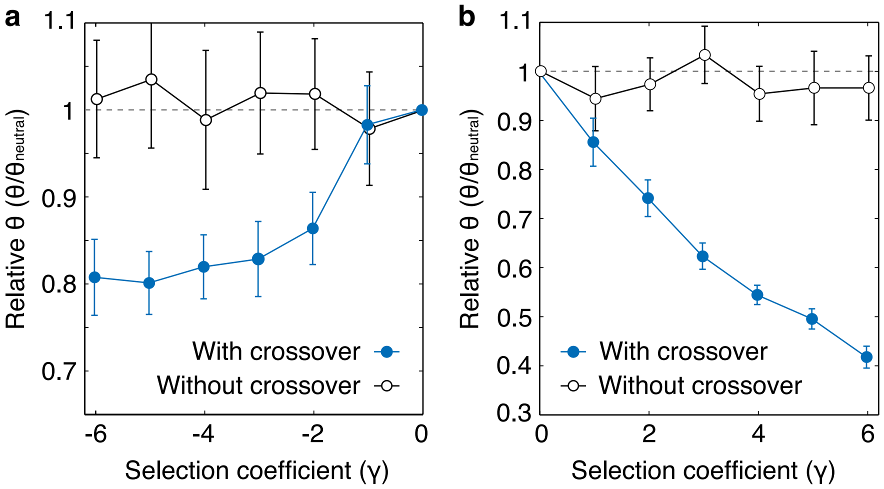

***Figure 3.11**. Simulations of the effect of (**a**) background selection and (**b**) hitchhiking on neutral diversity under crossover and non-crossover recombination. We simulated 7 coding loci linked in a genome, with neutral synonymous sites and non-neutral non-synonymous sites. Black lines represent simulations with gene conversion and blue lines represent simulations with crossover recombination. Relative θ values ($\theta/\theta_{neutral}$) are shown for 100 independent simulations (circles represent the median and error bars the 95% confidence intervals of the medians).*

In any case, hitchhiking could never explain the pattern of θs that we observe here. Assuming that it could be a relevant force, at most, hitchhiking would predict lower θs values in regions undergoing positive selection, while as we will describe in Chapter 4 we see lower θs values in genes under stronger purifying selection. Moreover, two independent observations also indicate that hitchhiking has a minor role in the overall variation of θs: (i) functional categories expected to undergo frequent positive selection do not appear to have lower or higher θs values (Tables 4.1 and 4.2) and (ii) we do not see any significant correlation between θs and the frequency of amino acids under positive selection (Figure 4.2).

Thus, neither background selection nor hitchhiking appear to have any effect in the gene-to-gene variation of θs in *E. coli*.

3.5. Non-selective factors affecting synonymous diversity

So far we have shown that despite the action of some selection factors at synonymous sites, the observed gene-to-gene variation of within-species synonymous diversity in *E. coli* is largely neutral. This suggests that the neutral mutation rate in *E. coli* varies substantially among genes. However, factors other than selection may also have an impact on θs, most importantly homologous recombination and regional variations in sequence composition.

Here we analyse the role of major non-selective factors in the variation of synonymous diversity in *E. coli*, to provide a more reliable representation of the variation of the neutral mutation rate along the genome.

3.5.1. Homologous recombination

The best known ways by which homologous recombination can impact on nucleotide diversity are hitchhiking and background selection. Having ruled out these factors, homologous recombination would only be expected to preferentially affect nucleotide diversity at certain loci if the rate of recombination neutrally varies along the *E. coli* genome. In such case, the main ways by which recombination could affect θs are:

1. Pervasive homologous recombination among *E. coli* genomes that are at least as similar to each other as the strains in this dataset. This would tend to erase variation among strains, leading to a negative correlation between the local rate of recombination and θs.

2. Homologous recombination among species or distant strains. This would lead to an increase in diversity with higher local recombination. However, events introducing a substantial number of changes would have left obvious traces in the alignment and should have been detected by our filters (Section 2.3.1).

3. Finally, if the process of homologous recombination is mutagenic itself, for example by error-prone recombinational double-strand-break repair (Bull et al., 2001), this could also lead to a positive correlation between local recombination rate and θs.

Here we find a positive correlation between the recombination rate and θs, as we describe below. In the eukaryotic literature this is often interpreted as an evidence for mutagenic recombination, hitchhiking or background selection (Andolfatto, 2001). Having shown that neither hitchhiking nor background selection are involved in the variation observed on θs in *E. coli*, the observed positive correlation between recombination and θs suggests that recombination is *mutagenic* (either directly by processes such as error-prone repair, or indirectly through more frequent recombination with more distant strains).

3.5.1.1. Quantification and removal of the recombination bias

For each gene all bi-allelic synonymous SNPs were used to calculate the degree of linkage disequilibrium (D, |D′| and r^2) using the *linkdisequ.m* function of the PGEtoolbox Matlab package.

Even when using the normalised measured |D′|, there are two main factors that can affect the comparability of linkage disequilibrium values among genes: the density of polymorphisms and the length of the genes. A higher density of polymorphisms is more likely to detect loss of linkage by gene conversion and longer genes might be expected to have lower linkage on average than shorter genes. To avoid these confounding effects, for each gene we used a sliding window of 300 bps (with overlapping windows shifted by 30 bps) and randomly sampled 4 SNPs inside each window. For each window we then calculated the mean |D′| for all pairs of sampled SNPs, and we used the average of all windows as the linkage disequilibrium value of a gene. This measure should naturally be independent of the density of polymorphisms and so it allows us to study the relationship between θs and recombination while avoiding spurious associations. We also explored alternative parameter settings (window size and number of SNPs sampled) and other measures for linkage disequilibrium (including the standard |D′| and r^2), obtaining analogous functional results.

As an alternative independent measure of homologous recombination, we exploited the underlying phylogenetic signal of the data by quantifying the fraction of bi-allelic synonymous SNPs that are consistent with the reference genomic tree. It has been shown that recombination in *E. coli*, although very common, does not obscure the phylogenetic signal when enough polymorphisms are used (Touchon et al., 2009). So the fraction of inconsistent alleles could be used as a measure of recombination in bacteria.

In theory, the reversal of mutations or two independent occurrences of the same substitution in the same site would produce phylogenetically inconsistent alleles, which would be interpreted as potential recombination events by any of our measures. To quantify the degree of linkage disequilibrium and the fraction of inconsistent alleles expected by these processes, we simulated the evolution of sequences along the genomic reference tree using the observed substitution matrices with INDELible (Fletcher and Yang, 2009). The simulations showed that only a very small fraction of the observed inconsistencies is expected in the absence of recombination.

To quantify the potential effect of homologous recombination on θs, we studied the association between θs and the normalised measure of linkage disequilibrium (Figure 3.12a) or the fraction of phylogenetically consistent sites (Figure 3.12b). Both measures show a negative correlation with θs (Spearman's rho=-0.36, P<10^{-82}; and rho=-0.21, P<10^{-27}, respectively). Repeating this analysis using θs corrected by GC content, genomic location and selection (see Section 3.5.3), produced lower but still strongly significant correlations (Spearman's rho=-0.28, P<10^{-48} for |D'|; and rho=-0.16, P<10^{-16} for the fraction of consistent SNPs). Such negative correlations reflect a positive association between the local frequency of recombination and θs in *E. coli* (note that higher values of linkage disequilibrium and of the fraction of consistent SNPs reflect lower recombination rates).

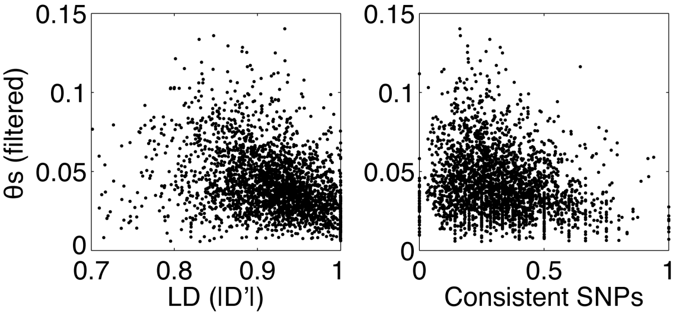

Figure 3.12. *Scatter plots showing the correlation of θs with two alternative measures of within-species homologous recombination: (**a**) a normalised measure of linkage disequilibrium (|D'|) and (**b**) the fraction of SNPs consistent with the reference genealogy.*

As discussed above, having ruled out hitchhiking and background selection, the best explanations for this negative correlation are: (i) a direct mutagenic effect of homologous recombination (Bull et al., 2001) or (ii) the introduction of diversity by more frequent recombination among strains that are more distant than those studied here (but not too distant to leave large artefacts in the multiple alignments). While the first explanation can be interpreted as a genuine source of increased local mutation rate, the latter could affect the estimation of the mutation rate based on θs values.

Therefore, to be conservative in our estimates of the variation in the mutation rate from θs, we handled homologous recombination as a potential source of bias. Linear regression suggests that homologous recombination explains around 10% of the variance in θs (R^2=0.097 using |D'| and uncorrected θs values). Using LOWESS regression we can eliminate the tendency for higher local recombination to increase local θs values. In particular, to be conservative, we used |D'| as the estimator of recombination rate since this measure showed higher correlation with θs variance. Despite fundamental differences between |D'| and the fraction of consistent SNPs, the removal of the association of θs with |D'| also removed the association with the fraction of consistent SNPs (R^2<0.01 after correction by |D'|).

3.5.2. Other non-selective biases

Selection and recombination are conceptually the most relevant biases as they could confound any functional interpretation of the genome-wide variation of θs. Nevertheless, we also evaluated the role of other factors that have been described to associate with synonymous substitution rates (Ochman, 2003).

3.5.2.1. Sequence composition bias (GC content)

The GC composition of each gene was measured as the GC% of the third position of the codons (GC3) using the *E. coli* K12 MG1655 sequence. We also used other measures, including the GC% using all sites of the reference genome, the average GC% of all sequences on each alignment or the GC3s (the GC3 at synonymous codon positions only, using CodonW -*http://codonw.sourceforge.net/*-). All measures provided analogous associations with θs and gave identical results.

Figure 3.13 shows the association of θs with GC3 (Spearman's rho=0.32, $P<10^{-65}$). This association has already been observed before (Ochman, 2003), and it explains more variance of θs ($R^2=0.10$) than any of the other factors studied here. A frequent explanation is that genes with high GC content show more synonymous polymorphisms because the C-to-T transition is the most common base substitution in *E. coli*. However, since θs has been calculated using a codon model and represents the rate of synonymous transversions (Wilson and McVean, 2006), the observed correlation between θs and GC content is more unexpected.

Part of this association may indeed result from the genuine variation of the neutral mutation rate along the genome. Nevertheless, to be conservative, we removed any possible effect of GC content on θs using LOWESS regression (see below).

3.5.2.2. Distance to the origin of replication

Some comparative genomic studies have detected a slight increase in synonymous substitution rates (dS) with increasing distance from the origin of replication in several bacterial species (Sharp et al., 1989, Mira and Ochman, 2002). This trend is believed to be caused by the increased occurrence (or lack of repair) of certain substitutions as replication approaches the terminus, although an experimental study found no evidence of this process (Hudson et al., 2002).

Here we found a near insignificant association between θs and the distance from the origin of replication (using the central coordinate of the *oriC* in the K12 MG1655 genome: coordinate 3,923,845) (Spearman's rho=-0.06, *P*=0.0029). This very weak association became insignificant after the effect of local GC composition (GC3) was removed from θs (Spearman's rho=0.015, *P*=0.45).

3.5.3. Adjustment for the effect of selective and non-selective factors

As described above, some of the associations of θs with certain factors might indeed be a consequence of the genuine variation of the neutral mutation rate along the genome. Nevertheless, since the null hypothesis of this study is that the neutral mutation rate is uniform, to avoid any confounding effect from any of the previously mentioned factors, we used LOWESS (Locally Weighted Scatterplot Smoothing) regression to remove their potential effect on θs.

Using the 2,930 θs values that passed the filters (Section 2.3), we performed LOWESS regression (smoothing factor = 0.5) of GC3 on θs. For each θs value, we subtracted the predicted θs value given its GC3 value and then added the genomic average of θs. These residuals provide estimates of θs corrected by GC content. Similarly, we removed the other biases by applying LOWESS sequentially on: GC3, genomic location, CUB, TDsyn and linkage disequilibrium (using the unbiased calculation of |D'|). The order of the factors did not affect the results and alternative smoothing factors yielded analogous results.

This analysis was applied to the 2,659 alignments that showed sufficient level of variation to allow the calculation of TDsyn and normalised linkage disequilibrium. The resulting residuals represent an estimate of θs unaffected by all the above factors, which we call θs'. Figure 3.13 shows the association of the five potential biases with θs before and after the LOWESS correction of each factor.

Alternatively, we also corrected θs simultaneously for all factors using multiple linear regression, obtaining analogous results in all the remaining analyses.

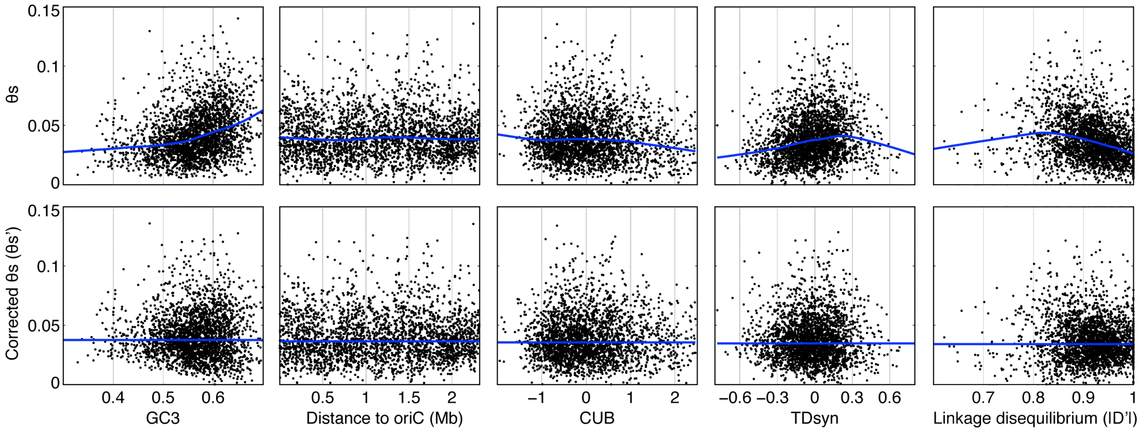

Figure 3.13. *Scatter plots showing the correction of the four potential biases using LOWESS regression. The first row shows the association of a given factor with uncorrected or partially corrected θs values and the second row shows the association after each correction.*

These corrections assume that any associations between the potential biases and θs reflect direct causal confounding biases. It is possible however that part of these associations are caused by the genuine variation of the local mutation rate, making our approach very conservative. In other words, it is likely that this approach overestimates the extent of the biases on θs, causing θs′ to underestimate the actual variation in the neutral mutation rate in the *E. coli* genome.

In summary, θs′ represents a conservative estimate of the local neutral mutation rate, unaffected by purifying selection, recombination and sequence composition. It must be stressed that all these factors combined explain a minority (~20%) of the variance in θs. Thus corrections have a relatively small effect on θs values, and so there is a very high correlation between θs and θs′ (Pearson's r=0.85, Figure 3.14).

Performing all the remaining analyses with either θs or θs′ yields qualitatively identical results. Nevertheless, to be conservative and avoid ambiguity in the interpretation of the results, in Chapter 4 we use θs′ as an unbiased measure of the local neutral mutation rate at gene resolution in the *E. coli* genome.

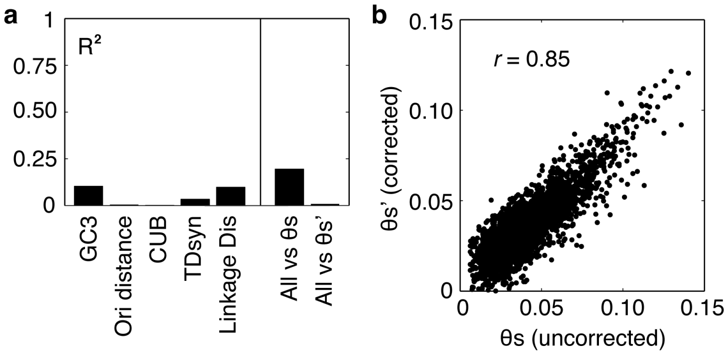

Figure 3.14*. Impact of major selective and non-selective biases on the gene-to-gene variation of θs in E. coli. (****a****) Bar plots displaying the proportion (R^2) of the variance of θs explained by five selective and non-selective potential biasing factors. All factors combined explain around 20% of the variance in θs. Their effect is eliminated from θs′ after LOWESS correction. (****b****) Scatter plot showing the correlation between θs vs θs′. Removal of the potential biases has only a moderate effect on the overall variation of θs. Qualitatively identical functional associations are observed with or without the adjustment.*

A fundamental tenet of evolutionary biology is that mutations are random events. This tenet does not mean that mutation rates are unaffected by environmental factors or that all portions of the genome are equally susceptible to mutations. (...) Rather, the randomness of mutation refers to the supposition that the likelihood of any particular mutational event is independent of its specific value to the organism.

—Richard E. Lenski and John E. Mittler, *The Directed Mutation Controversy and Neo-Darwinism* (1993)

4

Non-random variation of the neutral mutation rate in *E. coli*

4.1. Summary of the chapter

This chapter directly tackles the fundamental question of whether mutations are truly random in *E. coli*. The concept of randomness in this context is nicely explained in the citation above, by Lenski and Mittler.

As explained in the introduction, this question has attracted great interest for decades (Kimura, 1960, Kimura, 1967, Levins, 1967, Leigh, 1970, Cairns et al., 1988, Lenski and Mittler, 1993, Moxon et al., 1994, Radman et al., 1999, Wright, 2000, Sniegowski et al., 2000, Tenaillon et al., 2001, Bjedov et al., 2003, Denamur and Matic, 2006, Moxon et al., 2006, Pal et al., 2007). Nevertheless, currently there is little evidence that local mutation rates have been optimised during evolution (Hodgkinson and Eyre-Walker, 2011), with the limited exceptions of bacterial contingency loci and somatic hypermutation in the vertebrate immune system (Section 1.2.2.2).

Chapters 2 and 3 provide extensive evidence supporting θs′ as a conservative unbiased estimate of the local point mutation rate, unaffected by selection, recombination and sequence composition biases. This now allows us to study the association between the local mutation rate and gene function in *E. coli*. To do so, we use gene functional annotations and a variety of functional genomics data, from experimental information on the fitness cost of gene knockouts to RNA-sequencing and RNA-polymerase ChIP-sequencing. In addition, the last part of the chapter studies the role of known sources of mutation rate heterogeneity in the variation observed in *E. coli*.

4.2. Association of the mutation rate with gene function

4.2.1. Association with selection on the protein sequences

One of the best measures of the fitness cost of a mutation at a given gene is the strength of purifying selection on non-synonymous sites (*i.e.* on the protein sequence). Thus, here we use Tajima's D from non-synonymous polymorphisms (TDnon), as an estimate of purifying selection on protein sequences. Unfortunately, due to the very low protein-sequence diversity among *E. coli* strains (a median of 8 non-synonymous SNPs per alignment), TDnon values are expected to be highly noisy.

Nevertheless, Figure 4.1a provides initial evidence of the association between the neutral mutation rate and the local strength of selection on the protein sequence: namely, proteins under stronger purifying selection appear to experience a lower neutral mutation rate.

4.2.2. Association with gene essentiality

To build on this observation, we then studied the relationship between the mutation rate and gene-essentiality. A single-gene knockout library for most genes in the *E. coli* genome (the "Keio collection") was published in 2006 (Baba et al., 2006). Out of 4,288 genes targeted, non-lethal mutants were recovered for 3,985 genes in a rich medium, which identified 303 "essential" genes (whose inactivation is lethal in any or most conditions).

Using this list we observed that essential genes show a significantly lower neutral mutation rate than non-essential genes (*P*=0.001, Wilcoxon rank-sum test). The difference however is relatively small (12% lower median mutation rate in essential genes). Alternatively, Figure 4.1b shows the fraction of essential genes among those displaying very low (<5th percentile), low (5-25), medium (25-75), high (75-95) and very high (>95) θs′. Again, genes with lower θs′ are enriched in essential genes compared with those of higher θs′.

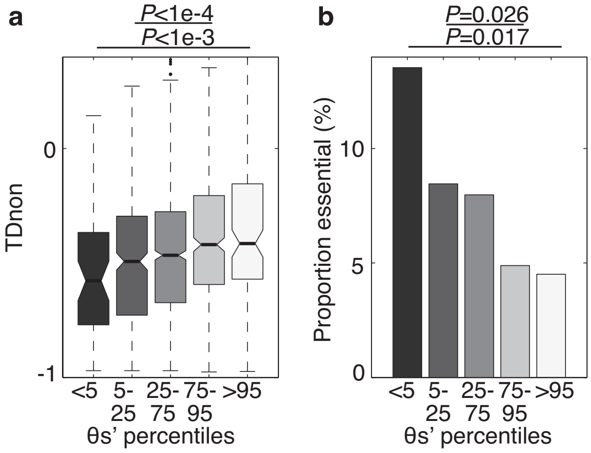

Figure 4.1. Variation in the mutation rate shows functional dependence. *Box- and bar-plots displaying the relationship between gene function and mutation rate. Genes were classified as displaying from very low (<5th percentile) to very high (>95) θs′ values. P-values correspond to: a. Wilcoxon rank-sum test; b. Fisher's exact test. (**a**) Genes with higher mutation rate tend to experience weaker purifying selection at the protein level (measured using TDnon for genes with at least 5 non-synonymous SNPs). (**b**) Genes with lower mutation rate show greater tendency to be essential for survival in rich media.*

4.2.3. Association with gene function

To study the variation of θs′ among different functional groups, we used the *Multifun* functional annotation of the *E. coli* genome (Serres and Riley, 2000), as it provides detailed manually curated annotation with experimental support for most genes. The annotation was downloaded from the *GenProtEC* database on February 2010 (*http://genprotec.mbl.edu/files/multifunassignments.txt*). In this hierarchical annotation system, a given gene can be assigned to multiple functions at each level of the hierarchy.

To identify functional groups with significantly different θs′ values, each θs′ value was assigned to all functional categories in which the gene was annotated. For categories with a minimum of five genes, a Wilcoxon rank-sum test was performed comparing the median θs′ of genes in the category against genes in all other categories.

Given the nature of this analysis, with genes belonging to several functions and with different categories having very different sample sizes, common approaches for assessing significance and for multiple testing may not be appropriate. To minimise the problem of variable sample sizes we used a double threshold approach to assess significance: (1) *P*-value must be lower than 0.01 (ensuring significant statistical effects) and (2) the deviation from the genomic median must be higher than 30% (ensuring significant size effects). To ensure that this approach has a low rate of false positives, we repeated the functional enrichment analysis with identical thresholds using 100 random permutations of the θs′ values. Using the combination of two thresholds described above in the randomly permuted datasets yielded a false discovery rate of 1.1 significant categories on average, with a standard deviation of 1.0. The two thresholds above were chosen after systematic evaluation of the effect of different combinations of thresholds, to provide a representative and reliable functional enrichment analysis.

Using this approach we found 21 categories with θs′ values significantly different from the genomic median, 8 cold (Table 4.1) and 13 hot (Table 4.2). It is important to note that, because of the different filters and controls, the final set of 2,659 genes is depleted in certain functions, such as those characterised by a very low synonymous diversity (*e.g.* ribosomal proteins) and those highly affected by gene loss and HGT.

P-value	*Median θs'*	*# genes*	*Functional annotation (Multifun ID)*
8.56E-06	0.0142	17	Oxidoreduction-driven Active Transporters (4.3.D)
7.49E-05	0.0142	13	Na+ transport (4.S.130)
0.000307	0.0145	14	H+ or Na+ -translocating NADH Family (4.3.D.1)
0.00641	0.0218	9	H+ transport (4.S.82)
0.00468	0.0218	15	Pilus (6.5)
0.00231	0.023	15	Protein transport (4.S.160)
0.000383	0.0235	38	Energy production/transport: Electron acceptor (1.4.2)
0.00737	0.0239	37	Energy production/transport: Electron donor (1.4.1)

Table 4.1. Cold functions. *Multifun functional categories showing significantly lower median values of θs' (using the final set of 2,659 genes).*

P-value	*Median θs'*	*# genes*	*Functional annotation (Multifun ID)*
0.000433	0.0439	38	ABC Superfamily + ABC-type Permeases (4.3.A.1.p)
0.00655	0.0444	16	Thiamine biosynthesis (1.5.3.8)
0.00832	0.0467	18	Catabolism of carbon sources other than carbohydrates, fatty acids, amino acids or amines (1.1.5)
0.00492	0.0488	8	Core region of the lipooligosaccharide (1.6.3.2)
0.000226	0.0488	27	Sulphur metabolism (1.8.2)
0.00862	0.0496	6	Anaerobic fatty acid oxidation pathway (1.1.2.4)
0.00171	0.0497	8	Ethanol degradation (1.1.5.2)
0.00616	0.0498	14	Isoleucine biosynthesis (1.5.1.19)
0.00218	0.0549	12	Leucine/valine biosynthesis (1.5.1.18)
0.00137	0.0566	6	Phenylalanine/ tyrosine transport (4.S.154)
0.00638	0.0576	6	Metabolism of other compounds: Sulphate assimilation (1.8.2.1)
0.00312	0.059	5	L-leucine/L-valine/L-iso-leucine transport (4.S.108)
0.00147	0.067	6	Histidine biosynthesis (1.5.1.16)

Table 4.2. Hot functions. *Multifun functional categories showing significantly higher median values of θs' (using the final set of 2,659 genes).*

This analysis reveals that mutationally cold genes generally encode for vital cellular functions such as processes related to central energy metabolism and the respiratory chain (Table 4.1). In contrast, mutationally hot genes are associated with metabolic pathways expressed at lower levels or used less frequently, such as amino-acid biosynthesis and catabolism of specific compounds.

The analyses above reveal that more essential genes under stronger purifying selection tend to show lower neutral mutation rates. This suggests that local mutation rates have indeed evolve to reduce the mutational burden in more essential genes.

In contrast, none of these analyses suggest a role for positive selection as a driving force for higher local mutation rates in genes under more frequent positive selection. Indeed, hot genes mainly represent dispensable functions in many conditions, rather than genes expected to experience frequent positive selection. Although certain antigenic genes (*e.g.* O-Antigen of the lipooligosaccharide) are depleted from the set of 2,659 genes, remaining genes expected to undergo more frequent positive selection do not appear as hot mutational functions. This contrasts with the very few previously known examples of functional variation of the local mutation rates, namely hypervariable repeats in bacteria and somatic hypermutation in the vertebrate immune system, which evolved in response to repeated local positive selection.

To shed more light on this question, in the next section we describe additional analyses on the role of positive selection in the variation of the neutral point mutation rate in *E. coli*.

To summarise the results so far, we were unable to detect any association between θs′ and the density of sites under positive selection. Thus, our observations suggest that purifying, rather than positive, selection has driven the evolution of the mutation rate along the core *E. coli* genome.

4.2.4. Association of the mutation rate with positive selection

A commonly cited evidence in support of the action of positive selection on a gene is a dN/dS value higher than 1. However, since purifying selection generally acts in most sites and more frequently than positive selection, dN/dS ratios for entire genes tend to be smaller than 1. In fact, in our dataset of 2,659 alignments the median pN/pS (ω_p) at gene resolution was 0.072, with a maximum value of 0.87 (estimated using CODEML, Section 2.2.4). Similar estimates were obtained using omegaMap (average ω 0.078, with a maximum value of 0.75).

An alternative, much more powerful method to identify positive selection is to estimate ω at codon resolution (Massingham and Goldman, 2005, Wilson and McVean, 2006). This test can be applied here despite having shown that θs is not a purely neutral reference, because purifying selection on synonymous sites is very weak in comparison with selection on protein sequences. Alternatively, McDonald-Kreitman's population genetic test can also be used to detect positive selection, but its application to bacterial sequences can be problematic (Hughes et al., 2008).

Therefore, to estimate the strength of positive selection on each protein along the radiation of *E. coli* strains, we used a site-wise Bayesian test based on a coalescent model that accounts for the effect of recombination (implemented in the omegaMap software) (Wilson and McVean, 2006). This method was designed for and tested with polymorphism data in bacteria (*Neisseria meningitidis*) and it was found to detect positively selected sites with good sensitivity and a low false positive rate.

Bayesian site-wise estimation of ω_p in a coalescent model

For each of the 2,659 alignments, the codon frequencies were inferred from the codon usage of the reference genome (*E. coli* K12 MG1655) and we ran a single MCMC chain of 150,000 iterations, which we sampled every 150 iterations after a burn-in of 20,000 iterations. The parameters used were as follows: *muPrior* = improper_inverse; *kappaPrior* = improper_inverse; *indelPrior* = improper_inverse; *omegaPrior* = inverse; *omegaParam* = 0.01, 100; *rhoPrior* = inverse; *rhoParam* = 0.01, 100; *muStart* = 0.1; *kappaStart* = 3.0; *indelStart* = 0.1; *omega_model* = variable; *oBlock* = 15; *rho_model* = variable; *rBlock* = 15. These parameters are identical to those used in the estimation of θs.

The output file was read using the "*summarize*" function. For each gene we calculated the fraction of positively selected amino acids, where an amino acid was considered to be under positive selection when the posterior probability for ω_{site}>1 was higher than 80%.

Results

Figure 4.2 shows a scatter plot relating the *fraction of positively selected sites* with θs′ at gene level. The plot shows two important features suggesting that positive selection is not very relevant to the evolution of local point mutation rates. First, only 2.9% of genes have one or more amino acids with evidence of positive selection. This, together with the very low ω_p values, indicate that purifying selection is much more frequent than positive selection in *E. coli*. Second, we found no positive association between the fraction or the absolute number of sites under positive selection and θs (both when including and excluding the bulk of genes with no amino acids under positive selection). Analogous results were obtained with θs (before bias correction). This suggests that genes under more frequent positive selection do not show higher neutral mutation rates (as would be expected under a bet-hedging model for the evolution of the point mutation rate).

This analysis, together with the functional associations of θs′, suggests that purifying, rather than positive selection, has driven the evolution of the local mutation rate along the *E. coli* genome. However, it should be noted that this analysis is restricted to the 2,659 core or near-core genes that passed our strict filters. Consequently, our analysis likely provides a conservative representation of the variation in the local mutation rate along the *E. coli* genome and the possibility of the evolution of mutation rates towards increasing mutability cannot be ruled out for genes excluded from this analysis.

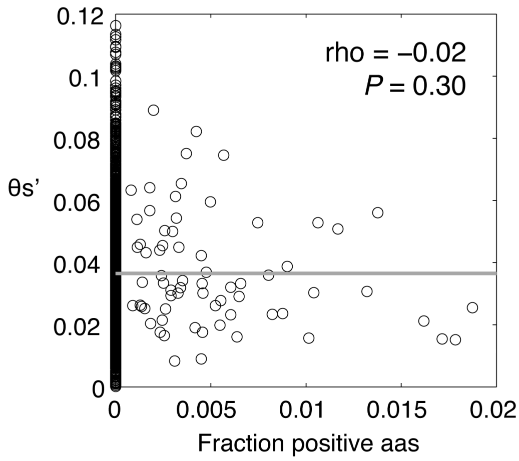

Figure 4.2*. Scatter plot showing the association between positive selection and θs. Positive selection is represented as the fraction of amino acids for which the posterior probability of ω_{site}>1 is higher than 80% according to omegaMap.*

4.2.5. Association of the mutation rate with gene expression

Having studied the association of θs′ with gene function, essentiality and with the strength of purifying and positive selection on the protein sequence, we then evaluated its relationship with gene expression. To do so, we used transcriptomic (RNA-sequencing) and RNA-polymerase-binding data (ChIP-sequencing).

These datasets were produced by our lab (Kahramanoglou et al., 2011). The datasets of RNA-sequencing were obtained for *E. coli* K12 MG1655 in rich media (LB) at two different time-points: early-exponential phase (EE) and transition-to-stationary phase (TS). The ChIP-sequencing dataset of the RNA polymerase binding was obtained in mid-exponential phase[15].

In the original publication (Kahramanoglou et al., 2011), reads were mapped to both strands of the *E. coli* K12 MG1655 genome using BLAT (Kent, 2002) allowing no gaps and up to two mismatches. Only reads that uniquely mapped to single regions of the genome were considered for further analysis. For each base position on the genome, the number of reads that mapped to that position was recorded. As an estimate of the mRNA level of each gene and the RNA polymerase occupancy along each gene body, we used the mean coverage along the entire gene. Alternatively, using the median yielded analogous results.

All results shown in the main text correspond to the early-exponential phase RNA-sequencing dataset, but similar results were obtained for the transition-to-stationary phase dataset.

Using these datasets, we found a clear negative correlation between θs′ and transcription levels (Figure 4.3), with the coldest genes (bottom 5% θs′) showing nearly four-fold higher transcription than the hottest genes (top 5% θs′).

[15] The raw read files, the processed mapping files and a full description of the protocols used are available in ArrayExpress under the accession number E-MTAB-387.

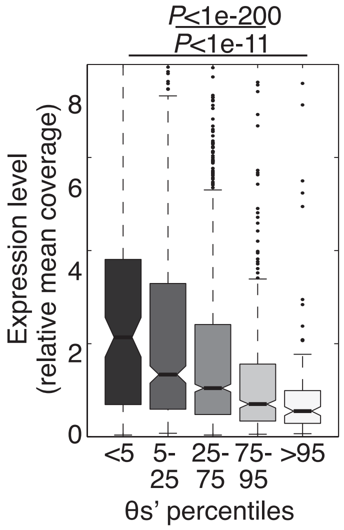

***Figure 4.3**. Boxplot displaying the relationship between the neutral mutation rate and the transcription level. Genes with lower mutation rate are generally more highly expressed (using RNA-Seq data from E. coli K12 in rich media and early-exponential phase). A Spearman's correlation test yields rho=0.22, P<10^{-29}.*

The association between the local mutation rate and the level of gene expression is a subject of special interest. A negative association between inter-species synonymous divergence (dS) and expression levels has been repeatedly reported previously, but it was generally accompanied by a correlation with CUB, and so it was typically interpreted as a consequence of the action of selection at synonymous sites (Drummond and Wilke, 2008). Interestingly, however, other studies showed that the negative association remained even after accounting for codon-usage bias (Eyre-Walker and Bulmer, 1995, Berg and Martelius, 1995) and interpreted the association as evidence for the variation of the neutral mutation rate. However, these observations were often dismissed as evidence of selection on synonymous sites by other causes (Ochman, 2003).

Having ruled out the role of CUB and unknown selection forces on the gene-to-gene variation of θs', our observation (Figure 4.3) indicates that the neutral mutation rate is indeed lower at highly expressed expressed genes in *E. coli*.

This conclusion is highly unexpected given that transcription is known to be mutagenic. Several experimental studies (Ochman, 2003, Beletskii and Bhagwat, 1996, Klapacz and Bhagwat, 2002) using gene reporters have shown that increasing the expression level of a gene also increases its mutation rate, which is consistent with the fact that, during transcription, DNA is transiently single-stranded and so more vulnerable to mutagenesis by deamination and dimerisation (Beletskii and Bhagwat, 1996).

Thus, if transcription is mutagenic, our observations imply that compensatory mechanisms must exist that preferentially protect or repair highly expressed regions of the genome. Indeed we do detect the action of the transcription-coupled repair (TCR) pathway (Francino et al., 1996) (Figure 4.4); however this process alone does not account for the decay in mutation rate with expression because the non-transcribed strand (in which TCR does not act) also displays a negative dependence between expression and mutation rate (Figure 4.4) (Pleasance et al., 2010). Moreover, single-gene experiments have demonstrated that transcription-induced mutagenesis occurs both in the presence and absence of TCR (Klapacz and Bhagwat, 2005). Therefore, the observed compensation must be mediated by mechanisms that generally target highly expressed genes, but are not directly coupled with the transcriptional process.

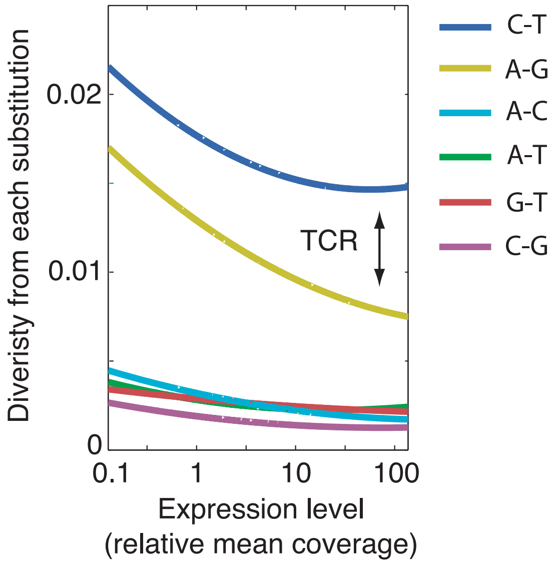

***Figure 4.4**. Fitted lines showing the relationship between $\theta s'$ per substitution type and expression level. The y-axis represents the average $\theta s'$ at each expression level multiplied by the fraction of synonymous substitutions corresponding to each substitution pair. The direction of the substitution is not shown as we used a reversible evolutionary model. Since the C-to-T transition (G-to-A in the opposite strand) is the most common mutation, the gap between the C-T (blue) and G-A (yellow) lines indicates the strand asymmetry caused by the action of transcription-coupled repair (TCR) on the transcribed strand.*

4.2.6. Additional controls for the robustness of the functional associations

In addition to the extensive and strict controls described in Chapter 3, here we show a few extra tests examining the impact of selection and recombination biases on the functional associations described above.

4.2.6.1. Selection

Section 3.3.1.4 and Figure 3.5 described a cross-validation test demonstrating the increasing ability of LOWESS regression to remove selection bias when increasing the minimum number of polymorphisms used to calculate TDsyn values.

Using the same cross-validation test, we can evaluate whether increasing removal of selection bias affects the functional associations described in the previous section. In particular, we focused on the correlations of θs′ with selection on the protein sequence and gene expression, as measured by RNA-sequencing (early log-phase) and RNA polymerase binding (Figure 4.5).

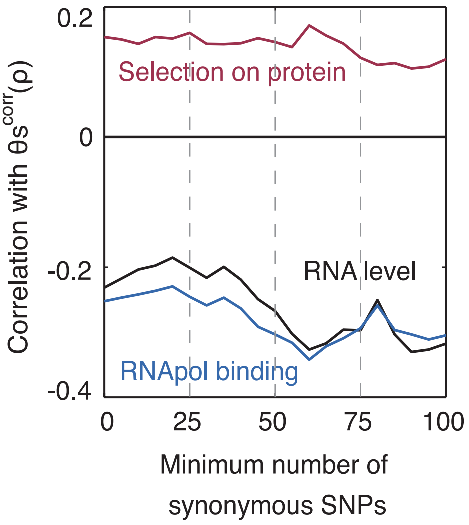

Figure 4.5. *Strength of the functional associations of θs values at different levels of selection bias correction. Complementing Figure 3.5, this figure shows that the functional associations of θs′ are unaffected by the removal of over 90% of the selection bias (as measured by TDsyn).*

The cross-validation test reveals that the strength of the functional associations of θs′ remains largely unchanged with increasing removal of selection bias. This result is consistent with the very weak selection bias observed in θs.

4.2.6.2. Recombination

As explained before, a higher local frequency of homologous recombination tends to increase the local synonymous diversity (θs). Since this directional effect was corrected by LOWESS regression, θs′ should be largely free of the recombination bias. Thus, the observed functional associations should also be independent of the impact of recombination.

Nevertheless, as an alternative control, we evaluated the correlation of θs with mRNA levels only in genes that are not affected by recombination or showing only a weak evidence of recombination. To do so, we binned our dataset of 2,659 alignments into five categories according to the normalised |D′| or to the fraction of SNPs consistent with the reference genealogy. For each bin, we calculated the correlation coefficient of θs vs RNA-Seq levels. Figure 4.6 shows the results of this analysis, and clearly demonstrates that the correlation of θs with mRNA levels is independent of the effect of recombination.

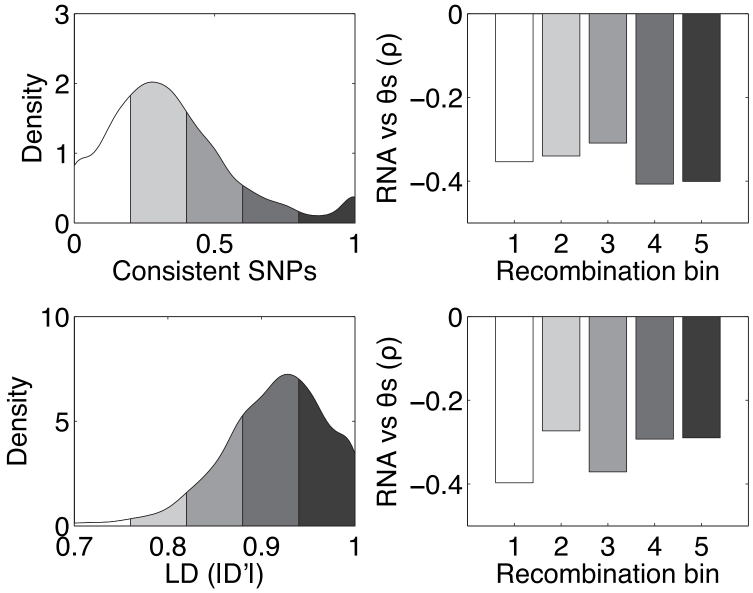

***Figure 4.6**. Impact of recombination on the correlation of θs and mRNA level as described in the text. The top two panels were obtained using the fraction of SNPs consistent with the reference phylogeny and the bottom two panels using the subsampled measure of linkage disequilibrium. On the left we show the frequency of genes for each value of recombination, and the five colours represent the five bins used (the darker the bin the lower the recombination frequency). On the right, we show the Spearman's correlation coefficient obtained between θs and RNA-Seq levels using only genes within a given bin. The figure evidences that very similar associations are observed between θs and transcription independently of the level of recombination.*

Using bins is particularly useful in the case of homologous recombination since, while low values of linkage or phylogenetic consistency could have different effects on θs at different genes, genes with high values (black bins in Figure 4.6) are largely free from recombination.

These analyses show that, consistently with the correction by LOWESS regression, recombination does not affect the observed anti-correlation of the mutation rate with transcription levels.

Finally, we also performed a functional enrichment analysis (see Section 4.2.3) of both our subsampled |D′| values and of the fraction of consistent SNPs. Both measures of homologous recombination showed none or very weak enrichments.

These analyses strongly support the independence of the functional associations from any bias associated with homologous recombination in *E. coli*.

4.2.7. Independence of the different functional associations of θs′

In this study we show that the neutral mutation rate is lower in highly expressed genes (Figure 4.3), essential genes (Figure 4.1b), genes that encode proteins under stronger purifying selection (Figure 4.1a) and genes that encode different components of the central energy metabolism (Table 4.1). In contrast, the neutral mutation rate is higher in lowly expressed genes and genes encoding for metabolic functions not needed under growth in rich media (Table 4.2). However, many of these observations are not independent. For example, essential proteins are generally highly expressed and under stronger purifying selection. Moreover, there are strong overlaps among significant hot and cold gene functions (Tables 4.1 and 4.2) and these are also associated with expression. Thus, it is unclear whether the functional associations observed could be explained by a few factors or even by a single dominant factor such as expression.

To evaluate the importance of each of these associations in the context of all the factors, we used a multiple linear regression, with all significant functional effects as predictor variables and θs′ as the predicted variable. We used the unsupervised stepwise method implemented in the *stepwisefit.m* function of the Statistics Toolbox of Matlab (version R14).

The predictor variables used are: transcript level in early-log phase (RNA-Seq), transcript level in late-log phase (RNA-Seq), RNA polymerase binding in mid-exponential phase (RPO-binding), purifying selection at the protein level (TDnon), essentiality in rich media (binary variable), and the 21 significantly hot or cold *Multifun* terms (Tables 4.1 and 4.2) (as binary variables). In total we used 26 starting predictor variables.

The resulting smallest model that contains most of the information from all the predictor variables is shown in Table 4.3.

Predictor	Coefficient (B)	SE	t	p-value
RPO binding	-0.00661	0.000852	-7.77	1.32E-14
TDnon	0.00802	0.00131	6.11	1.18E-09
4.3.D	-0.0242	0.00732	-3.3	0.000972
6.5	-0.0216	0.00505	-4.28	1.95E-05
1.4.1	0.00971	0.00455	2.13	0.033
4.3.A.1.p	0.00675	0.00308	2.19	0.0286
1.8.2	0.0105	0.00369	2.85	0.00441
1.1.2.4	0.0212	0.00781	2.71	0.00673
1.5.1.18	0.0209	0.00714	2.93	0.00348
1.5.1.16	0.0397	0.00873	4.55	5.83E-06

***Table 4.3**. Multiple linear regression identifies a reduced set of functional variables associated with θs′.*

The multiple regression yielded several interesting results:

1. RPO-binding alone (mid-exponential phase) is sufficient to explain the association between transcription and θs′, with the two RNA-Seq variables (from early- and late- exponential growth phase) strongly correlating with θs′ but not providing any additional explanatory power. This observation supports the precision of the gene expression datasets and the power of the multiple linear regression approach to remove redundant associations in the conditions of this study.

2. Purifying selection at the protein level (TDnon) is associated with θs′ even accounting for the effect of expression (the first variable introduced in the model). This is particularly remarkable given the noise in the calculation of Tajima's D due to the few non-synonymous polymorphisms per alignment.

3. The association of essentiality with θs′, however, can be explained by expression or even by TDnon alone. This is not surprising as essential genes are characterised by very high expression levels and are under stronger purifying selection.

4. Multiple functional categories from *Multifun* still appear as significantly hot or cold after accounting for the effects of expression and the strength of selection on the protein sequence. In addition, the multiple regression analysis accounts for the redundancy in the list of significantly hot and cold gene functions. The resulting list of non-redundant functions is a good representation of the original (redundant) list. According to this analysis, non-redundant cold functions are *oxidoreduction-driven active transporters* (central energy metabolism) and *pilus*. Non-redundant hot functions include *ABC superfamily*, *sulphur metabolism*, *anaerobic fatty acid oxidation pathway*, *leucine/valine biosynthesis* and *histidine biosynthesis*.

To evaluate further the impact of transcription levels on the other functional associations, we repeated the unsupervised stepwise procedure excluding the three expression-related variables (RPO-binding and the two RNA-Seq variables). Figure 4.7 shows a scatter plot of the p-values of the 23 remaining variables in the regression models including and excluding expression. It is clear that, although expression is a very important predictor, it has little impact on the association of the other functional variables with θs′.

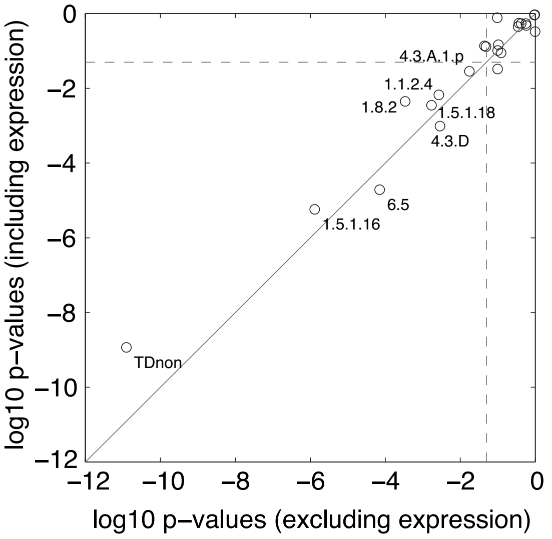

***Figure 4.7**. Scatter plot of the p-values of 23 predictor functional variables from a multiple regression model that includes a variable for expression (RPO-binding) and from a model that excludes variables measuring expression.*

4.2.8. Conclusions on the non-random variation of the mutation rate

All together, this study provides interesting and unexpected evidence for the non-random variation of the local mutation rate in *E. coli*. Some of the main conclusions from Chapters 2-4 are:

1. Synonymous diversity varies by more than 20-fold among genes in the *E. coli* genome; however, selective and non-selective factors explain only a small proportion of this variation. Instead, our observations suggest that the variation in synonymous diversity results from large heterogeneity in the underlying neutral mutation rate.

2. Contrary to the commonly accepted evolutionary tenet, we observe that the point mutation rate does not randomly vary across the genome. Instead, the data suggest that genes under stronger purifying selection - *i.e.* those for which spontaneous mutations are most likely to be more deleterious - display a lower mutation rate.

In Section 4.3 we explore the role of some factors known to affect the local mutation rate and in Section 4.4 we propose an evolutionary model that explains how non-random mutation rates can indeed evolve from random mutations and selection, to reduce the risk of deleterious mutations. Finally, a discussion and some preliminary results on the evolution of the local mutation rate in the human genome is presented at the end of Section 4.4.

4.3. Role of known sources of mutation rate heterogeneity

We can only speculate about the molecular mechanisms underlying the localised reduction in spontaneous mutations. DNA-binding proteins and DNA repair pathways are obvious candidates, especially as there is increasing evidence for locus-dependent control of the latter (Tu et al., 1996, Hoege et al., 2002).

Here we explore the role of some known sources of mutation rate heterogeneity in the *E. coli* dataset.

4.3.1. Context-dependent mutations

Several mutagenic mechanisms are influenced by the sequence flanking the affected base (Hodgkinson and Eyre-Walker, 2011). For example, the methylation of CpG motifs in mammals often leads to C-to-T transitions and even to an increased transversion rate of the guanine (G:C-to-T:A and G:C-to-C:G) (Ketterling et al., 1994). Another example is the formation of mutagenic pyrimidine dimers in CT and TC dinucleotides due to UV irradiation; these lesions often induce C-to-T transitions because of the tendency of translesion repair to introduce adenines opposite to the lesion (Choi et al., 2006).

To test the importance of context-dependent mutagenesis in our data, we studied the sequence enrichment surrounding each SNP. Since the currently known dependencies are generally specific to particular substitutions, we classified all biallelic synonymous substitutions into the six possible nucleotide pairs: A-C, A-G, A-T, C-G, C-T and G-T. For consistency with the use of unrooted trees and time-reversible substitution matrices, we did not distinguish the direction of these substitutions; however the classification should be sensitive enough to detect relevant biases. We then annotated the surrounding sequence of each SNP and calculated sequence enrichments using entropy (Schneider and Stephens, 1990).

The sequence logos (Figure 4.8) show modest enrichments for some substitutions, especially at positions -1 and -2 relative to the SNP. However, these enrichments can be explained by the constraints of the genetic code, and simulated random synonymous mutations generated using a Jukes-Cantor substitution matrix yield largely similar enrichments.

We also performed a statistical analysis of the enrichment in neighbouring nucleotides with an algorithm that uses a negative set of sequences (Vacic et al., 2006). Again, we observed only small differences in the two sets of sequences, with sequence enrichments and depletions that never deviate more than 15% from expected.

In summary, this analysis demonstrates that the vast majority of substitutions are context-independent in our dataset of *E. coli* polymorphisms. Only very small sequence enrichments, some possibly artefactual, were observed around specific types of synonymous substitutions. However, any interpretation of these small biases should consider a more realistic substitution matrix in the generation of a reliable set of control sequences, which is beyond the scope of this study.

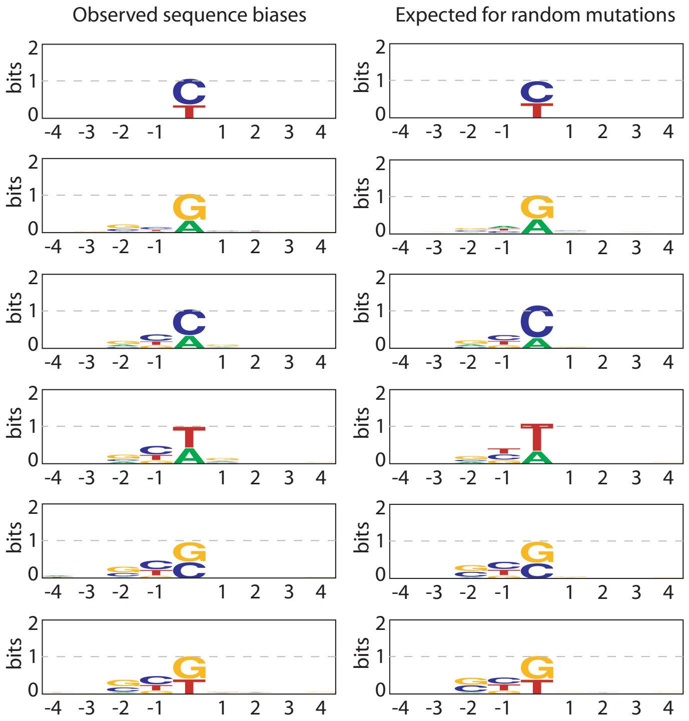

***Figure 4.8**. Sequence logos of enriched nucleotides around synonymous SNPs. The first column shows the enrichments observed around real synonymous SNPs and the second column shows the enrichments expected for randomly placed synonymous mutations in E. coli K12 gene sequences. The rows correspond to each of the six possible pairs of nucleotides: C-T, G-A, C-A, T-A, G-C and G-T. Logos adapted from the representation obtained from WebLogo (Crooks et al., 2004).*

4.3.2. Additional sequence composition effects

Sequence properties other than GC content could have an effect in the intrinsic mutability of a given base (Hoede et al., 2006). Here we explore the association of θs' with the frequency of dinucleotides and trinucleotides as well as with the formation of single-stranded DNA secondary structures during transcription.

4.3.2.1. Dinucleotides and trinucleotides frequencies

For each of the 2,659 genes, we calculated the frequency of each dinucleotide (16 in total) in the transcribed strand of the *E. coli* K12 MG1655 genome. We then used 16 simple linear regressions to study the contribution of each of these factors to the variance of θs′ separately and a multiple linear regression model to study the overall contribution of all dinucleotides. We found that the frequency of some dinucleotides correlates very weakly with θs′ and that all together they could explain up to 4.6% of the θs′ variance. Repeating this analysis for the 64 trinucleotides with a stepwise linear model shows that they explain around 7% of the θs′ variance.

It is unclear whether these weak associations reflect a direct effect of the sequence composition on the intrinsic mutability of the sequences or an indirect association caused by the action of other factors. In any case, these analyses indicate that the frequencies of dinucleotides and trinucleotides have very minor effects in the variation of θs′.

4.3.2.2. Single-stranded DNA and transcriptional mutagenesis

During transcription, the DNA is transiently single-stranded and therefore more vulnerable to mutagenesis. It has been suggested that transcription-induced mutagenesis can be reduced by the formation of local intra-strand secondary structures, and that selection may have favoured the formation of single-stranded DNA secondary structures in highly transcribed *E. coli* genes to minimise the impact of transcriptional mutagenesis (Hoede et al., 2006). In their paper, Hoede and colleagues showed that the observed distribution of codons in gene sequences might promote a higher frequency of local intra-strand secondary structure than expected if codons were distributed randomly. They noted that this deviation from random is higher in more highly expressed genes, suggesting the action of selection to reduce transcription-induced mutagenesis. This very interesting observation is consistent with the conclusions and the model of our study. Thus, we studied the impact of this mechanism in the genome-wide gene-to-gene variation of θs.

Hoede *et al.* defined a transcription-driven mutability index (TDMI) to reflect the overall mutability of a gene during transcription. To evaluate the effect of TDMI in the variation of θs, we used an analogous approach: (i) For each gene all 30-bp subsequences were folded using the program *hybrid-ss-min* from the UNAfold package (Markham and Zuker, 2005) (version 3.7, -n=DNA, t = 37, [Na+] = 1, [Mg++] = 0, maxloop = 30, prefilter = 2/2). (ii) For each sub-sequence we recorded the free energy (Z =

$e^{(-\Delta G/RT)}$) of the most stable structure and the paired/unpaired state of each base. (iii) For each base, the TDMI was calculated as the ratio of the sum of Z values for all subsequences that include a given base in a paired state over the sum of Z for all subsequences that include this base, independently of its paired/unpaired state. This provides a measure for the tendency of a base to be single-stranded, ranging from 0 (always paired) to 1 (always unpaired). (iv) Finally, we calculated the average TDMI for all bases of each gene as a measure of its transcription-driven mutability (Hoede et al., 2006).

To evaluate the impact of DNA secondary structures on the variation of θs, we explored the association of TDMI with θs and θs′. TDMI correlated positively weakly but significantly with θs (Spearman's rho=0.090, *P*<1e-5), but it showed no correlation (or at most a very weak negative correlation) with θs′ (Spearman's rho=-0.045, *P*=0.028). The weak positive association with θs can indeed be explained by an indirect association of both θs and TDMI with GC content, since correction of GC-bias on θs is enough to remove any association between θs and TDMI (Spearman's rho=-0.019, *P*=0.34).

This suggests that the formation of secondary structures during transcription does not contribute to the gene-to-gene variation of θs, and so this mechanism does not appear to be relevant for the observations reported here.

4.4. A risk management model for the evolution of local mutation rates by preferential protection

> *Natural selection, the blind, unconscious, automatic process which Darwin discovered, and which we now know is the explanation for the existence and apparently purposeful form of all life, has no purpose in mind. It has no mind and no mind's eye. It does not plan for the future. It has no vision, no foresight, no sight at all. If it can be said to play the role of watchmaker in nature, it is the blind watchmaker.*
>
> — Richard Dawkins, *The Blind Watchmaker* (1986)

> *Biologists must constantly keep in mind that what they see was not designed, but rather evolved.*
>
> — Francis Crick, *What Mad Pursuit* (1990)

4.4.1. Evolution of the local mutation rate

According to the observations in this chapter, the mutation rate appears to vary significantly among genes along the *E. coli* genome, with functionally more important genes experiencing lower neutral mutation rates. This suggests that purifying selection has driven the evolution of local mutation rates to reduce the rate of spontaneous deleterious mutations.

While extensive theoretical work has been devoted to the evolution of global mutation rates, comparably very few studies have tackled the evolutionary tuning of local mutation rates along a genome. Nevertheless, assuming that the mutation rate can vary along the genome and that this variation is heritable, much of the theory for global mutation rates also applies to local mutation rates.

With the exception of the few loci subject to repeated positive selection, such as those related to immune evasion in pathogens and immune response in hosts, positive selection is most likely too rare and unpredictable to drive the evolution of the local mutation rate at most loci. Further, bacteria can quickly generate local diversity by horizontal gene transfer and so the pressure to evolve local hypermutation is likely to be very relaxed.

In contrast, purifying selection acts on most spontaneous mutations in bacteria and it is relatively predictable, since certain genes show consistently stronger purifying selection over very long evolutionary periods (Figure 4.9).

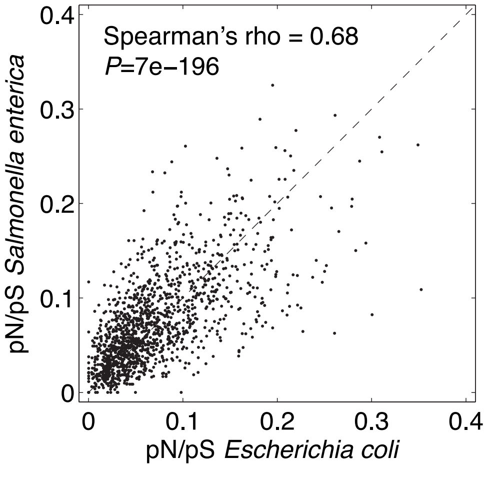

Figure 4.9. *Scatter plot showing the correlation of within-species pN/pS values obtained for orthologs of Escherichia coli (34 genomes) and Salmonella enterica (14 genomes). The species are believed to have separated around 100 My ago (Lawrence and Ochman, 1998) but they show remarkable similarity in the selection strengths of orthologous proteins.*

Since the cost of deleterious mutations varies remarkably among genes, while the cost of fidelity and the limit set by genetic drift affect all sites equally, the optimal and the lowest evolvable mutation rates are also expected to vary along the genome.

4.4.2. A risk management model for the evolution of local mutation rates

Here, the general model of evolutionary risk management developed by Wagner (Wagner, 2003) is used to explain the evolution of the local mutation rate by purifying selection. According to this simple model, selection against strongly deleterious mutations could lead to the preferential protection of the most important loci along a genome.

Wagner classified risks into those that affect all individuals of a given genotype simultaneously and those that only cause individual deaths. An environmental change is an example of the former while a spontaneous deleterious mutation is an example of the latter. To model the effect of risks causing individual deaths, Wagner considered two genotypes G_0 and G_r. G_r is the superior risk manager (here a genotype with a reduced frequency of lethal mutations). Their absolute frequencies are N_0 and N_r, respectively. Using a Malthusian model, the population growth rate of G_r can be defined as:

$$r_r = b_r - (d + d_r)$$

where b_r is the per capita birth rate, $d + d_r$ is the per capita death rate composed of d (the rate of death from other causes different from mutation) and d_r (deaths from lethal mutations). G_r is a superior risk manager if $d_r < d_0$. If the risk management strategy has a cost, then $b_r \leq b_0$. G_r will eventually dominate the population as long as:

$$r_r > r_0$$

or

$$b_0 - b_r < d_0 - d_r$$

If $N_r r_r - N_0 r_0 > 1$, then selection will effectively favour the fixation of the risk manager allele, otherwise drift will dominate its evolution.

Bacterial populations are generally large (Maynard-Smith, 1991, Guarner and Malagelada, 2003, Marteau et al., 2001, Charlesworth, 2009), with some studies estimating that the census population of *E. coli* in a typical human gut is around 10^{11}-10^{12} (Marteau et al., 2001, Guarner and Malagelada, 2003). The estimation of the effective population size (N_e) in *E. coli* is however more challenging, with different studies using different approaches with simplistic assumptions and yielding substantially different results. However, most estimates suggest that N_e is on the order of 10^7 (Charlesworth, 2009) or 10^8 (Hartl et al., 1994). This means that alleles with the capacity to reduce the rate of deleterious mutations even by very small factors (10^{-7} or even lower) could increase their frequency in the population effectively. For example, the genomic mutation rate in *E. coli* is roughly 0.0025 mutations per genome per generation (Drake et al., 1998), or around $5x10^{-7}$ per gene per generation. In that case, an allele able to reduce the mutation rate of an essential gene by as little as 10% would effectively provide a fitness advantage on the order of 10^{-7} or 10^{-8}. Modifier alleles with stronger effects and/or acting on larger genomic segments under similar selection pressure could provide much larger fitness advantage.

The model above is simplistic as it treats all mutations as lethal. Moderately and weakly deleterious mutations will have a lower immediate impact but will tend to accumulate and may cause a substantial fitness loss in the longer term (see Appendix 1 for a detailed evaluation of the fitness cost of mutations with different selection coefficients). Thus, despite of its simplicity, the model above provides an introduction to the power of selection and the limit of drift in the evolution of local mutation rates. It seems clear that, contrary to global mutation rates in microbes (Appendix 1), the evolutionary tuning of local mutation rates may be closely limited by genetic drift.

4.4.2.1. Mechanisms for the local control of mutation rates

At least two major types of mechanisms could allow bacteria to evolutionarily modulate their local point mutation rate.

First, certain sequence properties might make a given sequence less vulnerable to mutagenesis. An example would be the formation of more stable secondary structures in the non-transcribed strand of highly expressed genes (Hoede et al., 2006). This type of strategies could evolve if a single allele can affect the mutability of a segment of

DNA sufficiently. The evolution of these mechanisms is limited by genetic drift (Lynch, 2010), according to the strength of the anti-mutator alleles (the product between the reduction in the rate of strongly deleterious mutations per site and generation and the size of the fragment affected).

A second group of mechanisms includes those linked to protective or repair mechanisms. It is often believed that the observed investment on protective/repair factors is the result of a tradeoff between the cost of fidelity and the fitness cost of deleterious mutations. Since protective or repair proteins are limited, a possible risk management strategy could be to preferentially target protective or repair factors to regions with a higher risk of strongly deleterious mutations. Let us consider a hypothetical genome in which the action of protective or repair factors is reduced in a single more dispensable gene (where the risk of a strongly deleterious spontaneous mutation is much smaller) and increased in an essential gene (with higher risk). In such a case, the probability of strongly deleterious spontaneous mutations could be easily reduced by an amount on the order of 10^{-7} or 10^{-8}, without increasing the genome-wide investment in protection/ repair. The evolution of this second group of strategies is limited by the cost of the reduction of fidelity in other parts of the genome and by genetic drift. This suggests that a genome with a non-random distribution of protective or repair factors could evolve by the recurrent differential action of purifying selection. With a lower investment in protective factors, such a genome could reach the same rate of deleterious mutations as genomes in which protection or repair is uniform.

4.4.2.2. Increase in adaptability by cryptic variation

An interesting side effect of the risk management model is that, despite not being driven by positive selection, it could also facilitate adaptation in the event of an environmental change. Since the rate of deleterious mutations can be reduced efficiently by preferentially protecting more essential regions, more dispensable parts of the genome would tend to accumulate more diversity of little functional impact, which would increase the probability of adaptation to a new environment (Hayden et al., 2011). Effectively, targeting protective or repair factors to more strongly negatively selected regions could reduce the risk of deleterious mutations with a lower metabolic cost, while maintaining or even increasing the rate of adaptive mutations.

4.4.3. Evolution of local mutation rates by multilocus-modifier alleles in humans

The above explanations suggest that purifying selection can be strong enough to drive the evolutionary optimisation of the local mutation rate along a genome in species with large effective population sizes. Although genetic drift establishes a strict limit to the evolvability of local mutation rates, this strategy could also operate in much smaller populations, particularly if the deleterious genomic mutation rate is higher and/or if single local anti-mutator alleles are able to provide larger reductions in the risk of deleterious mutations. For instance, this could be achieved by affecting larger regions with similar needs for mutation, such as bacterial operons or gene-rich regions even in genomes with lower densities of coding sequences.

The question of whether evolution has optimised to some extent the local mutation rate in humans remains unanswered (Hodgkinson and Eyre-Walker, 2011). Nevertheless, as noted by Hodgkinson and Eyre-Walker two major factors make local modifier alleles very unlikely to provide any significant fitness advantage: first, the effective population sizes in mammals are on the order of 10^4; and second, the much lower density of functional elements in the genome further reduces the advantage of a local reduction of the mutation rate. Given these constrains it has been estimated that a modifier allele would need to reduce the mutation rate by more than an order of magnitude in a region of several megabases to be effectively selected (Hodgkinson and Eyre-Walker, 2011). Since the extent of this variation is much larger than any variation observed in the mutation rate along mammalian genomes, Hodgkinson and Eyre-Walker suggested that the variation of the local mutation rate in mammals is very unlikely to provide any evolutionary advantage.

The above is true for strictly local modifier alleles, whose action is limited to a linear segment of the genome. However, an unexplored possibility supported by the risk management model is the evolution of single modifier alleles affecting multiple loci of similar functional importance simultaneously (which we call *multilocus-modifier alleles*). For example, a single modifier allele coupling a repair pathway to the transcriptional machinery could provide significant fitness benefit as it would reduce the mutational burden preferentially in a large number of highly expressed genes. Similarly, given that certain epigenetic marks typically characterise transcriptionally active or even housekeeping genes (Filion et al., 2010, Roy et al., 2010, Gerstein et al., 2010, Ernst et al., 2011), single mutations coupling repair to certain chromatin remodelling factors could as well provide substantial fitness advantages. Interestingly, increasing experimental evidence supports the existence of several links between epigenetic marks and DNA repair (Tu et al., 1996, Hoege et al., 2002).

In theory, multilocus-modifier alleles could be effectively subject to selection in relatively small populations and potentially lead to the evolution of non-random mutation rates even in humans. This possibility, however, remains to be carefully investigated.

The imminent availability of large amounts of functional genomics and population genomics data for humans, from projects such as ENCODE (Myers et al., 2011) and 1,000 genomes (The 1000 Genomes Project Consortium, 2010), are likely to prove invaluable to tackle this question. With the mere goal of illustrating this, we can perform a very preliminary analysis of the association of chromatin remodelling factors and transcription factor binding with putatively neutral sequence variation.

A good starting point to study neutral sequence variation in humans, is using synonymous divergence between human and chimpanzee (*Pan troglodytes*) orthologs. This data is suitable for a preliminary analysis as it is naturally free from some of the main problems associated with the *E. coli* data (Chapters 2 and 3). First, owing to the very small population sizes in primates (on the order of 10^4 according to Charlesworth, 2009), synonymous substitution rates can be considered largely neutral. Second, owing to the lack of interspecies gene transfer, the identification of vertically divergent ortholog pairs is much simpler. Third, contrary to polymorphisms data, indirect selection (background selection and hitchhiking) cannot possibly affect substitution rates. Some limitations still remain, however, such as the role of CpG-methylation, recombination-associated biases, sequence composition biases and, unavoidably, the shadow of weak selection (Chamary et al., 2006). Thus, any result based on synonymous divergence as a neutral estimate of the local mutation rate should be considered preliminary and subject to further evaluation.

Regarding the functional genomics data, at present the ENCODE consortium has produced and processed 685 ChIP-Sequencing datasets (matched to input controls) characterising the genome-wide distribution of over a hundred transcription factors and chromatin-remodelling marks in several cell types (Myers et al., 2011). Some of the best studied transcription factors include: *c-Fos, c-Jun, c-Myc, CTCF, ERG1, GABP, GATA1-3, IRF1,3-4, JUND, MAFK, MAX, NFKB, P300, RAD21, STAT1-3, TAF1, USF1, YY1* and *ZNF274*. The list also includes RNA-polymerase II in multiple conditions as well as some of its better known modifications. Finally, the consortium also studied a comprehensive collection of chromatin marks, including most major histone methylations and acetylations.

These data allows a very preliminary exploration of the association of synonymous divergence with the local functional state of the DNA. In order to do so, $\text{dS}_{\text{human-chimp}}$ values were obtained from Ensembl BioMart (Kinsella et al., 2011) and the binding sites of each factor were kindly provided by the ENCODE consortium. For simplicity, for each of the 685 factors we defined a gene as putatively bound if there was a binding event of the factor within 1kb of the ends of the gene.

Interestingly, 663 (out of 685) factors showed significantly different $\text{dS}_{\text{human-chimp}}$ values between putatively bound and non-bound genes after FDR correction for multiple testing (at 0.05 false discovery rate). Out of these 663 factors, 648 showed significantly lower $\text{dS}_{\text{human-chimp}}$ values in bound genes. Since many of these datasets correspond to the same factors in different cell lines or to co-factors often bound together, to minimise redundancy in the set of dS-associated factors, we performed a multiple linear regression as described in Section 4.2.7. This analysis yielded a non-redundant set of 154 factors independently associated with $\text{dS}_{\text{human-chimp}}$.

Figure 4.10 shows the median $\text{dS}_{\text{human-chimp}}$ of the genes bound by each of the non-redundant set of factors. Transcription-factors and components of the transcriptional machinery (including RNA-polymerase II) are shown in black, major transcriptional-activating histone modifications (including all histone acetylations, H3k4me3 and H3k36me3) are shown in green, major transcriptional-repressing histone modifications (H3k27me3 and H3k9me3) are shown in red, and other histone modifications are shown in blue. The figure suggests that the binding of different factors is associated with genes with lower synonymous divergence. Further, as transcriptional-activating marks appear to be associated with lower $\text{dS}_{\text{human-chimp}}$ values and transcriptional-repressing marks are linked to higher values, Figure 4.10 suggests that open chromatin may be associated with lower mutation rates.

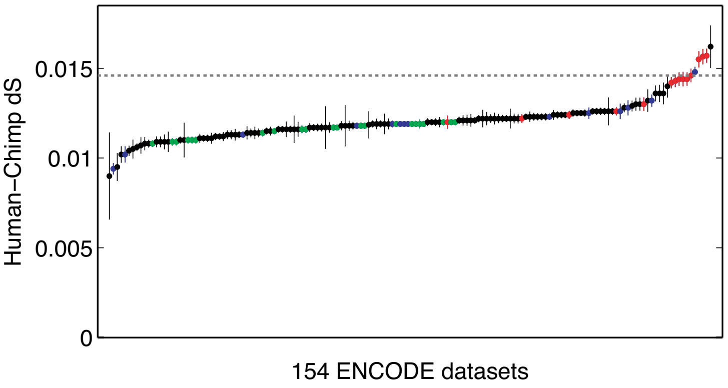

***Figure 4.10**. Association of transcription factor binding and chromatin remodelling with $dS_{human-chimp}$ values (estimated for whole genes). The median $dS_{human-chimp}$ for bound-genes, together with its 95% confidence intervals (error bars), is shown for the 154 ENCODE datasets that showed a non-redundant association with $dS_{human-chimp}$ according to the multiple linear regression analysis. The colour code is explained in the text.*

In the absence of further testing, this analysis has to be taken as very preliminary and merely suggestive. In order to obtain a more reliable understanding of the gene-to-gene variation of the mutation rate in humans, this preliminary analysis should be greatly expanded and complemented with additional factors (such as actual expression levels and functional gene annotations, as well as DNase I and FAIRE datasets).

Also, and most importantly, to avoid the ambiguity of previous studies in humans (Chuang and Li, 2004, Ying et al., 2010), potential biases such as selection, recombination and sequence composition should be carefully accounted for (Hodgkinson and Eyre-Walker, 2011). To do so, a similar strategy to the one described in Chapter 3 could be used. Fortunately, the recent availability of genome-wide polymorphism data from the HapMap project and the 1,000 genomes project should allow a quantitative evaluation of the role of selection on synonymous variation in humans (Bustamante et al., 2003), avoiding the classical limitation of studies of local mutation rates.

This future analysis may provide an unbiased picture of the gene-to-gene variation of the mutation rate along the human genome and shed light on the factors determining this variation.

Why is it important to understand selection? There are some important practical reasons: what we call purifying selection in the formal literature is disease when the mutation affects humans.

— Lawrence D. Hurst, *Genetics and the understanding of selection* (2009)

5

Repression of intronic *Alu* elements is essential for human transcriptome integrity

5.1. Summary of the chapter

Genomic integrity is first maintained by a low global mutation rate, established through the action of DNA repair pathways and DNA protection mechanisms. In addition, the previous chapter presented a strategy to reduce the risk of deleterious mutations by preferentially protecting the most functionally important regions of a genome. Finally, once mutations occur and are not repaired, at a species level the functional integrity of a genome is maintained by purifying selection against deleterious mutations.

Chapter 5 introduces an additional layer of complexity, describing a novel mechanism capable of buffering the deleterious impact of a certain type of mutation. Specifically, we describe a safeguarding mechanism by which the deleterious effect of *Alu* elements residing in introns is avoided by the action of an RNA-binding protein. Through this action, thousands of otherwise deleterious intronic *Alu* elements can silently persist in the genome with little damage to the organism.

The insertion of an *Alu* retrotransposon in an intron poses a serious threat to the stability of the host gene: cryptic splice sites and splicing signals in the *Alu* sequence can be recognised by the splicing machinery leading to its exonisation and the subsequent disruption of the mature transcript. This chapter shows how the RNA-binding protein hnRNP C, strongly binds to intronic *Alu* elements blocking their recognition by the splicing machinery and so protecting the integrity of the transcripts. The mechanism is shown to have a major role in maintaining human transcriptome integrity. This has implications for human health, as disruption of this repression at a single locus can cause a genetic disorder. Thus, this represents a new type of safeguarding mechanism for functional genomic integrity.

From an evolutionary point of view, this process is very different from repair or purifying selection, since instead of directly removing a harmful mutation it only minimises its potential functional impact. Thus, like the risk management strategy, this process selectively reduces the deleterious impact of mutations while maintaining or even fostering their evolutionary potential.

Contrary to the previous three chapters which were exclusively the result of my individual work, Chapter 5 describes my work as part of a collaborative project between our group and the group of Dr. Jernej Ule at the MRC-Laboratory of Molecular Biology in Cambridge (LMB). The project was led by Dr. Kathi Zarnack (EBI) on the computational side, and Dr. Julian König (LMB) on the experimental side.

Section 5.2 provides an introduction to the project and a description of the experimental data used. After this introduction, Sections 5.3 and 5.4 describe in detail the main results that I contributed to the project (Figures 5.5-5.13).

A manuscript describing this work has been submitted for publication.

> Zarnack K†, König J†, Tajnik M, **Martincorena I**, Stévant I, Reyes A, Anders S, Luscombe NM*, Ule J*. 2012. Direct competition between hnRNP C and U2AF65 controls the exonization of hundreds of *Alu* elements. *Submitted.*
>
> (†) denotes equal contribution. (*) denotes corresponding author.

5.2. Introduction: splicing regulation and the threat of pseudoexonisation

5.2.1. Alternative splicing

Alternative splicing is a crucial mechanism for gene regulation and for generating proteomic diversity within cells and among cell types. It is estimated that the great majority (95-100%) of human multi-exon genes are subject to alternative splicing (Nilsen and Graveley, 2010).

Splicing is carried out by the spliceosome, a large complex comprising five small nuclear ribonucleoproteins (snRNPs) and multiple auxiliary proteins, which cooperate to recognise the splice sites and to perform the splicing reaction accurately. The assembly and action of the spliceosome involve the following sequential steps (Chen and Manley, 2009), as depicted in Figure 5.1:

1. E′ complex: U1 snRNP recognises the 5′ splice site and the splicing factor 1 (SF1) binds to the branch point.
2. E complex: The U2AF (U2 Auxiliary Factor) heterodimer binds to the polypyrimidine tract and the 3′ AG. The assembly up to this point is ATP-independent.
3. A complex: SF1 is replaced by the U2 snRNP at the branch point.
4. B complex: Recruitment of U4/U6-U5 tri-snRNP complex. This large complex then undergoes extensive conformational changes and remodelling, including the loss of U1 and U4, resulting in the formation of the catalytically active spliceosome (C complex).

The decision of which exons are included or removed depends on four types of cis-regulatory sequence elements: exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs) and intronic splicing silencers (ISSs). Enhancers and silencers affect the binding of different factors that determine the balance between exon inclusion and exclusion. For example, ESEs are commonly bound by members of the SR protein family (Shepard and Hertel, 2009) while silencer elements (ESSs and ISSs) are usually bound by heterogeneous nuclear ribonucleoproteins (hnRNPs).

Alternative splicing decisions are believed to be made most frequently during splice site recognition and early spliceosome assembly, although different mechanisms have been described to act at different stages.

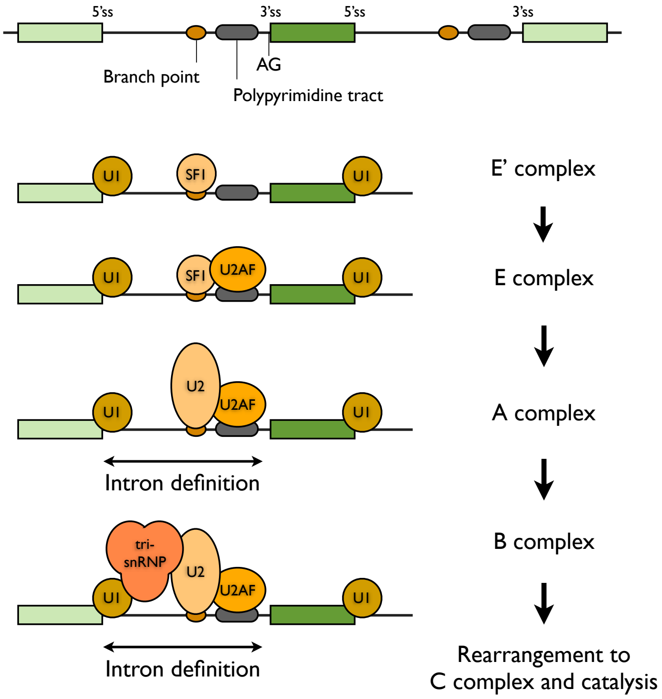

Figure 5.1. *Spliceosome assembly and splicing. Steps explained in the text.*

5.2.2. Aberrant splicing and human disease

The characteristics that make a nucleotide sequence an exon (recognised by the spliceosome) remain poorly known. Indeed, eukaryotic genes contain a large number of sequences that look like perfect exons but are somehow ignored by the splicing machinery (Dhir and Buratti, 2010). This is important: erroneous recognition of non-functional intronic sequences as exons (hence pseudoexons) would generally lead to disruption of the resulting transcript by (a) causing a frameshift of the downstream exons, (b) introducing premature stop-codons or (c) introducing non-functional amino acid sequences into the resulting proteins. Thus, tight control of the accessibility of cryptic splice sites might be required to ensure transcriptome integrity.

The importance of this threat is becoming more obvious as clinical studies have already identified more than 60 pathological pseudoexonisation events causing different genetic disorders (Dhir and Buratti, 2010). Most of these pathological pseudoexon inclusion events originate from point mutations or small indels creating a new donor or acceptor splice site, or creating/deleting splicing regulatory sequences.

A relatively neglected question in this context is the role of trans-acting factors in pseudoexonisation events. For example, the rat *α-tropomyosin* gene contains a pseudoexon whose induced inclusion leads to the truncation of the protein and activation of the nonsense-mediated decay pathway to ensure the fast degradation of the mRNA. Interestingly, repression of this exon was observed following the overexpression of PTB (Polypyrimidine tract-binding protein), a well-known splicing repressor with a major role in alternative splicing regulation. Recently, PTB was also shown to repress the inclusion of a pathological pseudoexon in humans, suggesting that silencer sequences may be commonly used to repress the undesirable exonisation of intronic sequences (Dhir and Buratti, 2010).

5.2.2.1. Alu elements are naturally prone to accidental exonisation

An important class of pseudoexons in humans originates from *Alu* elements inserted in introns. As described in the introduction, *Alu* elements account for more than 10% of the human genome, appearing in the introns of a large fraction of genes.

Alu elements inserted in an intron in the antisense orientation may pose a significant threat since the antisense sequence contains 3′ cryptic splice sites downstream of long U-tracts (Figure 5.2), which are typical sequence features required for recognition by the splicing machinery. Indeed, exons derived from *Alu* elements are estimated to account for more than 5% of alternatively spliced exons in humans (Lev-Maor et al., 2003). Importantly, while there are some examples of functional exons that evolved from *Alu* elements, the vast majority of accidentally included *Alu* sequences will have deleterious effects.

In fact, *Alu*-derived pseudoexons have been described in at least four genetic disorders (Dhir and Buratti, 2010). While this already highlights the clinical importance of aberrant *Alu* exonisation, it is likely to underestimate their real impact. For example, a systematic screen of 78 disease-causing pseudoexons (affecting 51 genes) revealed that eleven emerged from *Alu* sequences (of which ten were inserted in the antisense orientation).

Figure 5.2 shows the canonical sequence of antisense *Alu* elements, highlighting some of the sequence elements required for the recognition by the splicing machinery.

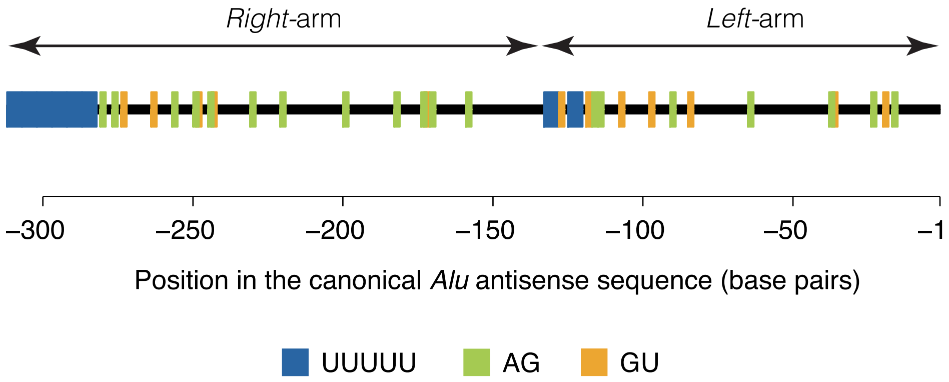
Right-arm
Left-arm
–300
–250
–200
–150
–100
–50
–1
Position in the canonical *Alu* antisense sequence (base pairs)
UUUUU
AG
GU

***Figure 5.2**. Schematic representation of some sequence elements along the canonical Alu antisense sequence that could be relevant for splicing. U-tracts (≥5 nucleotides) are shown in blue. There are two major U-rich regions, a long U-tract at the beginning of the sequence, corresponding to the poly-A tail of the Alu RNA ("terminal U-tract"), and a U-rich region in between the two arms ("linker U-tract"). These U-tracts could act as splicing signals. In addition, potential 3′ and 5′ splice sites are shown in green and orange, respectively. "Right" and "left" arms refer to their position in the sense representation, following the standard nomenclature in the literature.*

5.2.3. The role of hnRNP C as a splicing regulator

In 2010, König and colleagues described the role of the RNA-binding protein hnRNP C in splicing regulation using iCLIP, a new protocol to study protein-RNA interactions on a genome-wide scale at single-nucleotide resolution (see Section 5.2.4.2).

Using binding data, König revealed that hnRNP C has a strong affinity for continuous uridine tracts (U-tracts). To examine the possible role of hnRNP C in splicing regulation, they performed an *HNRNPC* knockdown and used a splice-junction microarray to identify alternatively spliced exons. This resulted in a total of 1,340 differentially spliced exons upon *HNRNPC* knockdown. Studying the binding of hnRNP C at these exons, they demonstrated that hnRNP C silences exon inclusion by binding to U-tracts upstream of the 3′ splice site (Figure 5.3).

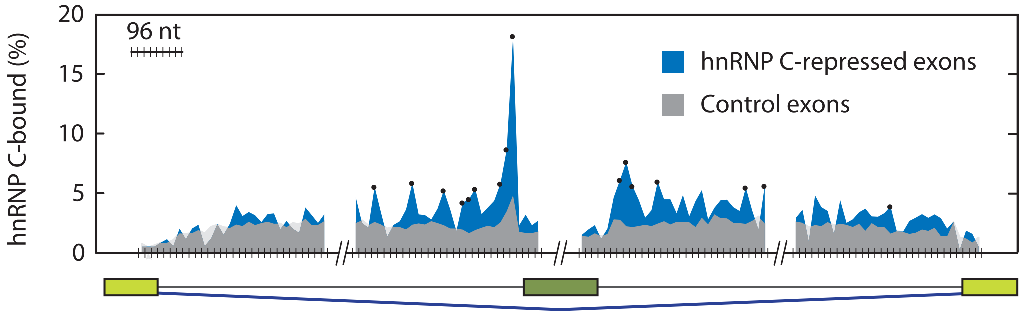

Figure 5.3. Averaged RNA-binding profile showing the binding frequency of hnRNP C around a set of hnRNP C-repressed exons (over-expressed upon HNRNPC knockdown - background blue) and around a set of control exons (unaffected by the knockdown - foreground grey). hnRNP C binding at each window is calculated as the percentage of exons that have one or more crosslinking events. Black dots mark statistically significant enrichment of binding over the control exons (P<0.01, Fisher's exact test). Figure adapted from König et al., 2010.

5.2.4. New analyses on the role of hnRNP C

The fact that hnRNP C represses exon inclusion by binding to the U-tract upstream to the 3′ splice site suggests that it may act by blocking U2AF-recruitment, so aborting the assembly of the spliceosome. To test this hypothesis, after their original experiments König performed iCLIP on the U2AF subunit 65 (U2AF65).

In addition, in their original analysis König and collaborators described that hnRNP C binding peaks often occur in pairs separated by ~165 nucleotides. This surprising feature was originally interpreted as a possible consequence of the known tetrameric organisation of hnRNP C (König et al., 2010). However, this spacing is also consistent with the distance between the two U-rich sequences in *Alu* elements (Figure 5.2). This insight led Zarnack and König to evaluate the contribution of *Alu* sequences to their original observations, the part of the work in which I became involved.

Thus, the project described here stems from König's original work with the aim to study the impact of hnRNP C on U2AF65 binding and on the exonisation of Alu elements. This project combines a variety of experimental methods, including multiple molecular techniques and two genome-wide approaches that are described below: RNA-sequencing and iCLIP. The genome-wide datasets were generated by König and processed by Zarnack.

5.2.4.1. RNA-sequencing

In order to investigate the impact of hnRNP C on splicing regulation, RNA-sequencing was performed in *HNRNPC* knockdown and control HeLa cells. The reproducibility of the knockdown was tested using two different siRNAs (KD1 and KD2), and biological and technical variance was controlled by generating duplicates of each dataset. The resulting RNA-Seq libraries were sequenced using Illumina GA IIx technology, generating a total of 240 million 72-nt paired-end sequence reads.

The reads were mapped to the hg19 version of the human genome using the splicing-aware alignment software TopHat (Trapnell et al., 2009). More than 80% of the reads mapped uniquely. Cufflinks was then used on the TopHat mapping for *de novo* annotation of transcript structures, thereby allowing the discovery of non-annotated exons (Roberts et al., 2011). Cufflinks predicted a total of 178,029 exons within 14,091 previously annotated genes. Importantly, this list included 16,143 novel exons that showed no overlap with any exon annotated in the Ensembl database. Of these, 1,903 novel exons were derived from intronic *Alu* elements.

In order to identify whether these exons were regulated by hnRNP C, Kathi Zarnack used the DEXSeq software (Anders et al., in preparation), identifying a high-confidence set of 3,052 exons (out of 178,029) that significantly changed their inclusion levels upon *HNRNPC* knockdown. This list included 363 exons derived from *Alu* elements.

5.2.4.2. iCLIP

iCLIP is one of the newest and most precise experimental techniques for characterising protein-RNA interactions (König et al., 2010, König et al., 2012). It employs ultraviolet crosslinking followed by immunoprecipitation with an antibody specific to the protein of interest. Unlike traditional ChIP-sequencing for protein-DNA interactions, which sequences bound DNA after fragmentation leading to a resolution on the order of hundreds of bases, iCLIP performs several library preparation steps that provide single-nucleotide resolution and strand-specificity (König et al., 2010). See Figure 5.4 for a schematic representation of the iCLIP protocol.

Here, we performed two replicate iCLIP experiments from untreated HeLa cells, obtaining a total of 14 million unique crosslinking events. Using a peak-finding algorithm we identified a total of 438,360 binding sites. Consistent with previous results (König et al., 2010), hnRNP C binding primarily occurs within introns and shows a strong preference for continuous U-tracts. Further, using these new datasets, the strength of hnRNP C-binding can be seen to increase with the length of the U-tract, with the strongest occupancy occurring at U-tracts longer than nine continuous uridines.

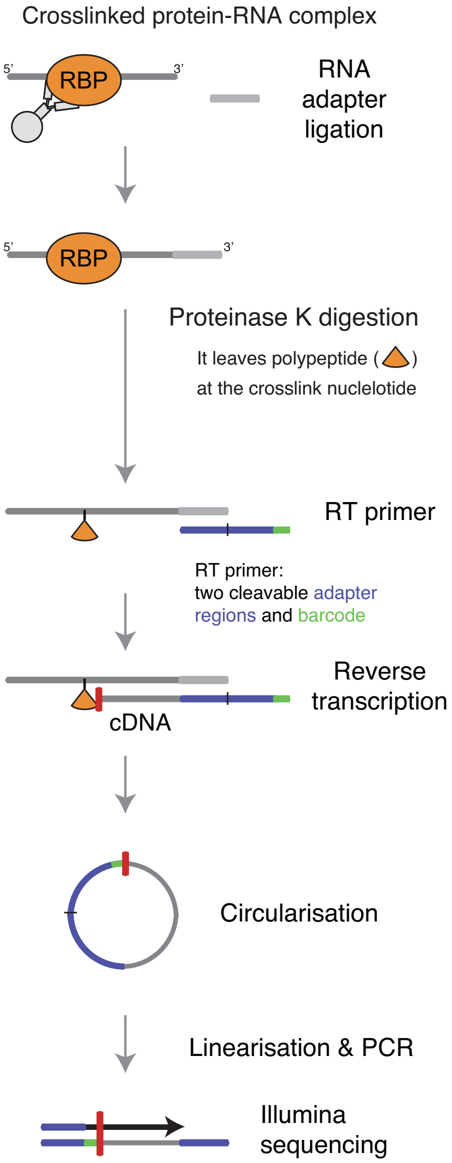

***Figure 5.4**. Schematic representation of the iCLIP protocol. UV irradiation is used to covalently crosslink the RNA-binding protein to the RNA. This is followed by a co-immunoprecipitation of the RNA and the RNA-binding protein. The RNA is then ligated to an RNA adapter. Digestion with proteinase K leaves a short polypeptide covalently bound to the crosslinking nucleotide, which causes premature truncation of the reverse transcription (RT) at the crosslinking site. In the figure, the red bar represents the last nucleotide added during the reverse transcription. Resulting cDNA molecules are then circularised, linearised (by a restriction endonuclease site in the adapter), PCR amplified and Illumina sequenced. The first nucleotides in each read contain a barcode (to allow multiplexing) and a random tetramer (which allows the identification of PCR duplicates). These sequences are then followed by the last nucleotide that was added during the reverse transcription, which provides single-nucleotide resolution and strand-specificity to the iCLIP protocol. Figure adapted from König et al., 2010.*

In addition, we performed iCLIP experiments of U2AF65 in control HeLa cells and *HNRNPC* knockdown cells, to study the changes in the binding of U2AF65 in the absence of hnRNP C repression. This yielded a total of 12 million crosslinking events which grouped into more than half a million of significant binding sites. Using these datasets Zarnack demonstrated that, as hypothesised, hnRNP C represses exon inclusion by competing with U2AF65 at U-tracts acting as splicing signals. This was then experimentally validated by König using synthetic gene constructs.

In this context, the following sections describe my analyses on the impact of hnRNP C-regulation on Alu pseudoexonisation.

5.3. hnRNP C represses the aberrant exonisation of thousands of intronic *Alu* elements

5.3.1. Knockdown of *HNRNPC* globally derepresses intronic *Alu* elements

The RNA-sequencing data revealed a list of 3,052 exons that significantly changed their inclusion upon *HNRNPC* knockdown. Of these, 1,807 exons (59%) showed a higher inclusion in the knockdown and 1,245 a lower inclusion level. In contrast, out of the 363 significantly changed *Alu* exons, 361 (99.4%) showed increased inclusion in the knockdown. Since hnRNP C is a repressor of exon inclusion, this suggests that while the changes observed in many non-*Alu* exons may be due to indirect effects, *Alu* exons are directly repressed by hnRNP C.

Interestingly, the repression of *Alu* exons by hnRNP C is not restricted to these 361 exons, as extending the analyses to all 1,903 *Alu* exons detected in the RNA-Seq data reveals that they are globally upregulated by the knockdown of *HNRNPC* (Figure 5.5). The reason most of these exons were not detected as significant using DEXSeq is probably because of the low RNA-Seq read-count from most of them. Indeed, out of 20 *Alu* exons that were flagged as non-significant by DEXSeq, 16 were confirmed as significantly de-repressed using quantitative PCR. Thus, *HNRNPC* knockdown causes a global derepression of hundreds, and most likely thousands, of intronic *Alu* elements in the human genome.

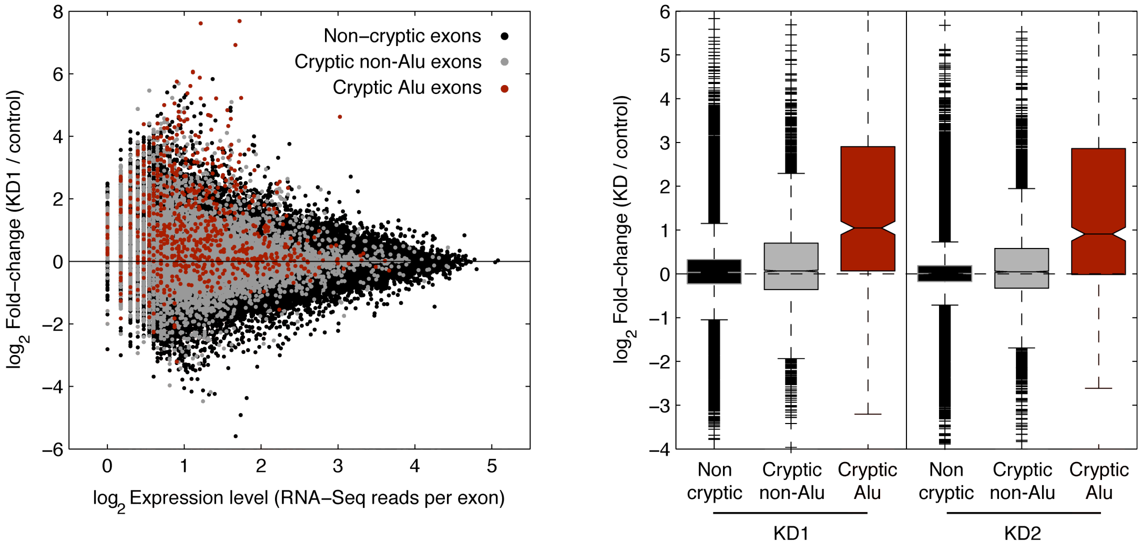

***Figure 5.5**. The knockdown of HNRNPC causes a global derepression of intronic Alu sequences giving rise to Alu exons disrupting hundreds of transcripts. The figure shows the fold-change in the expression level of exons upon knockdown of HNRNPC. Exons depicted in black correspond to de novo identified exons overlapping with exons annotated in Ensembl. Cryptic exons (non-overlapping with Ensembl annotated exons) are classified into non-Alu (grey) or Alu exons (red). The figure contains two alternative representations evidencing that the HNRNPC knockdown leads to the global upregulation of Alu exons: (**left**) an MA-plot shows the upregulation of Alu exons particularly at very low expression levels, (**right**) Boxplot showing the distribution of fold-change values in each group of exons in both knockdown experiments (KD1 and KD2). The differences between groups are extremely significant in all cases ($P<10^{-100}$, Wilcoxon rank-sum test).*

The analysis in Figure 5.5 is based on the $\log_2$ fold-change expression of each exon in the knockdown and the control cell lines. By definition, this analysis requires a non-zero expression of the exon in the control sample. However, it is likely that some deleterious exons may be fully repressed in the presence of hnRNP C, and so they will be missed by this analysis. To tackle this issue, Figure 5.6 represents the fraction of exons with zero reads in each condition. The analysis further support our hypothesis that *Alu* exons are strongly repressed in the presence of hnRNP C with a substantial fraction of them below the detection limit of RNA-Seq at this coverage.

Together, Figures 5.5 and 5.6 demonstrate that hnRNP C represses nearly 2,000 *Alu* exons in the human genome. However, this number is likely to underestimate the real scale of this phenomenon, since most *Alu* exons are very close to the detection limit by RNA-Seq. In fact, the iCLIP data indicates that hnRNP C binds to ~72,000 intronic antisense *Alu* elements (21% of all antisense *Alu* elements in the transcriptome, in contrast to 0.03% of all sense *Alu* elements). Taken together, these analyses highlight the role of hnRNP C as a global repressor of intronic *Alu* exonisation.

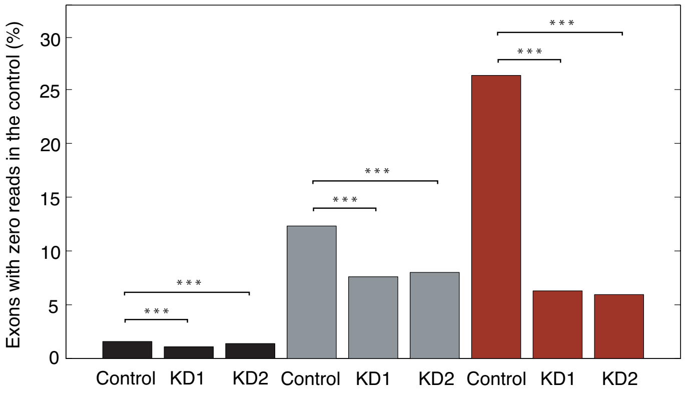

Figure 5.6. *Barplot showing the percentage of exons in each group with no reads in the RNA-Seq dataset of the control (i.e. no exonisation in the presence of hnRNP C). *** indicates statistically significant differences with P<0.0001, Pearson's chi-squared test. The figure shows that a large majority of non-cryptic (overlapping with Ensembl) exons (black) are detected in the control HeLa cell. In contrast, a substantial fraction of cryptic exons are only detected upon HNRNPC knockdown, in particular Alu exons (depicted in red).*

The inclusion of intronic *Alu* sequences as exons is likely to disrupt the expression of the corresponding transcripts or the function of the encoded protein, which may have very deleterious effects in the organism. Analysing the *Alu* exons detected in the RNA-Seq data reveals that 74.3% introduce premature stop codons in all three frames and/or cause a frameshift in the downstream exons. The remaining *Alu* exons are also likely to alter the function of the encoded protein by introducing non-native polypeptides in its sequence.

5.3.2. Sequence elements involved in the formation of *Alu* exons

The human genome comprises around 1.2 million *Alu* elements (Jurka et al., 2005, Cordaux and Batzer, 2009), as a result of their mobilisation over the past ~65 million years. Given their relatively recent evolution, the sequence divergence among paralogous copies is limited (Figure 5.8) and most *Alu* elements maintain the two ancestral arms. Thus, it is conceivable that exonisation and binding at intronic *Alu* elements might occur at certain preferred sites in most cases. In order to understand the origin of exons from intronic *Alu* elements, the possible preference for certain cryptic splice sites in the ancestral sequence (Figure 5.2), and the bindings of hnRNP C and U2AF65 within *Alu* elements, we aligned the sequences of *Alu* elements in the human genome to the canonical *Alu* sequence from the RepeatMasker annotation (Smit et al., 2012).

5.3.2.1. Multiple sequence alignment of Alu elements

In particular, we aligned to the canonical *Alu* sequence 856,791 *Alu* elements in the human genome whose lengths are within 15% of the canonical sequence length. The analysis was restricted to near-complete *Alu* elements to avoid alignment artefacts and to avoid confounding the interpretation of the analysis by including partial *Alu* copies.

To ensure high quality alignments, we use multiple-sequence alignments rather than pairwise ones, aligning *Alu* sequences in groups. This allows to exploit the phylogenetic relatedness of the sequences to obtain more reliable alignments. In particular, we aligned the sequences in 21,969 multiple alignments of 40 *Alu* elements using PRANK with default parameters for non-coding DNA (Loytynoja and Goldman, 2008).

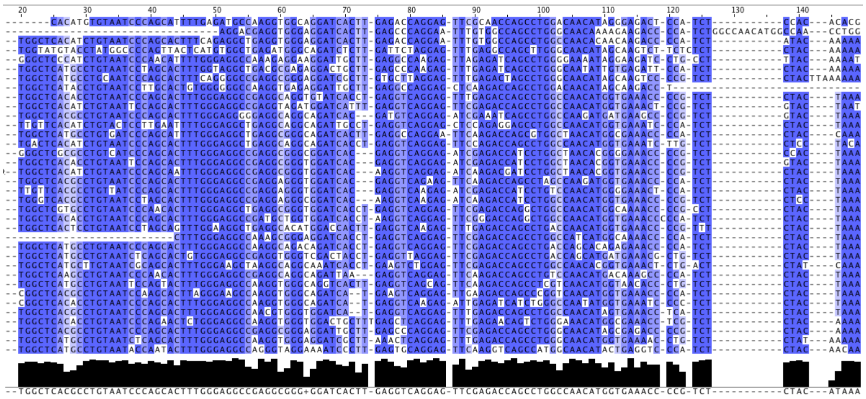

***Figure 5.7**. Representative segment of a PRANK multiple sequence alignment of paralogous Alu copies. The colour of each nucleotide represents the percentage identity to the consensus base at each site. Apparently very variable sites most often correspond to variable sites between major Alu subfamilies, as well as to some CpG sites along the sequence. A histogram representing the level of conservation per site is shown in black at the bottom of the alignment.*

Comparing every sequence to the canonical sequence we annotated the base substitutions, insertions and deletions in each *Alu* element. The frequency of these events along the reference is shown in Figure 5.8. As expected, the profiles of single-base substitutions or point mutations show some very variable sites, which correspond to variable sites among the different *Alu* subfamilies that propagated clonally at different times during primate evolution. Excluding variable sites between families, which do not reflect independent mutation events, reveals that the level of sequence divergence among paralogous *Alu* elements is low (notice the low baseline in Figure 5.8, top panel).

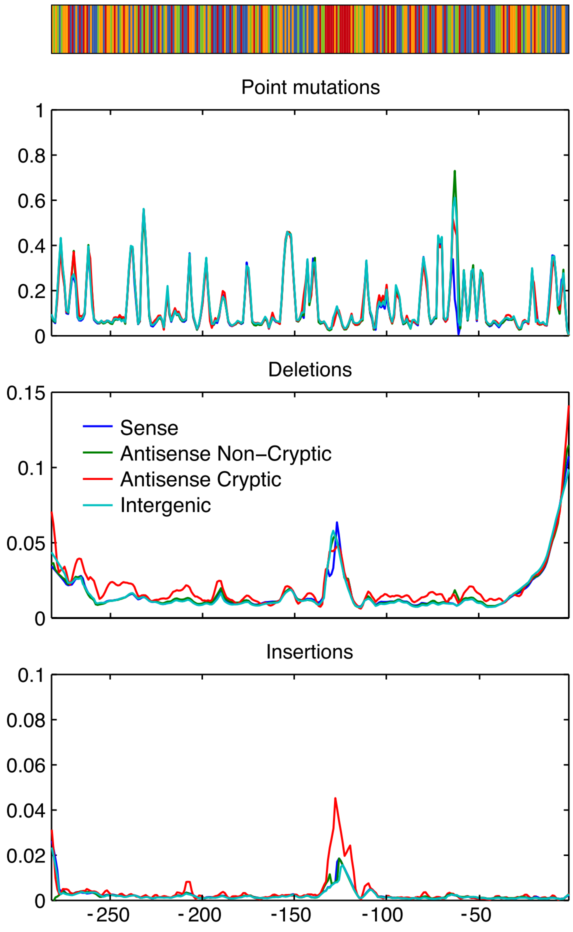

Figure 5.8. *Comparative genomic analysis of the sequence divergence of complete or near-complete Alu elements in the human genome. The figure represents the frequency of single-base changes (top), deletions (middle) and insertions (bottom), per base, with respect to the canonical reference Alu sequence. The lines represent the fraction of Alu elements in each category showing these events at each position. To aid the interpretation of the profiles, the figure includes a colour-coded representation of the canonical Alu sequence on top (A=green, C=blue, G=orange, T=red). Note that the peak of deletions and insertions around -125 corresponds to the linker T-rich region. The terminal long T-tract was excluded from this analysis as their alignment is naturally unreliable owing to their flanking position and minimal complexity.*

In Figure 5.8, *Alu* elements are classified into four categories, according to their position with respect to Ensembl transcripts: *intergenic* (for those outside transcripts), *sense* (intronic *Alu* elements in the sense orientation relative to an annotated transcript) and *antisense* (classified as *cryptic* or *non-cryptic*). Cryptic antisense *Alu* elements refer to intronic antisense *Alu* elements overlapping with a new exon detected in the RNA-Seq of the *HNRNPC* knockdown and not annotated in Ensembl.

The profiles of point mutations, deletions and insertions are very different, and yield very interesting observations:

1. **Similar divergence.** The profile of point mutations is nearly identical for the four categories, suggesting that there are no systematic differences in the age or distribution of families of the *Alu* elements in the four categories.
2. **Low frequency of indels.** On average, the frequency of a given base being affected by a deletion is lower than 2% for most of the *Alu* sequence, and the frequency of a base immediately upstream of an insertion is lower than 0.5% for most of the *Alu* sequence.
3. **Higher frequency of indels in the linker T-rich region**. Both deletions and insertions appear to occur more frequently in the linker region, whose canonical sequence is TTTTTTGTATTTTT. This may not be surprising as homopolymeric tracts often show higher rates of short indels due to polymerase slippage (Leclercq et al., 2010).
4. **Signatures of incomplete reverse transcription.** Deletions are most frequent at the end of the *Alu* sequence (in antisense orientation, *i.e.* 5′ end in the typical sense orientation), which may be due to incomplete reverse transcription or improper integration (Salem et al., 2003).
5. **Higher frequency of insertions in the linker region of cryptic *Alu* exons.** This is unexpected given that the rate of insertions does not seem to be higher in the rest of the sequence. According to results described in Section 5.4, this may be a consequence of positive selection favouring longer T-tracts in the linker region of cryptic *Alu* exons, to ensure stronger repression by hnRNP C.

These observations provide an interesting summary of the sequence diversity among *Alu* elements in the human genome. Nevertheless, more importantly in the context of this study, the observation of the limited sequence divergence of intronic *Alu* elements justifies using the multiple sequence alignments to map the sequences of individual *Alu* elements to the reference sequence. Doing so allowed us to study the role of different sequence elements in the process of *Alu* exonisation.

5.3.2.2. Characterisation of splice site usage and exonised fragments

First of all, we evaluated the segments of the *Alu* sequence involved in the formation of exons. Based on the RNA-Seq data, we obtained a list of 2,085 exons with at least one of its splice sites mapping to an antisense *Alu* element. Using this list, we can quantify the usage of different donor (5′) and acceptor (3′) splice sites in the formation of exons in the *HNRNPC* knockdown (Figure 5.9).

Using 684 exons whose donor and acceptor splice sites were both identified by RNA-Seq, Figure 5.9a shows the segments of the *Alu* canonical sequence that became exonised. The figure clearly shows that most *Alu* exons are short (around 100 bp long) and originate from one of the two arms of intronic *Alu* elements (single-arm *Alu* exons). The figure also reveals a smaller set of *Alu* exons spanning both arms, as well as some exons, particularly left-arm *Alu* exons that extend downstream of the *Alu* element.

Figure 5.9a also shows that most exons start and end at a few preferred splice sites in the canonical *Alu* sequence. This is most clearly shown in Figure 5.9b which displays the frequency of splice site usage at nucleotide resolution. Most *Alu* exons start at one of three main 3′ splice sites present in the canonical *Alu* sequence. Interestingly, these sites are a few nucleotides downstream of the two U-rich regions in the *Alu* elements which are expected to act as splicing signals (see Figure 5.1).

In most cases, cryptic splice sites used during *Alu* exonisation are already present in the ancestral *Alu* sequence. An evaluation of the sequence enrichment at the exon boundaries reveals that all 3′ splice sites harbour the canonical AG motif and most of 5′ splice sites include the GU motif (Figures 5.9c and 5.9d). Interestingly, the two main 3′ splice sites had already been described before as typical sites used in *Alu* exonisation using a collection of *Alu* exons from public expressed sequence tag (EST) data (Lev-Maor et al., 2003).

In summary, this analysis combined a comparative genomic analysis of paralogous *Alu* copies and *de novo* discovery of exons and splice sites using RNA-Seq data to reveal a comprehensive and high-resolution picture of the emergence of *Alu* exons in the human transcriptome upon *HNRNPC* knockdown. Most *Alu* exons are shown to originate from a single arm of full-length intronic *Alu* elements, typically using a small number of ancestral cryptic splice sites located a few nucleotides downstream of the also ancestral U-rich regions. This suggests that the main sequence elements required for the aberrant exonisation of *Alu* elements are already present in the ancestral sequence. This observation underlines the importance of a global repression of *Alu* elements by hnRNP C, as the sequence elements needed for exonisation are common to a vast number of intronic *Alu* elements in the human genome.

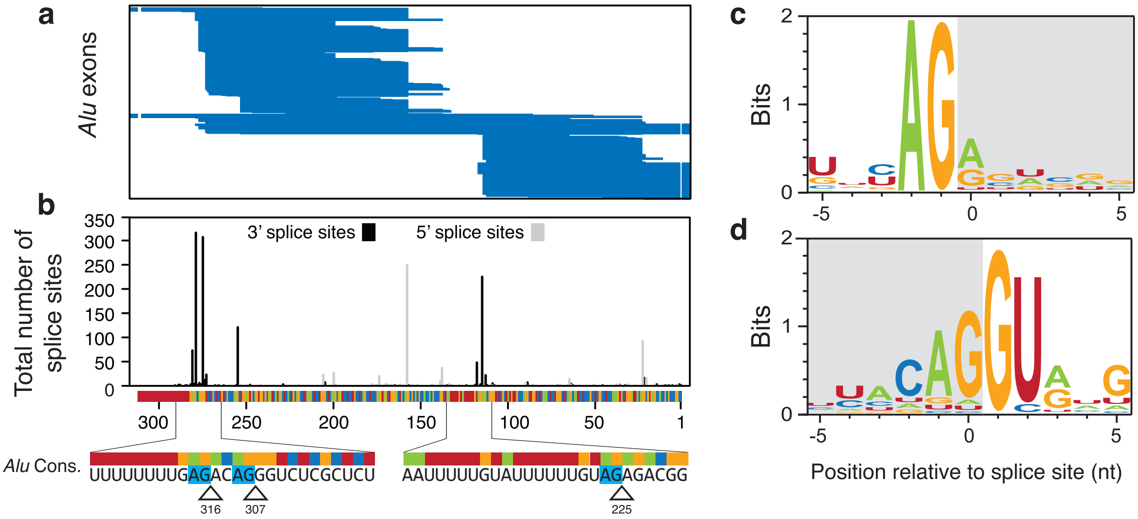

***Figure 5.9**. Origin of Alu exons and splice site usage along the canonical Alu sequence. (a) Schematic representation of the exonised segments of Alu elements using the set of Alu exons whose both splice sites could be identified unambiguously. The fragments of the canonical Alu sequence observed in the RNA-Seq data (exons) are shown in blue. Exons that extend beyond the Alu element (i.e. whose 3′ or 5′ splice sites fall outside an annotated Alu) are depicted as extending outside the range of the Alu coordenates (x-axis). (b) Barplot of the usage frequency of splice sites along the canonical Alu sequence. 3′ splice sites are shown in black and 5′ splice sites in grey. At the bottom of the panel the exact positions of the main 3′ splice sites are shown along the canonical Alu sequence. (c) We*blogo *representation of the sequences around the exon boundaries. The figure clearly shows that all 3′ splice sites contain the canonical AG motif and the vast majority of 5′ splice sites contain the GU motif.*

5.3.2.3. Repression of Alu exonisation by competition between hnRNPC and the core splicing machinery at the U-rich regions of Alu elements

The preferential usage of the cryptic 3′ splice sites immediately downstream of U-tracts suggests that these tracts are likely to act as splicing signals that recruit U2AF during the formation of the spliceosome (Figure 5.1). We hypothesised that strong hnRNP C-binding to ancestral U-tracts might help repress the aberrant exonisation of intronic *Alu* elements by preventing the binding of U2AF65 and so avoiding the assembly of the spliceosome.

To study this hypothesis we evaluated the hnRNP C- and U2AF65-binding within intronic *Alu* elements. Figure 5.10a shows the binding frequency of hnRNP C at each nucleotide of intronic *Alu* elements in HeLa cells. As expected, hnRNP C shows a strong preference for binding at U-tracts, and so the binding of hnRNP C is virtually restricted to the ancestral terminal and linker U-rich regions. The results in Figure 5.10 also highlight the extremely high resolution of the iCLIP data and the accuracy of the alignments, as we were able to resolve two distinct peaks at the two continuous U-tracts of the linker U-rich region (*UUUUUU*GUA*UUUUU*). This contrasts dramatically with the resolution shown by ChIP-Seq datasets, typically on the order of a few hundred bases.

Figure 5.10b shows the binding profile of U2AF65 within intronic *Alu* elements, in both the control and *HNRNPC* knockdown HeLa cells. Similarly to hnRNP C, U2AF65 preferentially binds to the U-tracts in *Alu* elements, confirming the role of the ancestral U-tracts as the main splicing signal for the binding of U2AF65 during spliceosome assembly at intronic *Alu* elements. However, in contrast to hnRNP C, a non-negligible level of U2AF65-binding was observed throughout the entire *Alu* sequence. This is consistent with the observation that U2AF65 is not as specific as hnRNP C for continuous U-tracts and instead binds other pyrimidine-rich sequences.

Importantly, a comparison of the normalised profiles in the *HNRNPC* knockdown revealed that U2AF65-binding to intronic *Alu* elements increases in the absence of hnRNP C. This supports the hypothesis that strong binding by hnRNP C to intronic *Alu* elements protects them from recognition by the splicing machinery, so minimising their accidental exonisation. Reiterating this observation, Figure 5.11 shows the ratio of U2AF65 binding within *Alu* elements in the knockdown *versus* the control, including the flanking regions around intronic *Alu* elements for reference. This analysis demonstrated that the repression of U2AF65 binding by hnRNP C spans the entire *Alu* sequence and does not extend far beyond it into other intronic regions.

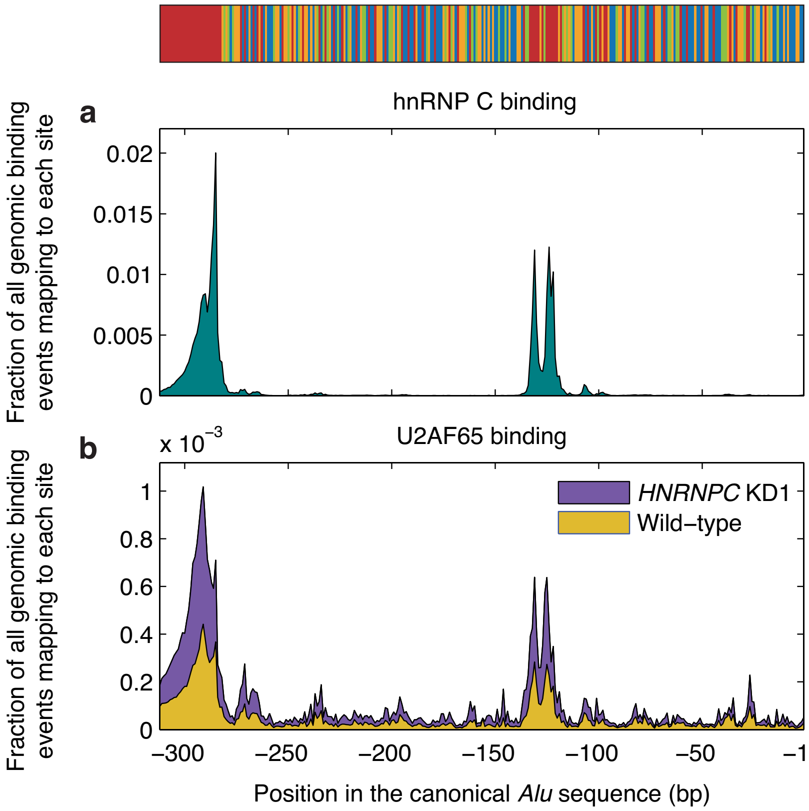

***Figure 5.10**. Binding profiles of hnRNP C and U2AF65 within intronic Alu elements. To obtain these profiles, all iCLIP crosslinking events (cDNA counts) occurring within annotated Alu elements in the genome were mapped to their corresponding position in the reference Alu sequence using the multiple sequence alignments. The y-axes represent the sum of all crosslinking events detected at a single site, divided by the total number of crosslinking events in the genome. (**a**) Binding profile of hnRNP C in control HeLa cells. (**b**) Binding profiles of U2AF65 in control HeLa cells (yellow) and in the HNRNPC knockdown (using KD1) (purple). Using the binding data from KD2 yielded analogous results. The figure clearly shows that U2AF65 binding within Alu elements increases substantially upon HNRNPC knockdown.*

Finally, a comparison of the scale of the y-axes of the binding profiles of hnRNP C and U2AF65 in Figure 5.10, reveals that hnRNP C has a much stronger preference for binding to intronic *Alu* elements than U2AF65. Binding at intronic *Alu* elements accounted for at least 24.44% of hnRNP C binding events in the human genome. This number is likely to be an underestimation given the difficulty of mapping reads uniquely to *Alu* elements (non-uniquely mapping reads were conservatively excluded from the analysis) and given that only near-complete *Alu* copies were considered in this analysis. In contrast, only 1.61% and 3.59% of U2AF65 binding events were detected inside *Alu* elements in the control and *HNRNPC* knockdown, respectively. This highlights that repression of *Alu* exons is a major function of hnRNP C.

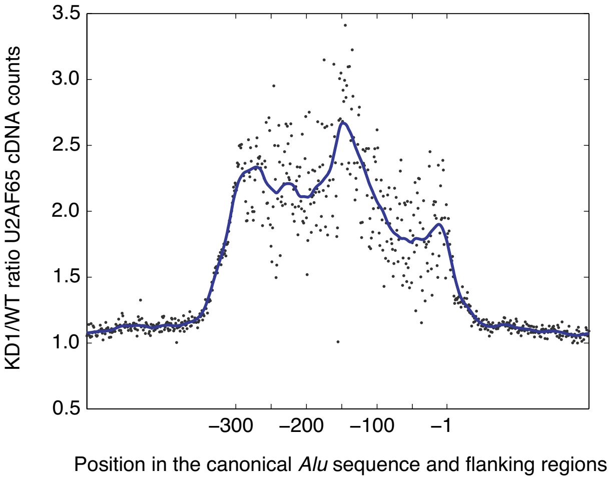

Figure 5.11. *Increase of U2AF65 binding at intronic Alu elements upon HNRNPC knockdown. The figure shows the ratio of U2AF65 binding (normalised by total number of binding events in the genome) in the knockdown versus the control. This demonstrated that the ~2.3-fold increase in U2AF65 binding is restricted to Alu elements and does not extend to their flanking regions, further highlighting the importance of hnRNP C on the repression of Alu exons.*

5.4. Evidence of positive and negative selection acting on hnRNP C binding to strengthen transcriptome integrity

Section 5.3.1 showed that hnRNP C represses the destructive exonisation of thousands of intronic *Alu* elements. Section 5.3.2 then provided a detailed description of the sequence elements involved in the formation of *Alu* exons. This revealed that hnRNP C strongly binds to intronic *Alu* elements at the U-tracts, blocking the binding of U2AF65 to these splicing signals, and so preventing exonisation.

Given that hnRNP C-binding preferentially occurs at continuous U-tracts we then evaluated if selection against aberrant *Alu* exonisation has favoured the conservation of long continuous U-tracts in *Alu* elements capable of exonisation, thus maintaining hnRNP C repression.

The detection of selection signatures in this context is challenging for several reasons. First, even if *Alu* elements are likely to evolve almost neutrally, the evolutionary divergence among paralogous copies is limited given their relatively recent expansion. Second, any analysis must consider that the indel rate at U-tracts is likely to be much higher than the mutation rate of the flanking sequences, owing to processes such as polymerase slippage. Third, different types of mutations (point mutations, insertions and deletions) may lead to analogous changes in the length of continuous U-tracts, and so they may have similar phenotypic consequences. And finally, the strength of selection may not be sufficiently strong to leave a clear signal, particularly considering the small population sizes during the evolution of *Alu* elements and the fact that we are studying mutations at intronic repeats, which are most often considered neutral.

To account for these limitations, we analysed the frequency of unusually short and unusually long U-tracts, with respect to the ancestral lengths originally shared by all *Alu* elements. Intronic *Alu* elements that lead to exons in the absence of hnRNP C are good candidates for selection to maintain hnRNP C repression, and so we compared the relative frequency of unusually short and long U-tracts in these exons to these frequencies in a control set of non-exonised *Alu* elements. Since the control set is matched with the set of exonised *Alu* elements (these are non-exonised *Alu* elements from transcripts where at least another *Alu* is exonised in the *HNRNPC* knockdown experiment), both sets should undergo similar mutation rates. Based on the level of divergence from the canonical *Alu* sequence, we also believe that both sets of *Alu* sequences are of similar age. This makes the comparison of the gain and loss of U-tracts particularly informative. Since control *Alu* elements are not exonised in the *HNRNPC* knockdown, mutations disrupting U-tracts are likely to be largely neutral in these exons.

Figure 5.12 shows the rates of U-tract elongation and U-tract loss at both the terminal and the linker U-rich regions. Since U-tracts upstream of the exon start are likely to be most important for hnRNP C-repression, we separated single-arm *Alu* exons into right- and left-arm exons. The analysis of statistically significant differences between control and regulated exons depicted in Figure 5.12 yielded several interesting observations:

1. Figures 5.12a and 5.12e reveal the existence of strong positive selection for the elongation of the U-tracts in the linker region. The ancestral length of the long-

est continuous U-tract in the linker region is 6 bp. Based on the explanation above, the frequency of longer U-tracts in the control set is assumed to reflect the neutral frequency of U-tract elongation (roughly 5%). In contrast, hnRNP C-regulated *Alu* exons show a higher rate of elongated U-tracts than neutrally expected. The rate of elongated U-tracts is particularly strong for left-arm *Alu* exons, with more than 25% exons showing elongated U-tracts. Since U-tracts in the linker region are expected to act as splicing signals in left-arm exons, in which hnRNP C blocks the binding of U2AF65, this reflects a strong positive selection for stronger repression by hnRNP C. Weaker, yet significant, positive selection is also observed for right-arm *Alu* exons, consistent with our experimental observation that longer U-tracts in the linker region also strengthen the repression by hnRNP C of right-arm *Alu* exons (*König, data not shown*).

2. Figure 5.12c shows evidence of a stronger conservation of long terminal U-tracts in right-arm *Alu* exons, but not in left-arm *Alu* exons. Again, this observation is consistent with our expectation, as the terminal U-tract is believed to act as a splicing signal only in right-arm *Alu* exons. The differences are however small, as expected given that ancestral terminal U-tracts are naturally long.

3. Figure 5.12d shows that the rate of complete loss of the terminal U-tract is much lower than neutrally expected at regulated right-arm exons, but not at left-arm exons. Again, this is consistent with the action of purifying selection to conserve hnRNP C regulation in the splicing signals of those exons.

4. Interestingly, we did not find any regulated *Alu* exon without a U-tract in the linker region (Figures 5.12b and 5.12f), consistent with the idea that these sequence elements are important for the repression of *Alu* exons. However, this observation is not statistically significant given the low neutral rate of U-tract loss in the linker region and the relatively small sample sizes.

Remarkably, all these observations are strictly consistent with our model of the repression of *Alu* aberrant exonisation by hnRNP C.

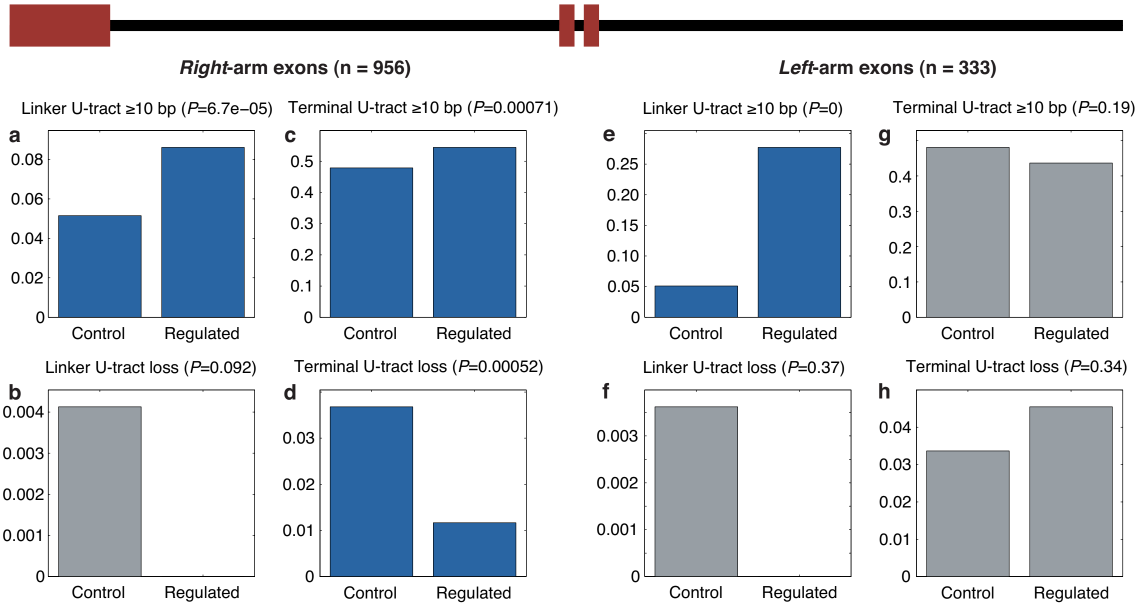

Figure 5.12. The frequencies of unusually long and unusually short U-tracts reveal the action of positive and negative selection on hnRNP C repression of Alu exonisation[16]*. Blue bars are used for statistically significant differences between hnRNP C-regulated exons and their matched set of control exons and grey bars are used for non-significant differences. As in Figure 5.2, right and left arm refer to the standard sense orientation of the canonical Alu sequence.*

The analyses here described reveal surprisingly strong signals of positive and negative selection to strengthen the hnRNP C repression of *Alu* exons. One of the strongest and most striking observations is positive selection promoting the elongation of U-tracts in the linker region. In the analysis above we used an arbitrary threshold of 10 bp for elongated U-tracts in the linker region (compared with 6bp in the ancestral *Alu* element). To generalise this observation, Figure 5.13 shows the frequency of unusually long U-tracts in hnRNP C-regulated *Alu* exons relative to control exons (y-axis) for a wide range of thresholds (x-axis). As the non-exonised (control) exons are assumed to

[16] The study of selection using relative frequencies of unusually long U-tracts in hnRNP C-regulated and control *Alu* exons is based on the assumption that each occurrence of a long U-tract reflects an independent event. However, it is conceivable that an elongated U-tract may have emerged in a mobile *Alu* element and subsequently spread to other parts of the genome. This possibility is, however, unlikely to be relevant since most *Alu* elements derived from a limited number of master copies (evidenced by the existence of over 30 major *Alu* subfamilies), and in all cases their sequence had a short linker U-tract (6 or 7 bp) (Smit et al., 2012). More importantly, however, even if multiple unusually long U-tracts derived from a single ancestral mutation, given that each subsequent *Alu* integration event was independent, the enrichment of long U-tracts in hnRNP C-regulated *Alu* exons would still reflect the action of selection on hnRNP C regulation. Thus, this possibility does not affect the interpretation of our results.

evolve neutrally with respect to hnRNP C regulation, the ratios shown in Figure 5.13 are conceptually analogous to the standard dN/dS ratio for point mutations in coding sequences.

Figure 5.13 consistently shows a strong positive selection for elongated U-tracts in left-arm exons. Longer U-tracts are expected to lead to stronger hnRNP C-binding and so stronger repression of aberrant exonisation. Again, a weaker signal of positive selection is also detectable for right-arm exons, in which U-tracts in the linker region do not act as main splicing signals. At these positions, hnRNP C could interfere with the binding of factors enhancing 5′ splice site usage, such as TIA1/TIAL1 that have been previously reported to act at *Alu* sequences (Forch et al., 2002, Gal-Mark et al., 2009).

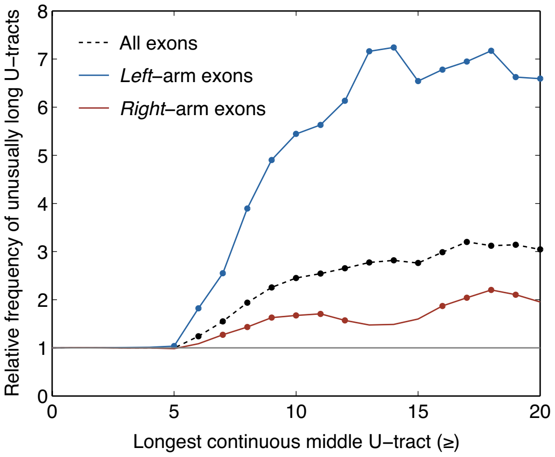

***Figure 5.13**. Relative rates of unusually long U-tracts in the linker U-rich region of intronic Alu elements. The lines show the ratio of the fraction of exons with a U-tract equal or longer than x in the set of regulated exons vs control exons. A ratio higher than 1 suggests the action of positive selection favouring longer U-tracts in regulated exons. The ratios are shown for all hnRNPC-regulated exons (black, n=1869), only left-arm exons (blue, n=333) and only right-arm exons (red, n=956). The control sets are matched non-exonised Alu elements from the same transcripts to the regulated exons. Ratios significantly higher than 1 are indicated with a solid dot (Pearson's chi-squared test P<0.01).*

The detection of a very strong positive selection for stronger repression of *Alu* exons by hnRNP C is surprising, particularly as these are repeated elements located in deep intronic positions, which are typically considered to evolve neutrally. In fact, the strength of positive selection highlights the importance of the threat of aberrant *Alu* exonisation during primate evolution. The results indicate that stronger repression of a single *Alu* exon provided a survival benefit significant enough to allow selection to effectively favour longer U-tracts. As stronger repression of *Alu* exons is not expected to have adaptive value, but merely reduce the fitness cost of aberrant *Alu* exonisation, the observed positive selection demonstrates that the incomplete repression enabled by the shorter ancestral U-tracts had (and probably still has) a substantial health burden.

5.5. Conclusions and discussion

5.5.1. A global protection mechanism for human transcriptome integrity

This study reveals a new role for hnRNP C as a major safeguarding mechanism against *Alu* pseudoexonsation. By performing RNA-sequencing in an *HNRNPC* knock-down cell line we showed that loss of hnRNP C triggers the deleterious incorporation of at least hundreds and likely thousands of *Alu* fragments into mature transcripts. *Alu* pseudoexons most often introduce premature stop codons or cause downstream frame-shifts. A number of disease-causing *Alu* pseudoexons have been described in the literature (Vorechovsky, 2010), highlighting the clinical relevance of *Alu* pseudoexonisation. However, this study unravels the genomic scale of this threat to transcriptome integrity in the absence of hnRNP C.

Using iCLIP, a single-base resolution genome-wide protocol for detecting protein-RNA binding, we observed that hnRNP C prevents pseudoexonisation by directly binding at continuous U-tracts, so preventing their recognition by the splicing machinery. In particular, direct competition between hnRNP C and U2AF65 determines the regulation of splicing at *Alu* elements.

Finally, by performing an evolutionary comparison of the lengths of U-tracts along hnRNP C-regulated and non-regulated *Alu* elements, we were able to detect strong signals of positive and negative selection favouring hnRNP C repression of *Alu* pseudoexonisation. The detection of strong selection against the accidental exonisation of intronic *Alu* elements further demonstrates the evolutionary and clinical relevance of this poorly known threat to transcriptome integrity.

Thus, this study reveals the scale of the threat of *Alu* exonisation to the integrity of the human transcriptome and provides insights into the mechanisms of protection by hnRNP C. Importantly, although we detected *only* around 2,000 *Alu* exons, this number is likely to be much larger as most *Alu* exons are close to the detection limit of the RNA-Seq dataset. Further, although the present study focuses on *Alu* elements -as they represent the largest family of retrotransposons in the human genome- some preliminary evidence suggests that hnRNP C may also act on other groups of retrotransposons, as well as on other pseudoexons not associated with repetitive elements.

A particularly interesting group of retrotransposons in this context is the MIR (*Mammalian Interspersed Repeat*) family, another type of SINE (*Short INterspersed Elements*). A study of 78 disease-causing cryptic exons found 40 related to transposable elements: 11 were derived from *Alu* elements and another 11 from MIRs, despite being five times less common than *Alu* in the human genome (Vorechovsky, 2010). A reanalysis of the hnRNP C iCLIP data revealed comparable levels of binding to *Alu* elements and MIRs, which opens new possibilities for future studies and suggests a possibly even broader role of hnRNP C as a global repressor of retrotransposon pseudoexonisation. Interestingly, U-tracts constitute ideal targets for the silencing of retrotransposon-derived pseudoexons since they are common to all retrotransposons as the complementary polyA tail is required for retrotransposition.

In addition, other splicing repressors such as PTB and hnRNP E1 have been shown to repress pseudoexons in molecular studies of single genes (Dhir and Buratti, 2010). Thus, additional factors may have important roles as global safeguarding mechanisms of human transcriptome integrity. A methodology similar to the one used here may prove invaluable to discover these mechanisms and to unravel their mode of action at a genome-wide scale.

5.5.2. hnRNP C repression of *Alu* pseudoexons may also facilitate evolutionary innovation

More than 650,000 *Alu* elements reside within transcribed regions of the human genome, typically inside introns and UTR regions. While intronic *Alu* elements pose a serious threat in the event of a loss of hnRNP C repression, they are also believed to be a large resource for the evolution of novel functional exons (Sorek, 2007, Sela et al., 2007, Meili et al., 2009, Keren et al., 2010).

Exonisation of intronic sequences is commonly thought to be an important mechanism of evolutionary innovation. An intronic sequence evolving neutrally can develop sequence features sufficient for weak recognition by the splicing machinery, leading to their incorporation in a fraction of transcripts from a given gene. While these pseudoexons are generally non-functional, their leaky presence in some transcripts may allow selection to shape their sequence and eventually give rise to new functional exons.

Given that *Alu* elements are widespread in the human transcriptome and naturally carry key sequence features required for recognition by the splicing machinery, they are thought to be a very important source of potential novel exons. For example, it has been estimated that 5% of alternatively spliced exons in humans derive from *Alu* elements. Further, exonised *Alu* elements are a particularly common source of new exons in humans, as they *Alu*-associated exons are present in 53% of *orphan* (*i.e.* human-specific) genes (Keren et al., 2010).

While the evolutionary potential of *Alu* exonisation has attracted considerable interest, the sudden incorporation of *Alu* sequences into mature transcripts is likely to be deleterious in the vast majority of cases. In this context, the newly described role of hnRNP C as a global repressor of *Alu* exonisation may also have important implications for evolutionary adaptation. In the absence of hnRNP C, suddenly exonised intronic *Alu* elements will most likely be deleterious and selection would favour their removal from the genome, thus limiting their chances to evolve into new functional exons.

On the other hand, in the presence of hnRNP C exonised *Alu* elements can be repressed instead of being removed, allowing them to evolve nearly neutrally for longer evolutionary times. Deleterious *Alu* pseudoexons will tend to be more strongly repressed by hnRNP C, with longer U-tracts. If by chance a pseudoexon becomes less deleterious, for example by losing nonsense mutations, selection against its exonisation will be considerably reduced, which may allow higher levels of leaky exonisation and so stronger evolutionary testing by selection. Eventually, repression of *Alu* exonisation could be completely removed if an exon becomes functional and has adaptive potential.

Thus, repression by hnRNP C may stabilise a large number of *Alu* elements capable of exonisation, facilitating the evolutionary exploration of new functions.

I believe in intuition and inspiration. Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution.

— Albert Einstein, *Cosmic Religion: With Other Opinions and Aphorisms* (1931)

6

Conclusions and future perspectives

6.1. Summary of the main findings

This thesis describes two novel strategies that reduce the impact of deleterious mutations. In contrast to the traditional view of DNA repair, which reduces the rate of mutations indiscriminately, these two strategies specifically reduce the deleterious effect of mutations while maintaining or even fostering their evolutionary potential.

Briefly, the main results of the thesis are:

1. **Non-random mutation rates reduce the risk of spontaneous deleterious mutations in *E. coli*.** Combining phylogenetic and population genetic techniques, we found that the mutation rate varies substantially among genes in the *E. coli* genome. Further, this variation is not random, as more highly expressed genes and genes encoding for proteins under stronger purifying selection experience lower mutation rates. This contradicts the common assumption that mutations occur randomly with respect to their fitness effect, suggesting that second-order selection has shaped mutation rates to reduce the risk of spontaneous deleterious mutations within a genome. Current knowledge of factors influencing the

mutation rate does not explain these observations, suggesting that additional mechanisms must be involved.

2. **Repression of intronic *Alu* elements by hnRNP C is essential to maintain transcriptome integrity in humans.** Combining experimental and computational techniques, we discovered that the RNA-binding protein hnRNP C globally represses the aberrant exonisation of thousands of *Alu* elements. In the context of this collaboration, my work revealed how this accidental exonisation uses a specific set of cryptic splice sites and splicing signals present in the ancestral *Alu* sequence. hnRNP C represses the accidental exonisation of *Alu* elements by strongly binding to their U-tracts, so hiding the cryptic splice sites from the splicing machinery. Finally, we revealed the existence of strong selection during primate evolution to strengthen this repression by hnRNP C, highlighting the scale of the health burden of accidental *Alu* exonisation.

Both studies unravel new insights into the control of genomic and transcriptomic integrity and suggest exciting future avenues of research.

In particular, the non-random occurrence of point mutations in *E. coli* has interesting implications for our understanding of evolution and the control of mutations. First, it highlights that, in contrast to the widely accepted evolutionary tenet, mutations do not have to occur randomly with regard to their fitness effect since second-order selection can effectively act on local mutation rates more frequently than previously believed. Second, the observation that the mutation rate varies functionally along a genome opens two important questions: (a) to what extent this occurs in other organisms and (b) what are the mechanisms behind this phenomenon. Below we describe our ongoing work on both questions.

In addition, the discovery that hnRNP C avoids the destructive exonisation of thousands of intronic *Alu* elements, protecting a large number of human genes, has important implications. In doing so, the study changes our understanding of the scale of the threat of accidental pseudoexonisation and reveals mechanistic insights into the control of *Alu* exonisation. Interestingly, *Alu* repression by hnRNP C may be the first well-characterised example of a broader safeguarding strategy, as other retrotransposons may be similarly repressed (such as MIRs) and other splicing regulators may have a similar role (such as PTB and hnRNP E1). Thus, this study will encourage further investigation into these protection mechanisms and their role in disease.

6.2. Future perspectives

The observations of this thesis, together with the current technological revolution of massively parallel sequencing, open very exciting avenues of research.

6.2.1. A very exciting time

Massively parallel sequencing has triggered a technological revolution that is changing the way we perform biological research and understand biology. First, it has led to a boom in genome resequencing that is revealing an unprecedented picture of the mutational and evolutionary processes acting during human evolution and cancer progression. Second, the flexibility of the technology has also enabled the development of a profusion of functional genomic techniques such as RNA-sequencing, ChIP-sequencing and methylation sequencing, among many others.

A particularly exciting area in the context of this thesis is the sequencing of tumours. During the less than four years since the start of my PhD, dozens of tumours have been deeply sequenced revealing fascinating insights into the origin and control of mutations in cancer; these range from tumour-specific signatures of DNA damage and repair (Pleasance et al., 2010, Rausch et al., 2012, Berger et al., 2012) to the discovery of novel large-scale mutational processes (Stephens et al., 2011).

In the imminent future, similar studies will continue to transform our understanding of somatic mutation and evolution. Excitingly, much remains to be discovered since our knowledge about the relative rates of different somatic mutations in different tissues is still in its infancy (Coufal et al., 2009, Baillie et al., 2011). Further, current analysis pipelines of paired-end sequencing are still far from providing a fully comprehensive picture of all mutational processes, as they neglect important sources of mutations. For example, repetitive regions of the human genome remain poorly explored, including potentially very relevant sequences for disease, such as retrotransposons, telomeres and microsatellites. Similarly, current pipelines are largely insensitive to certain types of mutations such as insertions of novel sequence (Mills et al., 2011).

Also within the context of this thesis, genome resequencing is now transforming population genetics from a traditionally theoretical field into one overwhelmed by empirical data. This is opening new exciting areas of research in human evolution and genetic diseases. As an example, the study of intratumour heterogeneity by massively parallel sequencing has mostly been based on phylogenetic approaches (Campbell et al., 2010, Navin et al., 2011, Gerlinger et al., 2012). In this context, the development of realistic population genetic models of cancer evolution, accounting for limited dispersal and spatial constrains, might become a powerful tool to study selection and cellular migration within tumours.

Thus, the ongoing technological revolution in genomics is already transforming our understanding of mutations in disease, and provides a very exciting environment for research on somatic evolution, mutagenesis and repair.

6.2.2. New research avenues

In the context of this ongoing revolution, the observation of the variation of the mutation rate along the *E. coli* genome opens very interesting questions. First, it remains to be shown to what extent the mutation rate varies along the genomes of other organisms, and to what extent this variation is advantageous. Second, this study suggests that mutagenesis and or DNA repair may vary along the genome, but the contribution of different mechanisms to this remains to be discovered.

6.2.2.1. Functional variation of the mutation rate in humans

One of the most imminent questions stemming from this thesis is whether part of the variation of the mutation rate in humans is non-random. As described in the Introduction of this thesis, the mutation rate appears to vary substantially along the human genome at multiple scales. However, our current understanding of this variation and of the power of second-order selection on local mutation rates suggests that this variation is essentially random with respect to its fitness effect (Hodgkinson and Eyre-Walker, 2011).

Challenging this view, Section 4.4.3 introduces a model by which selection could effectively shape local mutation rates in humans, for example by linking repair pathways to transcription-activating epigenetic marks. In addition, Section 4.4.3 provides some suggestive preliminary evidence using ENCODE data. This preliminary analysis needs to be carefully followed up to revisit the possibility of the functional variation of the mutation rate in humans. To avoid traditional limitations, this analysis should take advantage of the availability of (a) high-quality estimates of synonymous substitution rates between human and chimpanzee, (b) genome-wide population genetic data from the 1,000 genomes project (The 1000 Genomes Project Consortium, 2010), and (c) the comprehensive collection of over 500 epigenetic datasets from the ENCODE Project (Myers et al., 2011). A possible strategy for this is outlined in Section 4.4.3.

This underlines how the rapidly increasing availability of genome-wide population genetic and functional genomic data for humans may change our understanding of the functional variation of the mutation rate along the human genome and shed light on the factors determining this variation.

6.2.2.2. Mutagenesis and repair go genomic

In addition, in my personal opinion, one of the most suggestive outcomes of this thesis is that different mutagenic processes and repair pathways may act differently along a genome. Interestingly, while much is known about these processes at a molecular level, our understanding of them in a genomic context is virtually nonexistent.

However, I believe that this is likely to change within the next few years thanks to the availability of multiple functional genomic tools. In principle, ChIP-sequencing could be used to study the potential heterogeneity of repair activity along a genome. Surprisingly, to the best of our knowledge, this possibility remains to be exploited. Therefore, motivated by our observations in *E. coli*, over a year ago we started an experimental project to explore the potential of ChIP-sequencing to study some of the main DNA repair pathways in *E. coli*:

1. Mismatch repair pathway (target protein MutS)
2. Nucleotide excision repair pathway (target protein UvrB)
3. Error-prone polymerases (target proteins PolIV and PolV)

Monoclonal antibodies against these four *E. coli* proteins are currently being produced at the EMBL-Monterotondo Monoclonal Antibodies Core Facility, and we expect to be able to perform ChIP-sequencing experiments within the next few months. We hope that this analysis will provide new insights into the activity of major repair pathways along along the *E. coli* genome, for which we also have genome-wide expression (Kahramanoglou et al., 2011), methylation (Kahramanoglou et al., 2012) and protein binding data (Kahramanoglou et al., 2011, Prieto et al., 2012).

In addition to ChIP-sequencing, we are currently working on a protocol to map DNA damage at genome-wide scale with high resolution. Being able to quantify the basal rate of mutagenesis along a genome would be a powerful technique in combination with ChIP-sequencing. Our strategy is based on detecting premutagenic lesions by using endonucleases that specifically cleave DNA lesions, followed by *Illumina* sequencing of the breakpoints (see Appendix 2 for a more detailed description of the protocol and the experimental design). This work is inspired by protocols developed two decades ago to measure mutagenesis and DNA repair at single-base resolution along a short sequence of DNA (Pfeifer et al., 1991, Tornaletti and Pfeifer, 1994, Gao et al., 1994, Tu et al., 1996).

Directly quantifying premutagenic DNA damage would have interesting advantages over ChIP-sequencing of DNA repair and over *de novo* discovery of mutations in mutation accumulation cell lines:

1. The quantification of DNA damage before and after DNA repair (which normally acts within a few hours after the damage) would allow to measure separately the rates of mutagenesis and DNA repair (Tornaletti and Pfeifer, 1994, Gao et al., 1994). This could be a powerful technique in combination to ChIP-sequencing of repair enzymes.
2. Contrary to mutations detected in long-term *in vitro* cultures or during cancer development, premutagenic lesions are free from any bias by natural selection.
3. The profile of local mutagenesis and repair could be studied in different genetic backgrounds, allowing to study the role of different factors.

We believe that the development of these techniques could provide an unprecedented view of mutagenesis and DNA repair at a genome-wide scale, facilitating the expansion of these traditionally molecular fields into the genomics era.

Further, the combination of genome resequencing with functional genomic techniques for mutagenesis and repair could transform our understanding of the origin and control of mutations, and the impact of these processes on evolution and disease.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

— Alan Turing, *Computing machinery and intelligence* (1950)

Appendix 1

Evolution of global mutation rates in bacteria and the cost of mutations

A1.1. Introduction

As described in the Introduction of the thesis, understanding what factors determine the evolution of mutation rates has attracted a great interest for many decades. To summarise, the main selective forces believed to drive the evolution of mutation rates are: the cost of deleterious mutations, the cost of fidelity and the need of adaptive mutations. In addition, genetic drift and physicochemical constrains have been suggested to impose a lower limit to the evolvability of mutation rates. Here, the role of these major factors is briefly discussed on the light of some empirical observations and some of the results presented in Chapters 3 and 4.

A1.2. The unexplained universal mutation rate in microbes and the evolution of global mutation rates

In 1991, John W. Drake collected the available information on the mutation rates per base per generation for several DNA microbes, including three types of DNA viruses (bacteriophages M13, λ, T2 and T4), a bacterium (*E. coli*), a yeast (*Saccharomyces cerevisiae*) and a filamentous fungus (*Neurospora crassa*). Comparing their mutation rates, Drake noticed that, despite their genomes sizes varied by ~6,500-fold and their average mutation rates per base per generation (μ) spanned four orders of magnitude, their mutation rate per genome per generation (μL) was constrained to a very narrow range of values from 0.0019 to 0.0046 mutations per genome per division (Drake, 1991).

This striking observation has fundamental implications for our understanding of the evolution of genomic mutation rates. First, as Drake himself noted, this observation alone proves that mutation rates have indeed evolved:

> A common mutation rate for such diverse organisms, whose genomes are composed of either single-stranded (phage M13) or double-stranded DNA, and which include both lysogenic and lytic bacteriophages, bacteria, and microbial eukaryotes, strongly implies that this rate is highly evolved. (Drake, 1991)

Furthermore, this observation provides very valuable clues about the main factors driving the evolution of mutation rates in these organisms:

1. First, the constancy of mutation rates per genome in such a diverse set of organisms strongly suggests that global mutation rates are not determined by the environmental mutagenic pressure.
2. Second, mutation rates are not driven by the intrinsic mutability of the genetic material, since the set of species studied included single-stranded and double-stranded DNA organisms.
3. Third, since the mutation rate per base varies from almost 10^{-6} to less than 10^{-10}, it seems clear that these rates have not evolved to a hypothetical strict physicochemical minimum.
4. Genetic drift can be discarded in these organisms as we discussed in the introduction: the observed mutation rate per genome of 0.003 is several orders of magnitude higher than the drift limit ($\sim 1/N_e$).
5. Finally, the need for adaptive mutations does not seem to be a critical force, since otherwise we would expect mutation rates to be higher in organisms facing more variable environments. Instead mutation rates per genome per generation show very similar values in organisms ranging from phages (under strong co-evolutionary pressure) to environmental species (Drake, 1991, Lynch, 2010).

Thus, Drake's simple observation is invaluable for our understanding of the factors driving the evolution of mutation rates in DNA microbes. Unfortunately, a satisfactory explanation for why the mutation rate per genome in these organisms has evolved to such a universal value remains to be found (Lynch, 2010). The question however is fascinating as it strongly suggests the existence of a simple and elegant explanation for the evolution of global mutation rates.

The fact that mutation rates per genome are much higher than their physicochemical and drift limits suggest that they are indeed in an evolutionary equilibrium. In such a case, two selective forces remain as potential major drivers: (1) the cost of deleterious mutations, which would push mutation rates to lower values, and (2) the metabolic cost of repair, protection and proofreading activities, which would push mutation rates to higher values. Since both factors reflect the action of purifying selection, they have the desirable property of having a roughly similar dependence with the effective population size. This property is likely to be a condition for the convergence of multiple diverse species to a single universal mutation rate. Similarly, both vary linearly with the length of the genome: all other things being equal, a genome twice as long as another will require twice as much investment in repair to maintain the same fidelity and will also suffer a twice as many deleterious mutations per genome.

These two forces, already suggested by Drake (1991), may provide a first step towards understanding the universality of the mutation rate in the face of large variations in the effective population size and the length of the genome. Nevertheless, this still does not clarify why mutation rates should vary around 0.003.

A1.3. The cost of mutations

A1.3.1. Lethal mutations

Since one of the two main factors driving the evolution of global mutation rates is the cost of deleterious mutations, it is important to understand the actual fitness cost associated with a given mutation rate.

In the case of immediately lethal mutations, the fitness cost of mutations corresponds exactly to the loss of mutants per generation. Since mutations are lethal, mutants are removed in every generation and only non-mutants populate the next generation. Thus, the fitness cost (F) of a given mutation rate per genome per generation (μL) where all mutations are lethal corresponds to the expected fraction of mutants in each generation:

$$F_{lethal}(\mu L) = 1 - e^{-\mu L}$$

This approaches $F_{\text{lethal}} \approx \mu L$ for low mutation rates as the chance of individuals suffering more than one mutation is very small. Thus, the loss of fitness due to a mutation rate of 0.003 lethal mutations per genome per generation approaches 0.003. According to the equation, at μL=1 (*i.e.* an average of 1 lethal mutation per genome per generation), ~63% of the offspring will carry one or more lethal mutations. This shows that mutation rates as high as μL=1 will have a hardly bearable cost if most mutations are lethal or nearly lethal. This imposes a higher bound for the evolution of global mutation rates.

A1.3.2. Non-lethal mutations

Contrary to lethal mutations, non-lethal deleterious mutations have a smaller fitness cost at the generation of occurrence: s > -1. However, as mutants are not necessarily removed in each generation, non-lethal mutations can propagate to the next generations. In the presence of free crossover recombination, the linkage between a mutation rate modifier allele and the deleterious mutations will last an average of two generations (Lynch, 2011). Thus, moderately or weakly deleterious mutations will impose a lower cost to the modifier allele than lethal mutations.

On the contrary, bacterial recombination, even at moderately high recombination rates, has a very small effect in the evolution of deleterious mutations (Section 3.4.1). In such case, the fitness cost of a non-lethal deleterious mutation on a modifier allele is extended over many generations. In the end, in the absence of interfering processes, mutations with a population-scaled selection coefficient γ=2$N_e s$<<-1 will be removed by selection, and so their cumulative total fitness cost would approach that of lethal mutations: $F(\mu L)$ = 1-$e^{-\mu L}$. This is in agreement with Lynch's equation of the selective advantage of a mutation-modifier allele in asexual organisms, which is independent of the average fitness effect of mutations (Lynch, 2011).

This can also be shown by simulation. Similarly to the forward simulations described in Section 3.3.1.6, here we use Wright-Fisher forward simulations with multiplicative fitness and Poisson distributed mutation probabilities per genome per generation. For simplicity we assume infinite sites, which should not be a problem given the limited number of mutations and generations simulated. To quantify the fitness loss associated with mutations of different fitness costs, we simulated multiple populations with N=1,000 where all mutations had the same selection coefficient (γ=2Ns).

Figure A1.1 represents the loss of absolute fitness after 1,000 generations in populations evolving with two different mutation rates (μL=0.03 and μL=0.003) and different fitness cost per mutation (γ). The blue dots in the figure represent populations evolving with the universal global mutation rate of 0.003 mutations per generation. They show that, as hypothesised above, even weakly deleterious mutations with γ<-3 have a total fitness cost similar to lethal mutations (and closely following the theoretical equation). As described above, this total cost is the result of a much smaller cost accumulated over multiple generations.

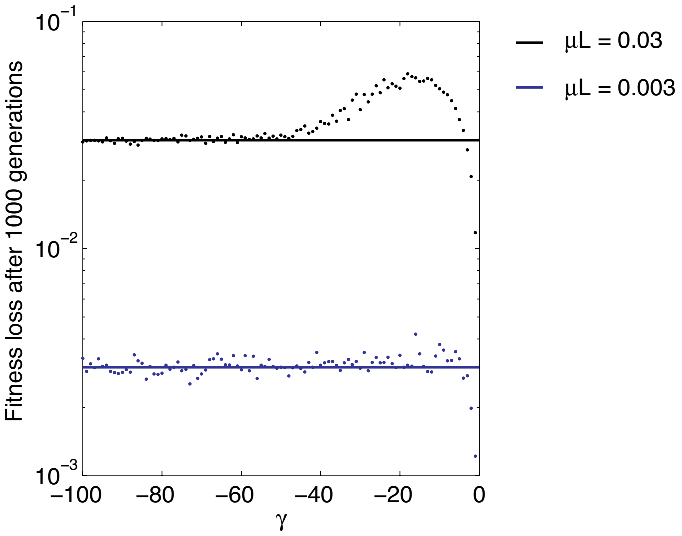

Figure A1.1. *Loss of mean fitness in populations with different genomic mutation rates (μL) and different selection coefficients (γ). To explore the cost of mutations with different selection coefficients, populations were simulated with N=1,000, with all mutations having the same fitness effect (γ). The dots in the figure represent the mean of the average fitness loss in 20 independent populations. The lines represent the expected fitness cost of mutations according to the equation F (μL) = 1-*$e^{-\mu L}$.

Figure A1.1 also evidences the phenomenon of mutational interference (Charlesworth, 2012). As weakly deleterious mutations may require multiple generations to be removed from a population, many individuals in such population will carry one or more mutations. When the mutation rate per genome per generation is sufficiently high, the lifespan of sufficiently weak deleterious mutations will overlap, and they will start to interfere. The black dots in Figure A1.1 evidence this phenomenon. Mutations with a fitness cost of γ>-40 survive long enough to interfere at moderately high mutation rates. If the populations are relatively small, as in the simulations (*N*=1,000), weakly deleterious mutations accumulate to such extent that at a given generation all individuals carry at least one mutation. In the absence of recombination or reverse mutation, selection will not be able to recover wildtype individuals with maximum fitness, which leads to an irreversible progressive loss of fitness in the population. This phenomenon is called Muller's ratchet and has been extensively studied before (Muller, 1964, Felsenstein, 1974, Haigh, 1978, Lynch et al., 1993, Gordo and Charlesworth, 2000, Soderberg and Berg, 2007). Eventually, Muller's ratchet will lead to a mutational meltdown and extinction (Lynch et al., 1993).

The simulations illustrate that, in the absence of mutation interference, the cost of a given mutation rate in bacteria can be approximated by the equation $F(\mu L) = 1\text{-}e^{-\mu L}$, for mutations with a fitness cost γ=2*Ns*<<-1. Strikingly, this suggests that moderately deleterious mutations are as costly as lethal mutations in bacteria. However, this is not strictly true when relaxing certain simplistic assumptions: first, weakly deleterious mutations will last for long periods of time, and so they have a higher chance to be removed by reverse mutation or gene conversion or to be compensated by newly occurring mutations; second, in changing environments, such as those faced by most microbial species, the fitness cost of a mutation will naturally change over time. In the event of even small environmental changes, previously weakly deleterious mutations may be weakly beneficial in the new environment. Thus, in realistic conditions and in the absence of mutational interference, more strongly deleterious mutations are likely to have a larger fitness cost also under strong linkage.

In summary, global mutation rates in bacteria are likely to be in an evolutionary equilibrium driven by the cost of mutations and the cost of fidelity. Increasing mutation rates quickly increases the cost of mutations, according to the equation $F(\mu L) = 1\text{-}e^{-\mu L}$. In addition, two processes can be of relevance to constrain global mutation rates to the observed value of ~0.003 mutations per genome per division. First, increasing the mutation rate one or two orders of magnitude will lead to significant mutation interference, increasing the cost of mutations over the expected value $F(\mu L)$. Second, even if regular selective sweeps are not a major driving force in the evolution of mutation rates in most bacterial species, small changes in the environment may favour populations with a moderate genetic diversity, thus reducing the pressure for lowering mutation rates beyond the already low observed values.

Appendix 2

A genome-wide protocol to study mutagenesis and repair at high resolution

A2.1. Goal

Our goal is to measure DNA damage and repair at a genome-wide scale with high-resolution. Such a protocol would be a powerful tool to study the heterogeneity and dynamics of mutagenesis and repair along a genome. During my PhD, I have devoted several months to this project working in the labs of Prof. Gordon Dougan and Dr. Michael Quail at the Wellcome Trust Sanger Institute.

The initial idea was inspired by several protocols developed in the 1990s, capable of quantifying the rates of DNA damage and DNA repair at nucleotide resolution for short fragments of a few hundred of bases of human DNA (Pfeifer et al., 1991, Tornaletti and Pfeifer, 1994, Gao et al., 1994, Tu et al., 1996). Having explored different alternatives and following discussions with Mike Quail, our current design is based on S1 nuclease cleavage at UV pyrimidine dimers followed by a library preparation protocol inspired on RAD (Restriction site Associated DNA) sequencing (Baird et al., 2008).

S1 nuclease is an endonuclease that cleaves single-stranded DNA, including single-stranded regions at double-stranded DNA such as those caused by a nick, a gap, a mismatch or a loop. As a result, this enzyme generates blunt-ended double-strand breaks at damaged sites such as pyrimidine dimers, a property that has already been used to quantify mutagenesis by some studies (Legault et al., 1997).

A2.2. Brief description of the protocol

The basic steps of the protocol are listed below and are depicted in Figure A2.1:

1. Irradiate *E. coli* cells with UV at the desired dose.
2. Extract DNA using a protocol to maximise its integrity (ideally the average size of DNA fragments should be >50kb). This is followed by phosphatase and nick-closing reactions to minimise the generation of background breaks.
3. S1 nuclease cleavage under stringent conditions. This step should leave phosphorylated blunt-ends ready for ligation at damaged sites. After systematic screening we found a range of conditions that yield high signal to noise ratios.
4. Library preparation following the RAD-Seq protocol. Briefly this involves:
    a. Ligation of a modified P1 adapter to the double-strand breaks created by S1 nuclease at pyrimidine dimers.
    b. Standard *Illumina* library preparation steps: DNA purification, sonication, end-repair and 3′dA overhang addition (A-tailing).
    c. P2 adapter ligation, amplification and *Illumina* sequencing.

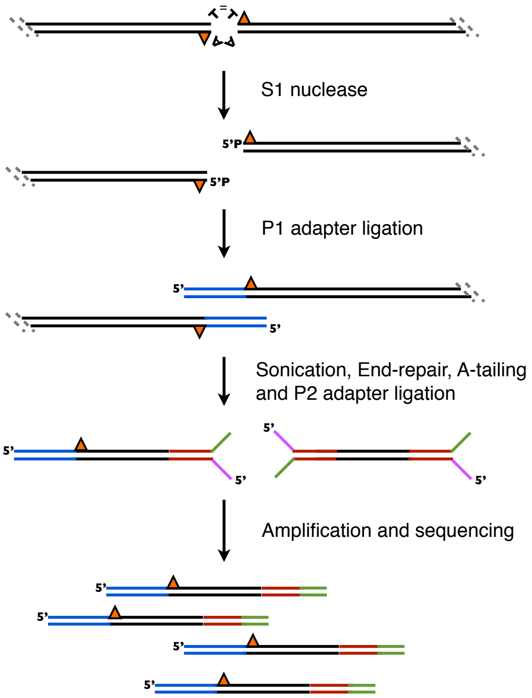
S1 nuclease
5'P
5'P
P1 adapter ligation
5'
5'
Sonication, End-repair, A-tailing and P2 adapter ligation
5'
5'
5'
5'
Amplification and sequencing
5'
5'
5'
5'

***Figure A2.1**. Schematic representation of the current design of the protocol. Briefly, UV-irradiated DNA is cleaved by S1 nuclease producing double-strand breaks at pyrimidine dimers. This is followed by the ligation of the P1 adapter, tagging the site of the damage. This continues with a largely standard Illumina library preparation and sequencing, following the RAD-Seq protocol (Baird et al., 2008). Only fragments carrying the P1 adapter will be sequenced, and so each read will correspond to the site of a pyrimidine dimer.*

At present, we have successfully implemented the first three steps of the protocol and we are working on the ligation of the P1 adapter and on the library preparation.

A2.3. Experimental design

If successful, this protocol will provide genome-wide high-resolution maps of UV damage for an organism. This could provide a powerful genomic tool for the fields of mutagenesis and DNA repair.

Once the protocol is working, we plan to perform the following experiments:

1. UV irradiation of *E. coli* followed by immediate processing to obtain a map of the heterogeneity of UV damage before the action of repair, which usually acts in a matter of hours (Tu et al., 1996).
2. UV irradiation of *E. coli*, incubation for DNA repair and then processing. When analysed together with the map above, this should reveal a map of the activity of DNA repair along the genome.
3. A control library from non-irradiated DNA to correct potential biases, including background cleavage by S1 nuclease and sequencing biases.

In addition, similar profiles could be obtained from mutant strains deficient in certain repair pathways (such as Mfd or UvrA knockouts), to gain additional mechanistic insights.

While this protocol is restricted to UV pyrimidine dimers, in principle, similar protocols could be designed for other types of DNA damage. For example, 8-oxoG can be removed by Fpg leaving a one-nucleotide gap that S1 nuclease should turn into a double strand break. Similarly, other types of damage are specifically cleaved by commercially available DNA glycosilases.

Bibliography

Achaz, G., Coissac, E., Netter, P. & Rocha, E. P. 2003. Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics,* 164**,** 1279-89.

Altenberg, L. 2011. An evolutionary reduction principle for mutation rates at multiple Loci. *Bulletin of mathematical biology,* 73**,** 1227-70.

Amaral, J. D., Xavier, J. M., Steer, C. J. & Rodrigues, C. M. 2010. The role of p53 in apoptosis. *Discovery medicine,* 9**,** 145-52.

Andolfatto, P. 2001. Adaptive hitchhiking effects on genome variability. *Current opinion in genetics & development,* 11**,** 635-41.

Andre, J. B. & Godelle, B. 2006. The evolution of mutation rate in finite asexual populations. *Genetics,* 172**,** 611-26.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology,* 2**,** 2006 0008.

Baer, C. F., Miyamoto, M. M. & Denver, D. R. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature reviews. Genetics,* 8**,** 619-31.

Baillie, J. K., Barnett, M. W., Upton, K. R., Gerhardt, D. J., Richmond, T. A., De Sapio, F., Brennan, P. M., Rizzu, P., Smith, S., Fell, M., Talbot, R. T., Gustincich, S., Freeman, T. C., Mattick, J. S., Hume, D. A., Heutink, P., Carninci, P., Jeddeloh, J. A. & Faulkner, G. J. 2011. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature,* 479**,** 534-7.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. & Johnson, E. A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one,* 3**,** e3376.

Belancio, V. P., Roy-Engel, A. M. & Deininger, P. L. 2010. All y'all need to know 'bout retroelements in cancer. *Seminars in cancer biology,* 20**,** 200-10.

Beletskii, A. & Bhagwat, A. S. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America,* 93**,** 13919-24.

Berg, O. G. & Martelius, M. 1995. Synonymous substitution-rate constants in Escherichia coli and Salmonella typhimurium and their relationship to gene expression and selection pressure. *Journal of molecular evolution,* 41**,** 449-56.

Berger, M. F., Hodis, E., Heffernan, T. P., Deribe, Y. L., Lawrence, M. S., Protopopov, A., Ivanova, E., Watson, I. R., Nickerson, E., Ghosh, P., Zhang, H., Zeid, R., Ren, X., Cibulskis, K., Sivachenko, A. Y., Wagle, N., Sucker, A., Sougnez, C., Onofrio, R., Ambrogio, L., Auclair, D., Fennell, T., Carter, S. L., Drier, Y., Stojanov, P., Singer, M. A., Voet, D., Jing, R., Saksena, G., Barretina, J., Ramos, A. H., Pugh, T. J., Stransky, N., Parkin, M., Winckler, W., Mahan, S., Ardlie, K., Baldwin, J., Wargo, J., Schadendorf, D., Meyerson, M., Gabriel, S. B., Golub, T. R., Wagner, S. N., Lander, E. S., Getz, G., Chin, L. & Garraway, L. A. 2012. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature,* advance online publication.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., Mc Henry, K. T., Pinchback, R. M., Ligon, A. H., Cho, Y. J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M. S., Weir, B. A., Tanaka, K. E., Chiang, D. Y., Bass, A. J., Loo, A., Hoffman, C., Prensner, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F. J., Sasaki, H., Tepper, J. E., Fletcher, J. A., Tabernero, J., Baselga, J., Tsao, M. S., Demichelis, F., Rubin, M. A., Janne, P. A., Daly, M. J., Nucera, C., Levine, R. L., Ebert, B. L., Gabriel, S., Rustgi, A. K., Antonescu, C. R., Ladanyi, M., Letai, A., Garraway, L. A., Loda, M., Beer, D. G., True, L. D., Okamoto, A., Pomeroy, S. L., Singer, S., Golub, T. R., Lander, E. S., Getz, G., Sellers, W. R. & Meyerson, M. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature,* 463**,** 899-905.

Bjedov, I., Tenaillon, O., Gerard, B., Souza, V., Denamur, E., Radman, M., Taddei, F. & Matic, I. 2003. Stress-induced mutagenesis in bacteria. *Science,* 300**,** 1404-9.

Blazquez, J. 2003. Hypermutation as a factor contributing to the acquisition of antimicrobial resistance. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America,* 37**,** 1201-9.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics,* 140**,** 783-96.

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V. & Kazazian, H. H., Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America,* 100**,** 5280-5.

Bull, H. J., Lombardo, M. J. & Rosenberg, S. M. 2001. Stationary-phase mutation in the bacterial chromosome: recombination protein and DNA polymerase IV dependence. *Proceedings of the National Academy of Sciences of the United States of America,* 98**,** 8334-41.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics,* 129**,** 897-907.

Bustamante, C. D., Nielsen, R. & Hartl, D. L. 2003. Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theoretical population biology,* 63**,** 91-103.

Bustamante, C. D., Wakeley, J., Sawyer, S. & Hartl, D. L. 2001. Directional selection and the site-frequency spectrum. *Genetics,* 159**,** 1779-88.

Cai, J. J. 2008. PGEToolbox: A Matlab toolbox for population genetics and evolution. *The Journal of heredity,* 99**,** 438-40.

Cairns, J., Overbaugh, J. & Miller, S. 1988. The origin of mutants. *Nature,* 335**,** 142-5.

Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. A., Morsberger, L. A., Latimer, C., Mclaren, S., Lin, M. L., Mcbride, D. J., Varela, I., Nik-Zainal, S. A., Leroy, C., Jia, M., Menzies, A., Butler, A. P., Teague, J. W., Griffin, C. A., Burton, J., Swerdlow, H., Quail, M. A., Stratton, M. R., Iacobuzio-Donahue, C. & Futreal, P. A. 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature,* 467**,** 1109-13.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution,* 17**,** 540-52.

Chamary, J. V., Parmley, J. L. & Hurst, L. D. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature reviews. Genetics,* 7**,** 98-108.

Charlesworth, B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature reviews. Genetics,* 10**,** 195-205.

Charlesworth, B. 2012. The effects of deleterious mutations on evolution at linked sites. *Genetics,* 190**,** 5-22.

Chen, F., Mackey, A. J., Vermunt, J. K. & Roos, D. S. 2007a. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS one,* 2**,** e383.

Chen, J. M., Cooper, D. N., Chuzhanova, N., Ferec, C. & Patrinos, G. P. 2007b. Gene conversion: mechanisms, evolution and human disease. *Nature reviews. Genetics,* 8**,** 762-75.

Chen, J. M., Cooper, D. N., Ferec, C., Kehrer-Sawatzki, H. & Patrinos, G. P. 2010. Genomic rearrangements in inherited disease and cancer. *Seminars in cancer biology,* 20**,** 222-33.

Chen, M. & Manley, J. L. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews. Molecular cell biology,* 10**,** 741-54.

Choi, J. H., Besaratinia, A., Lee, D. H., Lee, C. S. & Pfeifer, G. P. 2006. The role of DNA polymerase iota in UV mutational spectra. *Mutation research,* 599**,** 58-65.

Chuang, J. H. & Li, H. 2004. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS biology,* 2**,** E29.

Cordaux, R. & Batzer, M. A. 2009. The impact of retrotransposons on human genome evolution. *Nature reviews. Genetics,* 10**,** 691-703.

Coufal, N. G., Garcia-Perez, J. L., Peng, G. E., Yeo, G. W., Mu, Y., Lovci, M. T., Morell, M., O'shea, K. S., Moran, J. V. & Gage, F. H. 2009. L1 retrotransposition in human neural progenitor cells. *Nature,* 460**,** 1127-31.

Cox, E. C. 1972. On the organization of higher chromosomes. *Nature: New biology,* 239**,** 133-4.

Crasta, K., Ganem, N. J., Dagher, R., Lantermann, A. B., Ivanova, E. V., Pan, Y., Nezi, L., Protopopov, A., Chowdhury, D. & Pellman, D. 2012. DNA breaks and chromosome pulverization from errors in mitosis. *Nature,* 482**,** 53-8.

Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. 2004. WebLogo: a sequence logo generator. *Genome research,* 14**,** 1188-90.

Cruet-Hennequart, S., Gallagher, K., Sokol, A. M., Villalan, S., Prendergast, A. M. & Carty, M. P. 2010. DNA polymerase eta, a key protein in translesion synthesis in human cells. *Sub-cellular biochemistry,* 50**,** 189-209.

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

Deaconescu, A. M., Chambers, A. L., Smith, A. J., Nickels, B. E., Hochschild, A., Savery, N. J. & Darst, S. A. 2006. Structural basis for bacterial transcription-coupled DNA repair. *Cell,* 124**,** 507-20.

Denamur, E. & Matic, I. 2006. Evolution of mutation rates in bacteria. *Molecular microbiology,* 60**,** 820-7.

Dhir, A. & Buratti, E. 2010. Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. *The FEBS journal,* 277**,** 841-55.

Drake, J. W. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences of the United States of America,* 88**,** 7160-4.

Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. 1998. Rates of spontaneous mutation. *Genetics,* 148**,** 1667-86.

Drummond, D. A. & Wilke, C. O. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell,* 134**,** 341-52.

Ellegren, H. 2007. Characteristics, causes and evolutionary consequences of male-biased mutation. *Proceedings. Biological sciences / The Royal Society,* 274**,** 1-10.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. & Bernstein, B. E. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* 473**,** 43-9.

Eyre-Walker, A. 1994. Synonymous substitutions are clustered in enterobacterial genes. *Journal of molecular evolution,* 39**,** 448-51.

Eyre-Walker, A. 1996. Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? *Molecular biology and evolution,* 13**,** 864-72.

Eyre-Walker, A. & Bulmer, M. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic acids research,* 21**,** 4599-603.

Eyre-Walker, A. & Bulmer, M. 1995. Synonymous substitution rates in enterobacteria. *Genetics,* 140**,** 1407-12.

Eyre-Walker, A. & Keightley, P. D. 2007. The distribution of fitness effects of new mutations. *Nature reviews. Genetics,* 8**,** 610-8.

Eyre-Walker, A., Woolfit, M. & Phelps, T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics,* 173**,** 891-900.

Felsenstein, J. 1974. The evolutionary advantage of recombination. *Genetics,* 78**,** 737-56.

Fijalkowska, I. J., Schaaper, R. M. & Jonczyk, P. 2012. DNA replication fidelity in Escherichia coli: a multi-DNA polymerase affair. *FEMS microbiology reviews*.

Filion, G. J., Van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., De Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J. & Van Steensel, B. 2010. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell,* 143**,** 212-24.

Fletcher, W. & Yang, Z. 2009. INDELible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution,* 26**,** 1879-88.

Forch, P., Puig, O., Martinez, C., Seraphin, B. & Valcarcel, J. 2002. The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5' splice sites. *The EMBO journal,* 21**,** 6882-92.

Fousteri, M. & Mullenders, L. H. 2008. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell research,* 18**,** 73-84.

Francino, M. P., Chao, L., Riley, M. A. & Ochman, H. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science,* 272**,** 107-9.

Fu, Y. X. & Li, W. H. 1993. Statistical tests of neutrality of mutations. *Genetics,* 133**,** 693-709.

Gaffney, D. J. & Keightley, P. D. 2005. The scale of mutational variation in the murid genome. *Genome research,* 15**,** 1086-94.

Gal-Mark, N., Schwartz, S., Ram, O., Eyras, E. & Ast, G. 2009. The pivotal roles of TIA proteins in 5' splice-site selection of alu exons and across evolution. *PLoS genetics,* 5**,** e1000717.

Gao, S., Drouin, R. & Holmquist, G. P. 1994. DNA repair rates mapped along the human PGK1 gene at nucleotide resolution. *Science,* 263**,** 1438-40.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., Mcdonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A. & Swanton, C. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England journal of medicine,* 366**,** 883-92.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M. S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dose, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., Maccoss, M., Mackowiak, S. D., Mangone, M., Mckay, S., Mecenas, D., Merrihew, G., Miller, D. M., 3rd, Muroyama, A., Murray, J. I., Ooi, S. L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Ratsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science,* 330**,** 1775-87.

Goldman, N. & Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution,* 11**,** 725-36.

Gordo, I. & Charlesworth, B. 2000. The degeneration of asexual haploid populations and the speed of Muller's ratchet. *Genetics,* 154**,** 1379-87.

Gordon, D. J., Resio, B. & Pellman, D. 2012. Causes and consequences of aneuploidy in cancer. *Nature reviews. Genetics,* 13**,** 189-203.

Guarner, F. & Malagelada, J. R. 2003. Gut flora in health and disease. *Lancet,* 361**,** 512-9.

Guindon, S. & Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology,* 52**,** 696-704.

Haigh, J. 1978. The accumulation of deleterious genes in a population--Muller's Ratchet. *Theoretical population biology,* 14**,** 251-67.

Haldane, J. B. 1947. The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Annals of eugenics,* 13**,** 262-71.

Hanahan, D. & Weinberg, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.

Hartl, D. L., Moriyama, E. N. & Sawyer, S. A. 1994. Selection intensity for codon bias. *Genetics,* 138**,** 227-34.

Hayden, E. J., Ferrada, E. & Wagner, A. 2011. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature,* 474**,** 92-5.

Hernandez, R. D. 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics,* 24**,** 2786-7.

Hodgkinson, A. & Eyre-Walker, A. 2011. Variation in the mutation rate across mammalian genomes. *Nature reviews. Genetics,* 12**,** 756-66.

Hodgkinson, A., Ladoukakis, E. & Eyre-Walker, A. 2009. Cryptic variation in the human mutation rate. *PLoS biology,* 7**,** e1000027.

Hoede, C., Denamur, E. & Tenaillon, O. 2006. Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. *PLoS genetics,* 2**,** e176.

Hoege, C., Pfander, B., Moldovan, G. L., Pyrowolakis, G. & Jentsch, S. 2002. RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. *Nature,* 419**,** 135-41.

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J. M. & Tomonaga, K. 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature,* 463**,** 84-7.

Hudson, R. E., Bergthorsson, U., Roth, J. R. & Ochman, H. 2002. Effect of chromosome location on bacterial mutation rates. *Molecular biology and evolution,* 19**,** 85-92.

Hughes, A. L. & Friedman, R. 2005. Nucleotide substitution and recombination at orthologous loci in Staphylococcus aureus. *Journal of bacteriology,* 187**,** 2698-704.

Hughes, A. L., Friedman, R., Rivailler, P. & French, J. O. 2008. Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes. *Molecular biology and evolution,* 25**,** 2199-209.

Ide, H. & Kotera, M. 2004. Human DNA glycosylases involved in the repair of oxidatively damaged DNA. *Biological & pharmaceutical bulletin,* 27**,** 480-5.

Ikemura, T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of molecular biology,* 151**,** 389-409.

Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution,* 2**,** 13-34.

Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of molecular evolution,* 40**,** 190-226.

Janssen, A., Van Der Burg, M., Szuhai, K., Kops, G. J. & Medema, R. H. 2011. Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. *Science,* 333**,** 1895-8.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. & Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research,* 110**,** 462-7.

Kahramanoglou, C., Prieto, A., Khedar, S., Haase, B., Gupta, A., Benes, V., Fraser, G., Luscombe, N. & Seshasayee, A. 2012. Genome-scale analysis of DNA cytosine methylation in Escherichia coli K12 reveals its role in limiting stationary phase transcription. *Nature Communications (In press).*

Kahramanoglou, C., Seshasayee, A. S., Prieto, A. I., Ibberson, D., Schmidt, S., Zimmermann, J., Benes, V., Fraser, G. M. & Luscombe, N. M. 2011. Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic acids research,* 39**,** 2073-91.

Kang, J., Huang, S. & Blaser, M. J. 2005. Structural and functional divergence of MutS2 from bacterial MutS1 and eukaryotic MSH4-MSH5 homologs. *Journal of bacteriology,* 187**,** 3528-37.

Kazazian, H. H., Jr. 2004. Mobile elements: drivers of genome evolution. *Science,* 303**,** 1626-32.

Kent, W. J. 2002. BLAT--the BLAST-like alignment tool. *Genome research,* 12**,** 656-64.

Keren, H., Lev-Maor, G. & Ast, G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews. Genetics,* 11**,** 345-55.

Ketterling, R. P., Vielhaber, E. & Sommer, S. S. 1994. The rates of G:C-->T:A and G:C-->C:G transversions at CpG dinucleotides in the human factor IX gene. *American journal of human genetics,* 54**,** 831-5.

Kim, P. M., Lam, H. Y., Urban, A. E., Korbel, J. O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M. & Gerstein, M. B. 2008. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome research,* 18**,** 1865-74.

Kimura, M. 1960. Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. *Journal of Genetics,* 57**,** 21-34.

Kimura, M. 1967. On the evolutionary adjustment of spontaneous mutation rates. *Genetical Research,* 9**,** 23-24.

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P. & Flicek, P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database : the journal of biological databases and curation,* 2011**,** bar030.

Klapacz, J. & Bhagwat, A. S. 2002. Transcription-dependent increase in multiple classes of base substitution mutations in Escherichia coli. *Journal of bacteriology,* 184**,** 6866-72.

Klapacz, J. & Bhagwat, A. S. 2005. Transcription promotes guanine to thymine mutations in the non-transcribed strand of an Escherichia coli gene. *DNA repair,* 4**,** 806-13.

Kondrashov, A. S. 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *Journal of theoretical biology,* 175**,** 583-94.

König, J., Zarnack, K., Luscombe, N. M. & Ule, J. 2012. Protein-RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics,* 13**,** 77-83.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M. & Ule, J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology,* 17**,** 909-15.

Kooy, R. F. 2010. Distinct disorders affecting the brain share common genetic origins. *F1000 biology reports,* 2.

Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science,* 324**,** 255-8.

Lan, R., Alles, M. C., Donohoe, K., Martinez, M. B. & Reeves, P. R. 2004. Molecular evolutionary relationships of enteroinvasive Escherichia coli and Shigella spp. *Infection and immunity,* 72**,** 5080-8.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., Mcewan, P., Mckernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., Mcmurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., Mcpherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

Lange, S. S., Takata, K. & Wood, R. D. 2011. DNA polymerases and cancer. *Nature reviews. Cancer,* 11**,** 96-110.

Lawrence, J. G. & Ochman, H. 1998. Molecular archaeology of the Escherichia coli genome. *Proceedings of the National Academy of Sciences of the United States of America,* 95**,** 9413-7.

Leclercq, S., Rivals, E. & Jarne, P. 2010. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome biology and evolution,* 2**,** 325-35.

Legault, J., Tremblay, A., Ramotar, D. & Mirault, M. E. 1997. Clusters of S1 nuclease-hypersensitive sites induced in vivo by DNA damage. *Molecular and cellular biology,* 17**,** 5437-52.

Leigh, E. G. 1970. Natural Selection and Mutability. *The American Naturalist,* 104**,** 310-305.

Leigh, E. G. 1973. The evolution of mutation rates. *Genetics,* 73**,** 1-18.

Lenski, R. E. & Mittler, J. E. 1993. The directed mutation controversy and neo-Darwinism. *Science,* 259**,** 188-94.

Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. 2003. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science,* 300**,** 1288-91.

Levins, R. 1967. Theory of fitness in a heterogeneous environment. VI. The adaptive significance of mutation. *Genetics,* 56**,** 163-78.

Li, W. H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *Journal of molecular evolution,* 24**,** 337-45.

Lindahl, T. 1993. Instability and decay of the primary structure of DNA. *Nature,* 362**,** 709-15.

Lindahl, T. & Wood, R. D. 1999. Quality control by DNA repair. *Science,* 286**,** 1897-905.

Ling, A. & Cordaux, R. 2010. Insertion sequence inversions mediated by ectopic recombination between terminal inverted repeats. *PloS one,* 5**,** e15654.

Liu, Y., Prasad, R., Beard, W. A., Kedar, P. S., Hou, E. W., Shock, D. D. & Wilson, S. H. 2007. Coordination of steps in single-nucleotide base excision repair mediated by apurinic/apyrimidinic endonuclease 1 and DNA polymerase beta. *The Journal of biological chemistry,* 282**,** 13532-41.

Loeb, L. A. 2001. A mutator phenotype in cancer. *Cancer research,* 61**,** 3230-9.

Loeb, L. A. 2011. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature reviews. Cancer,* 11**,** 450-7.

Loytynoja, A. & Goldman, N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science,* 320**,** 1632-5.

Lynch, M. 2010. Evolution of the mutation rate. *Trends in genetics : TIG,* 26**,** 345-52.

Lynch, M. 2011. The lower bound to the evolution of mutation rates. *Genome biology and evolution,* 3**,** 1107-18.

Lynch, M., Burger, R., Butcher, D. & Gabriel, W. 1993. The mutational meltdown in asexual populations. *The Journal of heredity,* 84**,** 339-44.

Macdonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N., Macfarlane, H., Jenkins, B., Anderson, M. A., Wexler, N. S., Gusella, J. F., Bates, G. P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A.-M., Lehrach, H., Buckler, A. J., Church, D., Doucette-Stamm, L., O'donovan, M. C., Riba-Ramirez, L., Shah, M., Stanton, V. P., Strobel, S. A., Draths, K. M., Wales, J. L., Dervan, P., Housman, D. E., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J. J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F. S., Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D. & Harper, P. S. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell,* 72**,** 971-983.

Maher, C. A. & Wilson, R. K. 2012. Chromothripsis and human disease: piecing together the shattering process. *Cell,* 148**,** 29-32.

Markham, N. R. & Zuker, M. 2005. DINAMelt web server for nucleic acid melting prediction. *Nucleic acids research,* 33**,** W577-81.

Marteau, P., Pochart, P., Dore, J., Bera-Maillet, C., Bernalier, A. & Corthier, G. 2001. Comparative study of bacterial groups within the human cecal and fecal microbiota. *Applied and environmental microbiology,* 67**,** 4939-42.

Martincorena, I., Seshasayee, A. S. N. & Luscombe, N. M. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature,* 485**,** 95-98.

Massingham, T. & Goldman, N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics,* 169**,** 1753-62.

Maynard-Smith, J. 1991. The population genetics of bacteria. *Proceedings: Biological Sciences,* 245**,** 37-41.

Mcdonald, J. H. & Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature,* 351**,** 652-4.

Mcdonald, J. P., Tissier, A., Frank, E. G., Iwai, S., Hanaoka, F. & Woodgate, R. 2001. DNA polymerase iota and related rad30-like enzymes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences,* 356**,** 53-60.

Mcglynn, P. & Lloyd, R. G. 2002. Recombinational repair and restart of damaged replication forks. *Nature reviews. Molecular cell biology,* 3**,** 859-70.

Mcvean, G. A. & Charlesworth, B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genetics Research,* 74**,** 145-158.

Mcvean, G. a. C., B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genetics Research,* 74**,** 145-158.

Mcvean, G. T. & Hurst, L. D. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature,* 386**,** 388-92.

Mcvey, M. & Lee, S. E. 2008. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in genetics : TIG,* 24**,** 529-38.

Meili, D., Kralovicova, J., Zagalak, J., Bonafe, L., Fiori, L., Blau, N., Thony, B. & Vorechovsky, I. 2009. Disease-causing mutations improving the branch site and polypyrimidine tract: pseudoexon activation of LINE-2 and antisense Alu lacking the poly(T)-tail. *Human mutation,* 30**,** 823-31.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y., Leng, J., Li, R., Li, Y., Lin, C. Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stutz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., Mcvean, G., Sebat, J., Snyder, M., Wang, J., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., Mccarroll, S. A. & Korbel, J. O. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature,* 470**,** 59-65.

Mira, A. & Ochman, H. 2002. Gene location and bacterial sequence divergence. *Molecular biology and evolution,* 19**,** 1350-8.

Miyata, T., Yasunaga, T. & Nishida, T. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 77**,** 7328-32.

Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Current biology : CB,* 4**,** 24-33.

Moxon, R., Bayliss, C. & Hood, D. 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annual review of genetics,* 40**,** 307-33.

Muller, H. J. 1964. The Relation of Recombination to Mutational Advance. *Mutation research,* 106**,** 2-9.

Myers, R. M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R. C., Bernstein, B. E., Gingeras, T. R., Kent, W. J., Birney, E., Wold, B. & Crawford, G. E. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology,* 9**,** e1001046.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., Mcindoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., Mccombie, W. R., Hicks, J. & Wigler, M. 2011. Tumour evolution inferred by single-cell sequencing. *Nature,* 472**,** 90-4.

Nei, M. & Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution,* 3**,** 418-26.

Nevoux, M., Forcada, J., Barbraud, C., Croxall, J. & Weimerskirchi, H. 2010. Bet-hedging response to environmental variability, an intraspecific comparison. *Ecology,* 91**,** 2416-27.

Nielsen, R. & Yang, Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Molecular biology and evolution,* 20**,** 1231-9.

Nilsen, T. W. & Graveley, B. R. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature,* 463**,** 457-63.

O'fallon, B. D. 2010. A method to correct for the effects of purifying selection on genealogical inference. *Molecular biology and evolution,* 27**,** 2406-16.

Ochman, H. 2003. Neutral mutations and neutral substitutions in bacterial genomes. *Molecular biology and evolution,* 20**,** 2091-6.

Odegard, V. H. & Schatz, D. G. 2006. Targeting of somatic hypermutation. *Nature reviews. Immunology,* 6**,** 573-83.

Pal, C., Macia, M. D., Oliver, A., Schachar, I. & Buckling, A. 2007. Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature,* 450**,** 1079-81.

Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcher, C., Basham, D., Chillingworth, T., Davies, R. M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A. V., Moule, S., Pallen, M. J., Penn, C. W., Quail, M. A., Rajandream, M. A., Rutherford, K. M., Van Vliet, A. H., Whitehead, S. & Barrell, B. G. 2000. The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature,* 403**,** 665-8.

Pfeifer, G. P., Drouin, R., Riggs, A. D. & Holmquist, G. P. 1991. In vivo mapping of a DNA adduct at nucleotide resolution: detection of pyrimidine (6-4) pyrimidone photoproducts by ligation-mediated polymerase chain reaction. *Proceedings of the National Academy of Sciences of the United States of America,* 88**,** 1374-8.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., Mcbride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M. L., Ordonez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A. & Stratton, M. R. 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature,* 463**,** 191-6.

Plotkin, J. B. & Kudla, G. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics,* 12**,** 32-42.

Preston, R. J. 2005. Mechanistic data and cancer risk assessment: the need for quantitative molecular endpoints. *Environmental and molecular mutagenesis,* 45**,** 214-21.

Prieto, A. I., Kahramanoglou, C., Ali, R. M., Fraser, G. M., Seshasayee, A. S. & Luscombe, N. M. 2012. Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. *Nucleic acids research,* 40**,** 3524-3537.

Pupo, G. M., Lan, R. & Reeves, P. R. 2000. Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. *Proceedings of the National Academy of Sciences of the United States of America,* 97**,** 10567-72.

Radman, M., Matic, I. & Taddei, F. 1999. Evolution of evolvability. *Annals of the New York Academy of Sciences,* 870**,** 146-55.

Rausch, T., Jones, D. T., Zapatka, M., Stutz, A. M., Zichner, T., Weischenfeldt, J., Jager, N., Remke, M., Shih, D., Northcott, P. A., Pfaff, E., Tica, J., Wang, Q., Massimi, L., Witt, H., Bender, S., Pleier, S., Cin, H., Hawkins, C., Beck, C., Von Deimling, A., Hans, V., Brors, B., Eils, R., Scheurlen, W., Blake, J., Benes, V., Kulozik, A. E., Witt, O., Martin, D., Zhang, C., Porat, R., Merino, D. M., Wasserman, J., Jabado, N., Fontebasso, A., Bullinger, L., Rucker, F. G., Dohner, K., Dohner, H., Koster, J., Molenaar, J. J., Versteeg, R., Kool, M., Tabori, U., Malkin, D., Korshunov, A., Taylor, M. D., Lichter, P., Pfister, S. M. & Korbel, J. O. 2012. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell,* 148**,** 59-71.

Rice, P., Longden, I. & Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG,* 16**,** 276-7.

Riley, B. E. & Orr, H. T. 2006. Polyglutamine neurodegenerative diseases and regulation of transcription: assembling the puzzle. *Genes & development,* 20**,** 2183-92.

Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics,* 27**,** 2325-9.

Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., Washietl, S., Arshinoff, B. I., Ay, F., Meyer, P. E., Robine, N., Washington, N. L., Di Stefano, L., Berezikov, E., Brown, C. D., Candeias, R., Carlson, J. W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M. Y., Will, S., Alekseyenko, A. A., Artieri, C., Booth, B. W., Brooks, A. N., Dai, Q., Davis, C. A., Duff, M. O., Feng, X., Gorchakov, A. A., Gu, T., Henikoff, J. G., Kapranov, P., Li, R., Macalpine, H. K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S. K., Riddle, N. C., Sakai, A., Samsonova, A., Sandler, J. E., Schwartz, Y. B., Sher, N., Spokony, R., Sturgill, D., Van Baren, M., Wan, K. H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S. E., Brent, M. R., Cherbas, L., Elgin, S. C., Gingeras, T. R., Grossman, R., Hoskins, R. A., Kaufman, T. C., Kent, W., Kuroda, M. I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J. W., Ren, B., Russell, S., Cherbas, P., Graveley, B. R., Lewis, S., Micklem, G., Oliver, B., Park, P. J., Celniker, S. E., Henikoff, S., Karpen, G. H., Lai, E. C., Macalpine, D. M., Stein, L. D., White, K. P. & Kellis, M. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science,* 330**,** 1787-97.

Rydberg, B. & Lindahl, T. 1982. Nonenzymatic methylation of DNA by the intracellular methyl group donor S-adenosyl-L-methionine is a potentially mutagenic reaction. *The EMBO journal,* 1**,** 211-6.

Salem, A. H., Kilroy, G. E., Watkins, W. S., Jorde, L. B. & Batzer, M. A. 2003. Recently integrated Alu elements and human genomic diversity. *Molecular biology and evolution,* 20**,** 1349-61.

Schaeffer, S. W. 2002. Molecular population genetics of sequence length diversity in the Adh region of Drosophila pseudoobscura. *Genetical research,* 80**,** 163-75.

Schatz, D. G. & Swanson, P. C. 2011. V(D)J recombination: mechanisms of initiation. *Annual review of genetics,* 45**,** 167-202.

Schlacher, K., Pham, P., Cox, M. M. & Goodman, M. F. 2006. Roles of DNA polymerase V and RecA protein in SOS damage-induced mutation. *Chemical reviews,* 106**,** 406-19.

Schneider, T. D. & Stephens, R. M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic acids research,* 18**,** 6097-100.

Schubert, S., Darlu, P., Clermont, O., Wieser, A., Magistro, G., Hoffmann, C., Weinert, K., Tenaillon, O., Matic, I. & Denamur, E. 2009. Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species. *PLoS pathogens,* 5**,** e1000257.

Seki, M., Gearhart, P. J. & Wood, R. D. 2005. DNA polymerases and somatic hypermutation of immunoglobulin genes. *EMBO reports,* 6**,** 1143-8.

Sela, N., Mersch, B., Gal-Mark, N., Lev-Maor, G., Hotz-Wagenblatt, A. & Ast, G. 2007. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome biology,* 8**,** R127.

Serres, M. H. & Riley, M. 2000. MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. *Microbial & comparative genomics,* 5**,** 205-22.

Sharp, P. M. & Li, W. H. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research,* 15**,** 1281-95.

Sharp, P. M., Shields, D. C., Wolfe, K. H. & Li, W. H. 1989. Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science,* 246**,** 808-10.

Shepard, P. J. & Hertel, K. J. 2009. The SR protein family. *Genome biology,* 10**,** 242.

Smit, A. F. A., Hubley, R. & Green, P. 2012. RepeatMasker Open-3.0. *1996-2012 http://repeatmasker.org.*

Sniegowski, P. D., Gerrish, P. J., Johnson, T. & Shaver, A. 2000. The evolution of mutation rates: separating causes from consequences. *BioEssays : news and reviews in molecular, cellular and developmental biology,* 22**,** 1057-66.

Snyder, L. & Champness, W. 2007. *Molecular genetics of bacteria.* 3rd ed. Washington.

Soderberg, R. J. & Berg, O. G. 2007. Mutational interference and the progression of Muller's ratchet when mutations have a broad range of deleterious effects. *Genetics,* 177**,** 971-86.

Sorek, R. 2007. The birth of new exons: mechanisms and evolutionary consequences. *RNA,* 13**,** 1603-8.

Stankiewicz, P. & Lupski, J. R. 2010. Structural variation in the human genome and its role in disease. *Annual review of medicine,* 61**,** 437-55.

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., Mclaren, S., Lin, M. L., Mcbride, D. J., Varela, I., Nik-Zainal, S., Leroy, C., Jia, M., Menzies, A., Butler, A. P., Teague, J. W., Quail, M. A., Burton, J., Swerdlow, H., Carter, N. P., Morsberger, L. A., Iacobuzio-Donahue, C., Follows, G. A., Green, A. R., Flanagan, A. M., Stratton, M. R., Futreal, P. A. & Campbell, P. J. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell,* 144**,** 27-40.

Strachan, T. & Read, A. P. 2004. *Human Molecular Genetics.* 3rd ed. New York.

Subba Rao, K. 2007. Mechanisms of disease: DNA repair defects and neurological disease. *Nature clinical practice. Neurology,* 3**,** 162-72.

Svejstrup, J. Q. 2002. Mechanisms of transcription-coupled DNA repair. *Nature reviews. Molecular cell biology,* 3**,** 21-9.

Taboada, B., Verde, C. & Merino, E. 2010. High accuracy operon prediction method based on STRING database scores. *Nucleic acids research,* 38**,** e130.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123**,** 585-95.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. & Natale, D. A. 2003. The COG database: an updated version includes eukaryotes. *BMC bioinformatics,* 4**,** 41.

Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. 2010. The population genetics of commensal Escherichia coli. *Nature reviews. Microbiology,* 8**,** 207-17.

Tenaillon, O., Taddei, F., Radman, M. & Matic, I. 2001. Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Research in microbiology,* 152**,** 11-6.

The 1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing. *Nature,* 467**,** 1061-1073.

Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J. & Chen, J. Q. 2008. Single-nucleotide mutation rate increases close to insertions/ deletions in eukaryotes. *Nature,* 455**,** 105-8.

Tornaletti, S. & Pfeifer, G. P. 1994. Slow repair of pyrimidine dimers at p53 mutation hotspots in skin cancer. *Science,* 263**,** 1436-8.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M. E., Frapy, E., Garry, L., Ghigo, J. M., Gilles, A. M., Johnson, J., Le Bouguenec, C., Lescat, M., Mangenot, S., Martinez-Jehanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M. A., Pichon, C., Rouy, Z., Ruf, C. S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Medigue, C., Rocha, E. P. & Denamur, E. 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS genetics,* 5**,** e1000344.

Trapnell, C., Pachter, L. & Salzberg, S. L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics,* 25, 1105-11.

Tu, Y., Tornaletti, S. & Pfeifer, G. P. 1996. DNA repair domains within a human gene: selective repair of sequences near the transcription initiation site. *The EMBO journal,* 15, 675-83.

Vacic, V., Iakoucheva, L. M. & Radivojac, P. 2006. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics,* 22, 1536-7.

Vorechovsky, I. 2010. Transposable elements in disease-associated cryptic exons. *Human genetics,* 127, 135-54.

Wagner, A. 2003. Risk management in biological evolution. *Journal of theoretical biology,* 225, 45-57.

Waisertreiger, I. S., Menezes, M. R., Randazzo, J. & Pavlov, Y. I. 2010. Elevated Levels of DNA Strand Breaks Induced by a Base Analog in the Human Cell Line with the P32T ITPA Variant. *Journal of nucleic acids,* 2010.

Warnecke, T. & Hurst, L. D. 2010. GroEL dependency affects codon usage--support for a critical role of misfolding in gene evolution. *Molecular systems biology,* 6, 340.

Waters, L. S., Minesinger, B. K., Wiltrout, M. E., D'souza, S., Woodruff, R. V. & Walker, G. C. 2009. Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiology and molecular biology reviews : MMBR,* 73, 134-54.

Watson, J., Baker, T., Bell, S., Gann, A., Levine, M. & Losick, R. 2004. *Molecular biology of the gene. 5th ed.,* New York :, Pearson Benjamin Cummings.

Williamson, S. & Orive, M. E. 2002. The genealogy of a sequence subject to purifying selection at multiple sites. *Molecular biology and evolution,* 19, 1376-84.

Wilson, D. J. & Mcvean, G. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics,* 172, 1411-25.

Wilson, T. E., Grawunder, U. & Lieber, M. R. 1997. Yeast DNA ligase IV mediates non-homologous DNA end joining. *Nature,* 388, 495-8.

Wiseman, H. & Halliwell, B. 1996. Damage to DNA by reactive oxygen and nitrogen species: role in inflammatory disease and progression to cancer. *The Biochemical journal,* 313 (Pt 1), 17-29.

Wright, B. E. 2000. A biochemical mechanism for nonrandom mutations and evolution. *Journal of bacteriology,* 182, 2993-3001.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution,* 24, 1586-91.

Ying, H., Epps, J., Williams, R. & Huttley, G. 2010. Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. *Molecular biology and evolution,* 27, 637-49.

Zhai, W., Nielsen, R. & Slatkin, M. 2009. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Molecular biology and evolution,* 26, 273-83.

Zhu, L. & Bustamante, C. D. 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics,* 170**,** 1411-21.

Zhu, L., Wang, Q., Tang, P., Araki, H. & Tian, D. 2009. Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. *Molecular biology and evolution,* 26**,** 2353-61.