# Distribution and evolution of short sequence tandem repeats in eukaryotic genomes

Alice Ledda

TESI DOCTORAL UPF / ANY 2011

DIRECTOR DE LA TESI
Director: Mar Albà Departament Facultat de Ciències de la Salut i de la Vida

Dicono che gli esami non finiscono mai,
le tesi, per fortuna, si!!
A.L.

## Agraments Agraeixo....

# Abstract

Microsatellites are DNA sequences formed by tandem repetition of short motifs. Short sequence tandem repeats are ubiquitous in eukaryotic genomes both in coding and non-coding regions. They show a very high level of polymophism and interspecific divergence.

We investigated the use of next generation sequencing data, from the 1000 Genomes Pilot Prjects, to quantify microsatellite variability in the human population and discover putative new loci involved in trinucleotide repeat expansion diseases.

We analysed microsatellites phylogenetic conservation to learn about the role of selection in shaping microsatellite evolution. The first study concluded that in vertebrate lineages amino acid tandem repeats were more conserved than similar sequences located in non-coding regions. This lead us to the conclusion that evolution was preserving repeats in protein-coding regions. In a second stage we analzed the conservation of microsatellites in different genomic regions, comparing them with the conservation of microsatellite in intergenic region. We concluded that selection was not preserving microsatellites only in exons but also in other genomic regions.

# Resum

Els microsatèl·lits són seqüències d'ADN formades per repeticions en tàndem de motius curts. Les curtes seqüències repetides en tàndem són ubiqües en els genomes dels eucariotes, tant en les regions codificants com en les regions no codificants. Aquestes seqüències tenen un nivell molt elevat de polimorfisme i de divergència interespecífica. Hem investigat si les dades obtingudes mitjançant la seqüenciació de nova generació del Projecte Pilot dels 1000 Genomes són útils per quantificar la variabilitat dels microsatèl·lits en les poblacions humanes i per descobrir nous loci hipotèticament implicats en malalties causades per l'expansió de repeticions de trinucleòtids.

Hem analitzat la conservació fiologenètica dels microsatèl·lits per entendre el rol que juga la selecció en l'evolució dels microsatèl·lits. El primer estudi conclou que en els llinatges dels vertebrats, les repeticions en tàndem d'aminoàcids estan més conservades que altres seqüències similars localitzades a les regions no codificants. Això ens porta a concloure que l'evolució ha mantingut les repeticions a les regions codificants de les proteïnes. En una segona fase hem analitzat la conservació dels microsatèl·lits en diferents regions genòmiques, comparant-les amb la conservació dels microsatèl·lits a les regions intergèniques. Concloem que la selecció no manté només els microsatèl·lits als exons, sinó que també a altres regions genòmiques.

**Prefaci** *In crime movies we are used to see the detective collecting cigarette butts smoked from the suspect. The suspect's DNA extracted from the cigarette butt is then compared with DNA traces found on the crime scene and potentially belonging to the killer. If the DNA matches the suspect is arrested and charged with the crime. Although in real life the entire process is not as fast and as good as it looks in the movies, the movies are based on reality. In real life there are different ways to test whether a sample of DNA belongs to a person. In forensic science you need this test to be fast, cheap and reliable. This is achieved comparing little strings of repetitive DNA which is proved to be very variable among human populations. The comparison of these strings in as much as 13 selected loci of the human genome is enough to say whether two samples of DNA belong to two different individuals not related to each other [Butler, 2006]. Nonetheless the same comparison would not be sufficient to detect the killer if the suspects are father and son or two brothers. In such cases the test has not sufficient statistical power to discriminate and further examinations would be needed. The same test is used for paternity testing.*

*The strings of repetitive DNA are called microsatellites. Due to their characteristic features, microsatellites have been used in molecular biology for a vast range of applications ranging from forensic science and paternity testing as described in the example above, to genetic mapping, tumor and cancer research, linkage and association studies, evolution and as genetic markers in studies of disease mutations. Microsatellites are the subject of this thesis. In the rest of the thesis a summary of everything you wanted to know about microsatellites (and you never dared to ask) is given.*

# Contents

**List of figures** **xv**

**List of tables** **xvii**

**I Introduction** **3**

**1 INTRODUCTION** **5**

1.1 What's a microsatellite? . . . . . 6

1.2 Basic microsatellite mutational mechanisms . . . . . 7

1.3 Microsatellite distribution . . . . . 12

1.4 Amino acid tandem repeats . . . . . 13

1.4.1 Functional microsatellites in proteins . . . . . 14

1.5 Non coding microsatellites . . . . . 19

1.6 Microsatellites and diseases . . . . . 21

1.6.1 Microsatellites and cancer . . . . . 21

1.6.2 Microsatellites and expansion . . . . . 22

1.7 Phylogenetic conservation of microsatellites . . . . . 24

**II Objectives** **27**

**III Results** **31**

**2 MICROSATELLITES AND NEXT GENERATION SEQUENCING** **33**
2.1 Introduction . . . . . 34
2.2 Next Generation Sequencing . . . . . 36
2.3 Work Plan . . . . . 43
2.4 Methods . . . . . 44
2.5 Results . . . . . 45
2.5.1 Pilot study: the individual $NA12878$ . . . . . 45
2.5.2 Analysis of variability in 22 individuals. . . . . 50
2.5.3 Analysis of the disease related repeats . . . . 52
2.6 Discussion . . . . . 55
2.7 Conclusions . . . . . 57

**3 NATURAL SELECTION DRIVES THE ACCUMULATION OF AMINO ACID TANDEM REPEATS IN HUMAN PROTEINS** **59**

**4 PHYLOGENETIC CONSERVATION OF HUMAN MICROSATELLITES** **71**
4.1 Introduction . . . . . 73
4.2 Methods . . . . . 76
4.3 Results . . . . . 80
4.4 Discussion . . . . . 91
4.5 Conclusions . . . . . 95
4.6 Aknowledgements . . . . . 95

**IV Discussion** **97**

**5 DISCUSSION** **99**

5.1 Thesis overview . . . . . 100
5.2 Microsatellites variability . . . . . 101
5.3 Inferring microsatellite functionality . . . . . 103
5.3.1 Inferring microsatellite functionality experimentally . . . . . 103
5.3.2 Inferring microsatellite functionality from a bioinformatics point of view . . . . . 104

**V Conclusions 107**

**VI Appendix 111**

**6 GENOME-WIDE ANALYSIS OF HISTIDINE REPEATS REVEALS THEIR ROLE IN THE LOCALIZATION OF HUMAN PROTEINS TO THE NUCLEAR SPECKLES COMPARTMENT 113**

# List of Figures


1.1 Classification of microsatellites . . . . . 8
1.2 Slippage mechanism . . . . . 10
1.3 The genetic code . . . . . 14
1.4 Dog skulls are influenced by polyQ repeat length. . . 15
1.5 Green Fluorescence protein localizing into speckles . . 17
1.6 FAM76 paralogus proteins . . . . . 18

2.1 Work-flow: Sanger sequencing and Next Generation Sequencing . . . . . 37
2.2 Batch Effect . . . . . 42
2.3 Individual repeat length variation with respect to reference genome repeat length . . . . . 46
2.4 Length distribution in the individual genome for each length in the reference genome . . . . . 48
2.5 Length distribution of the repeats in reference and in the individual's genome . . . . . 49
2.6 Poisson distribution of the DNA fragments. . . . . 55
2.7 A sketch of how coverage affects the assembly of a genome . . . . . 56

4.1 Relative microsatellite abundance depending on genomic location . . . . . 84
4.2 Conserved microsatellites in different genomic regions. 89

# List of Tables

1.1 Functional microsatellites in non coding regions . . . 20
1.2 Major neurodegenerative disorders . . . . . 23

2.1 Main features of the 3 most used next generation sequencing platforms . . . . . 38
2.2 Repeat variability in the 22 individuals sequenced at low coverage . . . . . 51
2.3 Diseases caused by abnormal eleongation of $CAG$ repeats 53
2.4 Variability of 4 repeats involved in neuromuscular disorders . . . . . 54

4.1 Index of microsatellite abbreviations . . . . . 78
4.2 Average length, maximum length and total number of microsatellite repeats . . . . . 79
4.3 Dinucleotide and trinucleotide abundance in the human genome . . . . . 81
4.4 Distribution of microsatellites in different genomic regions . . . . . 83
4.5 Conservation of di- and triNRs in different genomic regions. . . . . 86
4.6 Stability values of different intergenic microsatellites. 88
4.7 Observed and expected conserved microsatellites. . . 90

# List of abbreviations

| | |
|---|---|
| Amino acid | aa |
| Short tandem repeat | STR |
| Simple sequence repeat | SSR |
| Base pair | bp |
| Mono-nucleotide repeat | mono-ntr |
| | monoNR |
| Di-nucleotide repeat | di-ntr |
| | diNR |
| Tri-nucleotide repeat | tri-ntr |
| | triNR |
| Tetra-nucleotide repeat | tetra-ntr |
| | tetraNR |
| Amino acid tandem repeat | AATR |
| Repeat units | ru |
| Nucleotide | nt |
| Point mutation | PM |
| high frequency microsatellite instability | MSI-H |
| Untranslated region | UTR |
| Next generation sequencing | NGS |

# Part I

# Introduction

# Chapter 1

# INTRODUCTION

## 1.1 What's a microsatellite?

Microsatellites are genetic loci where a short motif is tandemly repeated for varying number of times [Katti et al., 2001]. For this reason they are also known as short tandem repeats (STR) [Gemayel et al., 2010] or Simple Sequence Repeats (SSR) [Li et al., 2002a].

The most common definition of microsatellites states that the short motifs range from 1 to 6 base pairs (bps). Tandem repeats of longer motifs ($7 - 100$ bps) are called minisatellites, while tandem repetitions of motifs longer than 100 bps are called satellites. Minisatellites and satellites are believed to evolve in a different way than microsatellites. Depending on the motif length, we speak about mononucleotide repeats (mono-ntrs), dinucleotide repeats (di-ntrs), trinucleotide repeats (tri-ntrs),tetranucleotide repeats (tetra-ntrs), pentanucleotide repeats (penta-ntrs) and hexanucleotide repeats(hexa-ntrs).

Trinucleotide repeats are the most studied class of microsatellites [Subramanian et al., 2003a]. When translated into protein, in fact, they result in an homopolymeric tract and for this reason they are also known as amino acid tandem repeats (AATR). It should be noted that not all the homopolymeric tracts are trinucleotide repeats at the DNA level. For example histidine repeats are mostly encoded by a mixture of the two codons coding for histidines (CAC, CAT), so a histidine tract often it is not a microsatellite [Salichs et al., 2009]. On the other hand glutamine repeats, especially those involved in diseases, are often coded by only one of the two possible codons (CAG,CAA), and so glutamine tracts are often microsatellites at the DNA level [Butland et al., 2007].

There is not much agreement on the minimum length which defines a microsatellite, to be sure that the string considered is not a random assembly of nucleotides. Microsatellites are ideally defined as those short motif repetitive strings that, given the genomic composition in that locus, are statistically significantly rare [Karlin et al., 2002]. The commonly used thresholds to define microsatellites are 4

[Mularoni et al., 2010, Kelkar et al., 2008] or 5 [Huntley and Clark, 2007] repeat units (rus) or 12 nucleotides (nts) [Tóth et al., 2000, Subramanian et al., 2003b], although they depend strongly on what the focus of the article is (sometimes, to study features that can be better seen on longer repeats even 20 rus are used [Katti et al., 2001]).

Depending on the type of repeat sequence, microsatellites can be classified as perfect, imperfect, interrupted, composite and "synonymous codon mixture" [Bhargava and Fuentes, 2010, Mularoni et al., 2010]. An example of each type of repeat is shown in Figure 1.1. A perfect repeat is a stretch of DNA in which a motif of 1 to 6 basepairs is tandemly repeated without any kind of interruptions. In an imperfect repeat there are some bases between the repeated motifs that do not match the motif sequence. In the interrupted repeat instead there is a small sequence within the sequence that does not match the motif sequence. A composite microsatellite is constituted by two distinct adjacent repeated sequences [Bhargava and Fuentes, 2010]. finally a "synonymous codon mixture" repeat is mainly a trinucleotide repeat that at the DNA level does not look like a microsatellite, but due to the codon degeneration, when translated into a protein constitutes an homopolymeric tract [Mularoni et al., 2010].

From now on, unless otherwise specified in the text, when we will talk about microsatellites or repeats we will be always referring to perfect repeats.

## 1.2 Basic microsatellite mutational mechanisms

The mechanism through which microsatellites mutate their length is called replication slippage. This mechanism only applies to microsatellites due to their peculiar structure made by a short motif repeated many times. For this reason microsatellites are also defined as those DNA sequences that undergo replication slippage [Kelkar et al., 2010]. When the DNA strand is opened for replication the

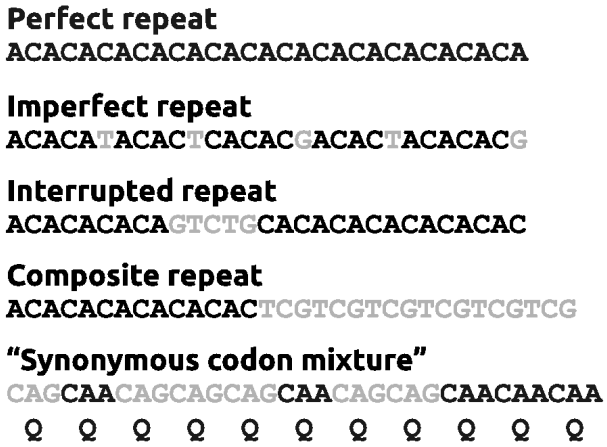



Figure 1.1: **Classification of microsatellites.** An example of perfect repeat, imperfect repeat, interrupted repeat, composite repeat and "synonymous codon mixture" is given. More details on their definition in the text.

short repeats belonging to the same strand may get attached forming some kind of secondary structure (mainly hairpins but also quaternary structures)[Bacolla et al., 2008]. If the loop is formed in the template strand, the resulting copied strand is shorter than the template itself, so the new strand has a shorter repeat than the template. By contrast, if the loop is formed in the growing strand the new repeat is longer than the template. The process is shown is figure 1.2, which is taken from [Gemayel et al., 2010].

Yet the process is not that simple; the repeats of different length produced by replication slippage undergo mismatch replication repair process, which is a cellular process especially designed to correct mistakes that could possibly occur during DNA replication. Many repeats with altered length are suppressed at this stage (for example by protein degradation), thus reducing the effective mutation rate that we are able to see. *In vitro* experiments, where mismatch repair mechanism (MMRM) doesn't work, show mutation rates between 1000 and 100 fold higher than *in vivo* experiments [Strand et al., 1993]. Hence the mutation rate of microsatellites observed *in vivo* is a balance between slippage mutation rate and the efficiency of the mismatch repair mechanism [Schlotterer, 2000].

The observed mutation rate for microsatellites is on average $10^{-4}$ per locus per generation, although depending on the site varies between $10^{-6}$ and $10^{-2}$ [Schlotterer, 2000]. It is much higher than the average point mutation rate which is $10^{-9}$ per locus per generation. It has been shown that the mutation rate is affected by the motif length, the motif sequence, the number of repeats, the purity of the repetition [Chakraborty et al., 1997, Ellegren, 2004, Kashi and King, 2006, Kelkar et al., 2008] as well as by the base stacking strength of the repeat [Bacolla et al., 2008, Woodside et al., 2006].

Still is not clear whether there is a minimum repeat length for replication slippage to start to work. It was commonly believed that the microsatellite had to get to a minimum length to be affected by slippage. This length was determined to be 8 nt independently of the repeat unit length [Rose and Falush, 1998]. A minimum length repeat

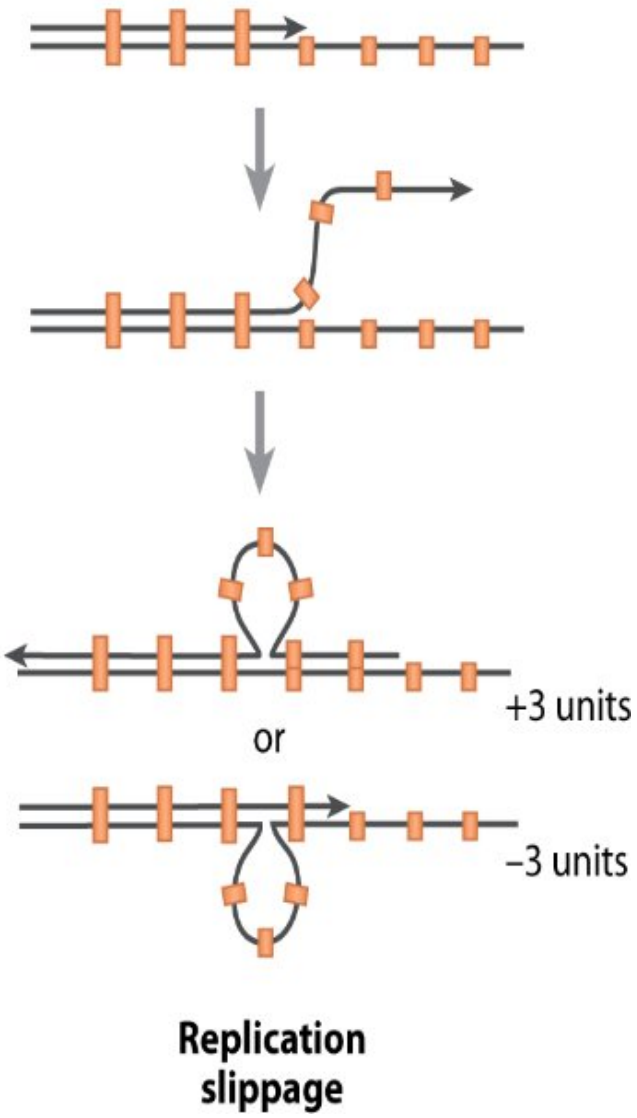



Figure 1.2: **Slippage mechanism.** If the hairpin is created on the template strand the new repeat will be shorter, if the hairpin is on the growing strand the resulting repeat will be longer than the template (taken from [Gemayel et al., 2010])

was created by point mutations in regions of low complexity. The process of creating a repeat long enough for slippage to work happens on evolutionary timescales. Once reached the threshold, the locus became hipervariable, with mutations occurring even on generational timescale [Messier et al., 1996]. Yet a recent article showed that no minimum length is required for slippage to work [Leclercq et al., 2010, Pupko and Graur, 1999]. The issue is still controversial.

Nonetheless, there is a dependence of the slippage mutation rate on the microsatellite length. Longer microsatellites have been shown to be more variable than short ones [Lai and Sun, 2003, Payseur et al., 2010]. The accumulation of point mutations (PM), that prevent microsatellites from an infinite growth [Ellegren, 2000], although there seem to be a bias towards repeat expansion [Rubinsztein et al., 1999].

As a matter of fact, although the main mutation mechanism of microsatellites evolution is replication slippage, point mutations have a very important role. Like the rest of the genome, microsatellites are not immune from PMs. But a PM falling in a perfect microsatellite disrupts the repeat cutting it in two shorter perfect repeats. Point mutations along with replication slippage on the template strand are the two ways for microsatellite shortening, while replication slippage on the new strand is the main mechanism for microsatellite elongation. These observations led to the idea of a "life cycle" for microsatellites [Buschiazzo and Gemmell, 2006], where a microsatellite is born out of a short repeat [Messier et al., 1996], elongated by slippage, then cut by point mutation and shortening slippage, leading again to a short repeat, which can be seen as the death of a microsatellite [Taylor et al., 1999].

For completeness we have to mention that unequal crossing over during recombination has also been proposed as an alternative mechanism of microsatellite evolution. Since unequal crossing over is the main mechanism of evolution of minisatellites and satellites, it is probably involved also in microsatelite evolution [Ellegren, 2004]. To support the hypothesis of unequal recombination being involved in microsatellite evolution there is an over-representation of genes con-

taining repeated sequences within recombination hotspots [Mirkin, 2007]. But repeated sequences are also found to promote recombination. Therefore it is difficult to differentiate causes and consequences in the existing relationship between recombination and microsatellites. Thus the importance of unequal recombination in microsatellites case is still controversial [Haerty and Golding, 2010b].

## 1.3 Microsatellite distribution

Microsatellites are present in one fifth of human proteins [Subramanian et al., 2003a] and are ubiquitous in eukaryotic genomes [Tóth et al., 2000]. Many studies have been done on the distribution of microsatellites across the human genome [Subramanian et al., 2003a, Payseur et al., 2010, Mularoni et al., 2006a] as well as across different taxonomic groups [Tóth et al., 2000, Li et al., 2002b, Sharma et al., 2007a, Katti et al., 2001, Bacolla et al., 2008]. Some common features have been found to be shared among taxa, for example trinucleotide repeats are the most common repeat type inside proteins [Tóth et al., 2000]. This is not surprising: in proteins there is a reading frame and disrupting it may cause harmful missenses in the protein. Trinucleotide repeats do not spoil the reading frame and so they are better integrated into the protein. For the same reason the second most common microsatellite type in proteins of all taxa is hexanucleotide repeat [Tóth et al., 2000].

Nonetheless, taxon specific variations can be detected in the frequency distribution of microsatellites. For example tetranucleotide repeats are more abundant than trinucleotide repeats in introns and intergenic regions of vertebrates, but not in other lineages studied. In invertebrates and fungi, for instance, tetra-ntrs are the least abundant class of micorstellites [Tóth et al., 2000]. An accurate study on 12 *drosophyla* species showed significative differences in the abundance and patterns of repeats, despite the closeness of lineages [Huntley and Clark, 2007]. These are only two examples, many more can be

found in [Bacolla et al., 2008, Katti et al., 2001, Li et al., 2002b] and a review of the best bioinformatics tool to investigate microsatellites in eukaryotic genomes is given in [Sharma et al., 2007a].

A recent study [Haerty and Golding, 2010a], focused on the differences in constitutively and alternatively spliced genes, found out that alternatively spliced genes are enriched in homopolymer sequences encoded by homocodons. Another study found an inverse correlation between long dinucleotide repeat abundance and decreased gene expression. According to the authors this lead to negative selection against dinucleotide repeats longer than 12 repeats, which, in fact, are underrepresented in the human genome [Sharma et al., 2007b].

## 1.4 Amino acid tandem repeats

Microsatellites in coding sequences have often been considered "junk" DNA [Ohno, 1972] or "selfish" DNA [Orgel and Crick, 1980]. That is DNA which had no function in the protein. Nevertheless, a growing number of experiments are assessing the functionality of microsatellites in their wild type state in coding sequences.

In the protein sequence we can have homopolymeric tracts. But amino acid tandem repeats do not always correspond to microsatellites when we look at their DNA code. As the genetic code is degenerate (as shown in figure 1.3), the DNA stretch coding for an homopolymeric tract can be a microsatellite, or not. In this last case, as said in section 1.1 and in chapter 3 we talk about "synonymous codon mixture". Microsatellites with functional effects in proteins are not just rare exceptions, although the relevant literature on them is dispersed across many disciplines, with many studies focused not on microsatellites themselves but rather on the function of a particular gene [Alvarez et al., 2003] or on the genetic bases of a determined phenotype [Kashi and King, 2006]. We will only go into two examples of aminoacid tandem repeat functionality in proteins.These two examples are particularly striking for the severity by which a change

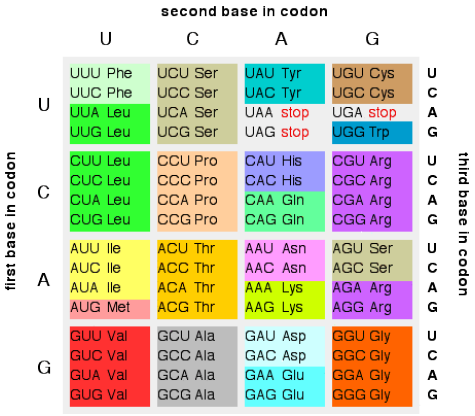



Figure 1.3: **The genetic code.** Apart from tryptophan all the other aminoacid are encoded by more than one codon.

in the repeat length affects the related phenotype. The first is the relation between repeat length and the skull form in dog breeds, the second is the importance of the presence of the histidine repeats for protein localization into nuclear speckles. In both cases amino acid tandem repeats are the focus of the study, not caring whether or not they are or not encoded by microsatellites at the DNA level.

### 1.4.1 Functional microsatellites in proteins

Fondon and Gardner showed that exonic tandem repeat expansion or contraction was a major source of phenotypic variation [Fondon and Garner, 2004]. Repeats in coding regions were studied in 92 different dog breeds. Two of the 5 repeats that showed most variability had particularly striking associated phenotypes.

A deletion of 51 bps (17aa) in a $PQ_n$ repeat in the gene $Alx{-}4$ was found only in one dog breed out of 89 analyzed and was associated with a double dewclaw phenotype present in that dog breed.

Even more striking is the phenotype variation associated to re-

Figure 1.4: **Dog skulls are influenced by polyQ repeat length.**Differences in the dog skull due to different ratios of glutamine repeat length to alanine repeat length in the protein $runx-2$. Picture taken from [Fondon and Garner, 2004].

peat length variation in gene $Runx-2$. In this gene a compound aminoacid repeat of alanine and glutamine is present. A variation in the ratio of the alanine repeat length and the glutamine repeat length is associated with a morphological variation in the dog skull, as shown in figure 1.4. Before the publication of this study, variation in microsatellite length was associated to disease phenotype. In fact polyQ repeats, which were the first aminoacid repeats to be studied, were associated to neurodegenerative diseases. This study showed that variation in microsatellite length could be associated to non-disease phenotype.

Nuclear speckles are nuclear compartments used for storage and modification of proteins. Not much is known about them, the process used to recruit proteins into them is still unclear although it is related to protein charge and phosphorilation [Lamond and Spector, 2003]. Following an article from Alvarez et al [Alvarez et al., 2003], which showed that the histidine tract in the protein DYRK1A was necessary and sufficient to send the protein to nuclear speckles, Salichs et al investigated whether the histidine tract was a general signal to send proteins to nuclear speckles [Salichs et al., 2009].

They showed that an histidine tract of length 6 or more residues is sufficient to target a chimeric protein composed by green fluorescence protein and the histidine tract to nuclear speckles. Their result is shown in figure 1.5. They also showed that the histidine tract is both a necessary and sufficient signal to target proteins having it to nuclear speckles. A particularly striking example is shown in figure 1.6. In the human genome there are two paralogus proteins, FAM76A and FAM76B, which are very similar apart from the histidine tract which is present in FAM76B but not in FAM76A, as shown in the alignment in figure 1.6 a. This only difference prevents the paralogue without the histidine tract to localize in nuclear speckles as shown in the third column of figure 1.6 b.

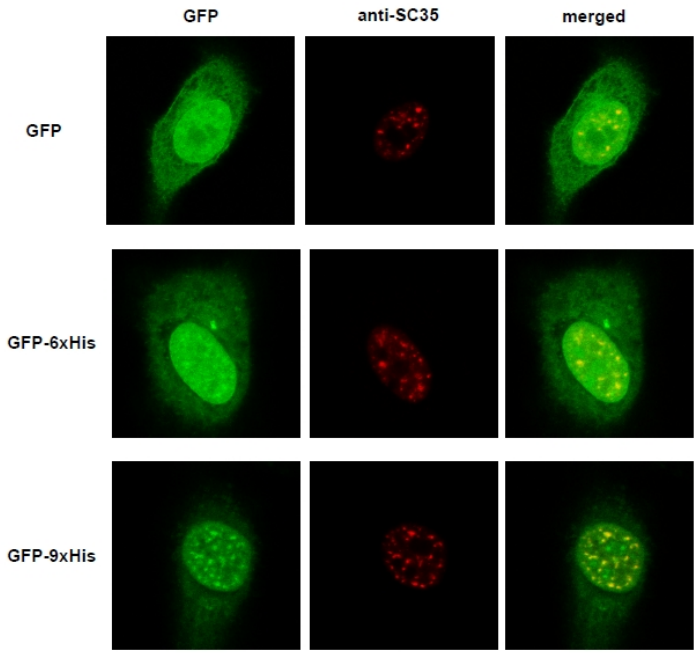
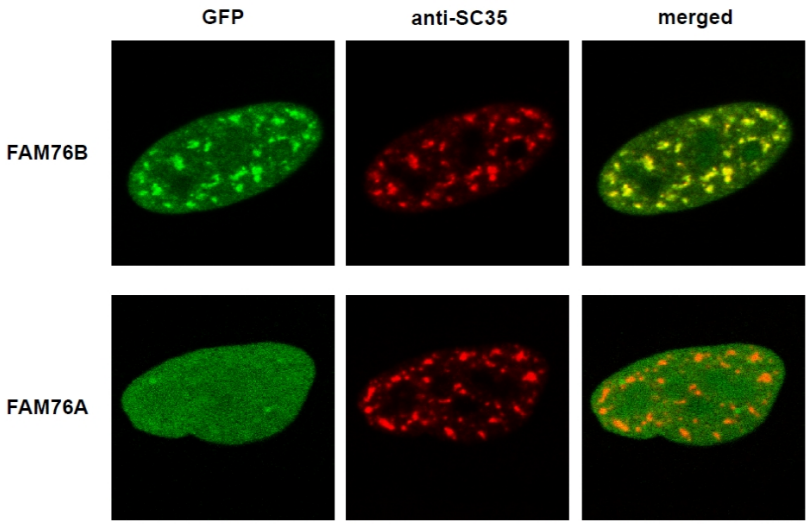



Figure 1.5: **Green Fluorescence protein localizing into speckles.**Localization into nuclear speckles of a chimeric protein formed by GFP and an histidine stretch. GFP alone is not able to localize to nuclear speckle (up), GFP with 6 histidines starts to localize (middle), GFP with 9 histidines show a much stronger signal than with 6 histidines. In the middle column a control with a speckles specific antibody in shown. Picture taken from [Salichs et al., 2009]

A)

```
FAM76B     4   SALYACTKCTQRYPFEELSQGQQLCKECRIAHPIVKCTYCRSEFQQESKTNTICKKCAQN   63
FAM76A     2   AALYACTKCHQRFPFEALSQGQQLCKECRIAHPVVKCTYCRTEYQQESKTNTICKKCAQN   61

FAM76B    64   VKQFGTPKPCQYCNIIAAFIGTKCQRCTNSEKKYGPPQTCEQCKQQCAFDRKEEGRRKVD  123
FAM76A    62   VQLYGTPKPCQYCNIIAAFIGNKCQRCTNSEKKYGPPYSCEQCKQQCAFDRKDD-RKKVD  120

FAM76B   124   GKLLCWLCTLSYKRVLQKTKEQRKSLGSSHSNSSSSSLTEKDQHHPKHHHHHHHHHRHS  183
FAM76A   121   GKLLCWLCTLSYKRVLQKTKEQRKHLSSSSRAGH----QEKEQ----------------  159

FAM76B   184   SSHHKISNLSPEEEQGLWKQSHKSSATIQNETPKKKPKLESKPSNGDSSSINQSADSGGT  243
FAM76A   160   -----YSRLSGGGHYN--SQKTLSTSSIQNEIPKKKSKFESITTNGDSFSPDLALDSPGT  212

FAM76B   244   DNFVLISQLKEEVMSLKRLLQQRDQTILEKDKKLTELKADFQYQESNLRTKMNSMEKAHK  303
FAM76A   213   DHFVIIAQLKEEVATLKKMLHQKDQMILEKEKKITELKADFQYQESQMRAKMNQMEKTHK  272

FAM76B   304   ETVEQLQAKNRELLKQVAALSKGKKFDKSGSILTSP  339
FAM76A   273   EVTEQLQAKNRELLKQAAALSKSKKSEKSGAI-TSP  307
```

(a) Alignment




(b) Localization into speckles

Figure 1.6: **FAM76 paralogous proteins.** The orthologus protein FAM76A and FAM76B. (a) The alignment of the two proteins clearly shows that the two proteins are very similar except for the histidine tract. (b) Only the paralogous with the histidine tract can localize into nuclear speckles. Picture taken from [Salichs et al., 2009].

## 1.5 Non coding microsatellites

Microsatellites effects are not restricted to coding sequences. Repeat number variation exerts a functional influence even when STRs are located in promoter, introns or other non-coding locations where they do not affect protein structure directly.

Simple sequence repeats in 5'UTRs have an important role in differentiating expression profiles between housekeeping and tissue-specific genes. In fact SSR densities in 5'UTR of housekeeping genes are 1.7 times higher than those in tissue-specific genes [Lawson and Zhang, 2008]. Length variation in the $HUMTHO1$ tetranucleotide repeat $(TCAT)_n$ located in the first intron of the $TH$ gene has quantitative effects on the expression of traits associated with the activity of the $TH$ gene [Albanese et al., 2001]. The tetranucleotide $(TTTA)_n$ in the fourth intron of the aromatase gene has been associated to spine features and bone mass [Carbonell et al., 2005]. The dinucleotide repeat $(AC)_n$ located in the promoter region of the matrix metalloprotinease 9 gene is heterogeneous in Japanese population. Length alteration of this dinucleotide repeat causes phenotypic differences in carcinoma cells [Shimajiri et al., 1999]. The polymorphic $(CA)_n$ dinucleotide in the first intron of the epidermal growth factor receptor gene affects the transcriptional activity of the gene. In particular the transcription activity declines with increasing number of $CA$ dinucleotides [Gebhardt et al., 1999].

A brief review of microsatellites in non coding regions that have been experimentally proven to be functional is given in table 1.1. The table contains the motif, the noncoding region in which it is located and the gene. A brief summary of the function and the phenotypic effect is given, as well as the reference to the experimental article in which it was discussed.

| Motif | Location | Gene | SSR function and phenotypic effect | Ref. |
|---|---|---|---|---|
| AC | promoter | Metalloproteinase-9 ($MMP-9$) | length alteration cause phenotypic differences in carcinoma cells | [Shimajiri et al., 1999] |
| CAG | 5'UTR | Human calmodulin-1 gene ($hCALM1$) | Required for HCALM1 full expression | [Toutenhoofd et al., 1998] |
| CTG | 5'UTR | $C/EBP\beta$ | Serve as protein binding sites and regulating gene translation and protein component | [Calkhoven et al., 1994] [Timchenko et al., 1999] |
| TCAT | Intron | Tyrosine Hydroxylase (TH) gene | Acts as transcription regulatory element and relevant to expression of pathogenesis | [Albanese et al., 2001] |
| CA | Intron | Epidermal grow factor receptor ($egfr$) | Enhances $egfr$ transcription and involved in breast carcinogenesis | [Gebhardt et al., 1999] |
| T | Intron | ATM gene | Shorter repeat tract leads to aberrant splicing and abnormal transcription in colon tumor | [Ejima et al., 2000] |
| AT | 3'UTR | Cytotoxic T-lymphocyte antigen 4 ($CTLA4$) | Longer repeats increase susceptibility to Rheumatoid Arthritis | [Rodriguez et al., 2002] |

Table 1.1: **Functional microsatellites in non coding regions.** The motif, the non coding region, the gene as well as the repeat function and its phenotypic effect is given. In the last column a reference to the article in which the phenotypic effect was studied is provided.

## 1.6 Microsatellites and diseases

Microsatellites are involved in a variety of diseases, including many types of cancers and neurodegenerative diseases.

### 1.6.1 Microsatellites and cancer

Cancer is often related to hypervariability in genomic loci. Since microsatellites are hypervariable, it is not surprising that tri-nt repeats containing genes are 5 times more prevalent in cancer genes [Haberman et al., 2008]. Moreover cancer related genes are longer than other genes [Haberman et al., 2008].

Nonetheless hypervariable microsatellite are not directly causative to cancer. Rather they are a side-effect of cancer. Various type of cancers that affect various organs and that are known by the name of Lynch syndrome, are all caused by deficiency of mismatch repair mechanism, which results in high frequency microsatellite instability (MSI-H) and in neoplastic cell evolution [Shah et al., 2010]. These cancers share lots of clinical features with other types of cancers, familiar cancers or cancers of sporadic origin. Tumors with MSI-H exhibit many differences in clinical, pathological and molecular characteristics relative to tumors without it. Diagnosis of Lynch syndrome is associated with a better prognostic factor and could affect the efficacy of chemotherapy. The standard testing procedure consists in the analysis of five microsatellites in the normal tissue and in the cancer tissue. Depending on the degree of similarity of microsatellite length in the normal and cancer tissue, Lynch syndrome can be diagnosed. The microsatellites markers in this case are 2 mononucleotide repeats and 3 dinucleotide repeats, although mononucleotide repeats show less false positives and perform better as a diagnostic tool [Imai and Yamamoto, 2008]. So in Lynch syndrome microsatellite elongation is a side effect of the inactivation of the mismatch repair mechanism, which is the real cause of cancer. On the other hand a simple analysis of the length of microsatellites in five loci

provides a cheap and very powerful diagnostic tool.

### 1.6.2 Microsatellites and expansion

Various repeats are associated with diseases [Mirkin, 2007, Butland et al., 2007, Cummings and Zoghbi, 2000, Ellegren, 2000, Gatchel and Zoghbi, 2005, Lavoie et al., 2003, Ranum and Cooper, 2006, Sumiyama et al., 1996]. In particular their aberrant elongation is associated with a gain of function of the RNA resulting in the disease. In those cases the repeat in its wild type state has no known function itself. Although most of those diseases were associated with trinucleotide repeats [Cummings and Zoghbi, 2000], diseases such as myotonic dystrophy type 2, can also result from expansion of the tetranucleotide repeat (CCTG)@(CAGG) [Ranum and Cooper, 2006]. Most of the microsatellites related to diseases are located in exons [Madsen et al., 2008], for example Huntington disease [Butland et al., 2007] and sympodactilyty [Lavoie et al., 2003]. Nonetheless expandable repeats can be located in various regions of their resident genes, as the case of Fragile X syndrome (FRAXA), located in the 5' untranslated region (UTR), myotonic dystrophy of type 1, located in the 3'UTR, myotonic dystrophy of type 2, Friedereich's ataxia and spinocerebellar ataxia of type 10, located in introns[Cummings and Zoghbi, 2000, Mirkin, 2007, Usdin and Grabczyk, 2000]. Most of the microsatellite-related neuromuscular diseases are autosomal dominant diseases, some of them, like fragile X syndrome, are X-linked and only Friedrich ataxia is autosomal recessive [Gatchel and Zoghbi, 2005].

Although all these disease share the same genetic cause, an aberrant elongation of the microsatellite in the gene, the way this elongation affects the gene, leading to the disease is different: it can be the loss of protein function (LoF) (as in FRAXA), the gain of function of the repeat at the RNA level (GoF), as in HD, an altered protein function or an altered RNA function (aRNAf) [Cummings and Zoghbi, 2000, Gatchel and Zoghbi, 2005, Usdin and Grabczyk, 2000]. A re-

| Disease | Gene | Repeat | Genomic location | Normal length | Pathogenic length | Disease mechanism |
|---|---|---|---|---|---|---|
| Spinocerebellar Ataxia type1 | ATXN1 | CAG | exon | 6-39 | 40-82 | GoF |
| Spinocerebellar Ataxia type2 | ATXN2 | CAG | exon | 15-24 | 32-200 | GoF |
| Spinocerebellar Ataxia type3 | ATXN3 | CAG | exon | 13-36 | 61-84 | GoF |
| Spinocerebellar Ataxia type6 | CACNAIA | CAG | exon | 4-20 | 20-29 | GoF |
| Spinocerebellar Ataxia type7 | ATXN7 | CAG | exon | 4-35 | 37-306 | GoF |
| Spinocerebellar Ataxia type17 | TBP | CAG | exon | 25-42 | 47-63 | GoF |
| Dentatorubral-pallidoluysian atrophy | ATN1 | CAG | exon | 7-34 | 49-88 | GoF |
| Spinobulbar muscular atrophy Kennedy disease | AR | CAG | exon | 9-36 | 38-62 | GoF |
| Huntington Disease | HD | CAG | exon | 11-34 | 40-121 | GoF |
| type1 myotonic dystrophy type2 | CNBP | CCTG | intron1 | $< 27$ | $> 75$ | aRNAf |
| Friederich ataxia | FRDA | GAA | intron1 | $7 - 32$ | $34 - 80$ (pre) $> 100$(full) | LoPF |
| Fragile X syndrome | FMR1 | CGG | 5'UTR | $6 - 53$ | $60 - 200$(pre) $> 230$(full) | LoPF |
| fragile XE syndrome | FMR2 | GCC | 5'UTR | $6 - 35$ | $61 - 200$(pre) $> 200$(full) | LoPF |

Table 1.2: **Major neurodegenerative disorders.** Summary of the mayor degenerative disorders caused by abnormal repeat length expansion. The disease and the gene involved in the disease in given, as well as the repeat whose aberrant elongation is involved in the disease and its genomic location. The repeat normal length and pathogenic length are shown. In the last column the mechanism through which the repeat elongation affects the gene is shown: loss of function (LoF), gain of function at the RNA level (GoF), or altered RNA function (aRNAf). Data from [Cummings and Zoghbi, 2000, Gatchel and Zoghbi, 2005, Mirkin, 2007, Usdin and Grabczyk, 2000]

view of the most common neurodegenerative diseases, with the gene and the repeat involved, the normal and pathogenic repeat length, and the disease mechanism is given in table 1.2.

The two main types of disease related aminoacid repeats are CAG-encoded polyglutamine repeats and polyalanine repeats. Polyglutamine disease related repeats result from an aberrant elongation of a CAG repeat [Butland et al., 2007]. They cause neurodegenerative neuromuscular diseases including Huntington disease and several types of ataxia [Cummings and Zoghbi, 2000]. Polyalanine disease related repeats instead, can be encoded by a mixture of synonymous codons [Lavoie et al., 2003]. They cause developmental diseases, because they affect transcription factors that are expressed during development [Cummings and Zoghbi, 2000].

## 1.7 Phylogenetic conservation of microsatellites

Two genes are called orthologus if they diverged after a speciation event while they are called paralogous if they diverged after a gene duplication event. Both ortology and paralogy are part of a most general definition, homology, which designates a general relationship of common descent between any entities, where the evolutionary scenario is not further specified.

Comparison of orthologus proteins from pairs of related species is a common method used to study the evolution of amino acid tandem repeats in vertebrates [Alba et al., 1999, Hancock et al., 2001, Faux et al., 2007, Mularoni et al., 2007, Mularoni et al., 2008, Simon and Hancock, 2009]. For example a study of inter-specific size variation of aminoacid tandem repeat in human and chimpanzee showed that 17% of the repeats in orthologus human and chimpanzee proteins differ in size in the two species [Mularoni et al., 2008].

Strongly conserved repeats tend to be encoded by a mixture of synonymous codons, while non-conserved repeats are more encoded

by runs of identical codons [Alba et al., 1999, Albà and Guigó, 2004, Mularoni et al., 2007]. This is consistent with a predominant role of replication slippage in the formation of new, nonconserved, repeats.

It has also been observed that well conserved amino acid repeats tend to be embedded in protein regions that evolve more slowly than regions containing non-conserved repeats [Hancock et al., 2001, Faux et al., 2007, Mularoni et al., 2007, Simon and Hancock, 2009]. This bias suggests that repeat evolution is influenced by selection.

A similar approach can be used to study all types of microsatellites, instead of only trinucleotide repeats. Moreover orthologus regions can be retrieved genomewide instead of only in protein coding regions, although retrieving orthologous non-coding regions becomes increasingly more difficult when the distance between species increases. For example Kelkar and coworkers studied genome-wide microsatellite evolution between human and chimpanzee [Kelkar et al., 2008], discovering many details of microsatellites mutagenesis.

A recent study on conservation of human microsatellites in other vertebrate species showed that microsatellites located in exons are more conserved than microsatellites located in other genomic regions. This study also reported that, among non-coding regions microsatellites, those located in UTRs are more conserved than those located in introns and intergenic egions. This points to a functional role of microsatellites in UTRs [Buschiazzo and Gemmell, 2010].

# Part II

# Objectives

The objectives of this PhD thesis can be summarized as follows:

1. Measure the strength of selection in the evolution of aminoacid tandem repeats;
2. Explore the utility of high troughput sequencing to identify microsatellite polymorphisms;
3. Analyse polymorphism in human genes related to polyglutamine repeats, including disease associated repeats;
4. Compare the properties of microsatellites in different genomic regions of the human genome;
5. Study the effects of microsatellite length, composition and genomic location on their conservation in eutherian lineages;

# Part III

# Results

# Chapter 2

# MICROSATELLITES AND NEXT GENERATION SEQUENCING

Next generation sequencing (NGS is the general name to indicate the new sequencing technology that became commercially available in the last years. This technology provides an automated routine to sequence big amounts of data. We used data from the 1000 Genomes Pilot Projects, which were sequenced using NGS platforms, to study CAG variability in the human population. Owing to the lack of data our results were not fully conclusive. Nonetheless we developed a pipeline to deal with data produced from NGS platforms. Moreover we identified the main points that need to be addressed in order to get the best of NGS data in the study of microsatellites: the coverage, which needs to be higher than 15x and the mapping programs, that need to be adapted to the study of microsatellites.

## 2.1 Introduction

Amino acid tandem repeats, also called homopolymeric tracts, are ubiquitous in eukaryotic genomes. Almost one fifth of human proteins contain amino acid tandem repeats [Mularoni et al., 2010]. A peculiar feature of homopolymeric tracts is that they are among the most variable types of DNA sequences in the genome and that their polymorphism derive mainly from variability in length rather than in the primary sequence [Ellegren, 2004].The mutation rate of microsatellites is several orders of magnitude higher than that of unique eukaryotic DNA: while the latter mutates at a rate of approximately $10^{-9}$ per nucleotide per generation, microsatellite mutation rate is often quoted in the range of $10^{-3}$ to $10^{-4}$ per locus per generation [Ellegren, 2000]. Since their discovery in the 1980s, homopolymeric tract have puzzled the scientists: they seemed not to have any function in the protein but still they were so ubiquitous. Yet some amino acid repeats have been described as to perform a function in the proteins that contain them [Fondon and Garner, 2004, Gerber et al., 1994, Lanz et al., 1995, Janody et al., 2001, Galant and Carroll, 2002, Buchanan et al., 2004, Brown et al., 2005].

What was clear since the early 1990s, when the molecular mechanism beneath the Huntington Disease was discovered, is that polyglutamine aberrant elongation is involved in neuromuscular diseases [Usdin and Grabczyk, 2000, Ranum and Cooper, 2006, Gatchel and Zoghbi, 2005]. By now at least 10 such diseases have been discovered [Usdin and Grabczyk, 2000], which share the same DNA mechanism: a $CAG/CTG$ repeat in an exon expands abruptly. The resulting abnormal expansion of the poly-glutamine tract in the protein causes conformational changes that confer toxic properties to the protein. All of these disorders also share similar clinical features which include selective neuronal degradation associated with a neurological phenotype. However their respective causative genes appear to have little functional or structural similarity. For this reason functional genomics approaches for identifying new gene-disease associations

will not be useful [Butland et al., 2007]. Despite recent advances in molecular diagnosis, the majority of clinically diagnosed diseases that share clinical features with the $CAG$ diseases do not have any identified mutation within the known disease associated genes [Ranum and Cooper, 2006]. By studying the $CAG$ encoded polyQ repeats length distribution in around 200 individuals, it has been shown that the disease associated repeats are very polymorphic, they show an higher length variance than most of the analyzed repeats. Nonetheless other CAG repeats in human genes show similar features, making them excellent candidate disease-causing genes to be associated to the neuromuscular diseases that still do not have a molecular cause [Butland et al., 2007]. In this panorama the 1000 Genomes Project [Durbin et al., 2010] provides an unique tool to study the variability of human trinucleotide repeats. The project, in fact, aims at resequencing the entire human genome of 1000 individuals of different geographic origin. With sequences of so many individuals a deep study of variability in human populations will be possible. Alleles as rare as $10^{-2}$ in the human population should be seen in such a study. Such a project should give us the opportunity to identify the most variable repeats and the spectrum of their segregating alleles in a detail never reached before in a genome-wide repeat variability study. In [Andrés et al., 2003, Butland et al., 2007] disease causing CAG repeats were identified to be the most variable ones. Data from 1000 Genomes Project, allowing us to identify the most variable genes, can also be used in this case to predict potentially disease causing genes.

The 1000 Genomes Project [Durbin et al., 2010] is very recent, it started in the end of the year 2008. In the first year of the 1000 Genomes Project the 3 initial pilot projects have been finalized. They will be used to define the final setup of the full project. Each pilot project is different. In pilot 1 180 people of three major geographic groups were sequenced at low coverage. In Pilot 2 two genetically related trios (mother, father and daughter) were sequenced at high coverage. One of the trios was of European origin and the other was of Yoruban origin. Pilot 3 sequenced 1000 genes of 1000 unrelated

people at high coverage. At the moment in which the studies reported in this thesis were performed, only data from the three Pilot Projects was available. New data has meanwhile been made available, which is more comprehensive than the the data in the three Pilot Projects. The studies described in this thesis may have different outcomes if performed on more recent data releases of the 1000 Genomes Project.

## 2.2 Next Generation Sequencing

The 1000 Genomes Project became possible only in recent years, due to the development of new, fast and cheap tools for genome sequencing, known as Next Generation Sequencing (NGS). Since the early 1990's, and until very recent years, DNA sequencing has almost exclusively been carried out with capillarity based Sanger sequencing(i.e. [Lander et al., 2001, Venter et al., 2001, Bovine et al., 2009]), that is considered as first generation technology. Only in 2004 the first next generation sequencing platforms became commercially available. Their release had a major impact on molecular biology: suddenly it become much easier, cheaper and quickier to sequence entire genomes [Shendure and Ji, 2008, Mardis, 2008]. Although several genome-wide studies were completed using only Sanger sequencing, the advent of NGS technologies changed completely the approach to genome sequencing, reducing time, work and people needed to fully sequence entire genomes [Shendure and Ji, 2008, Mardis, 2008]. The broadest application of NGS is the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease.

**First Generation Sequencing.** In the Sanger sequencing technique, DNA to be sequenced is randomly fragmented, then cloned into a high-copy-number plasmid which is then used to transform *E.coli*. The "amplified template" obtained is then sequenced in a

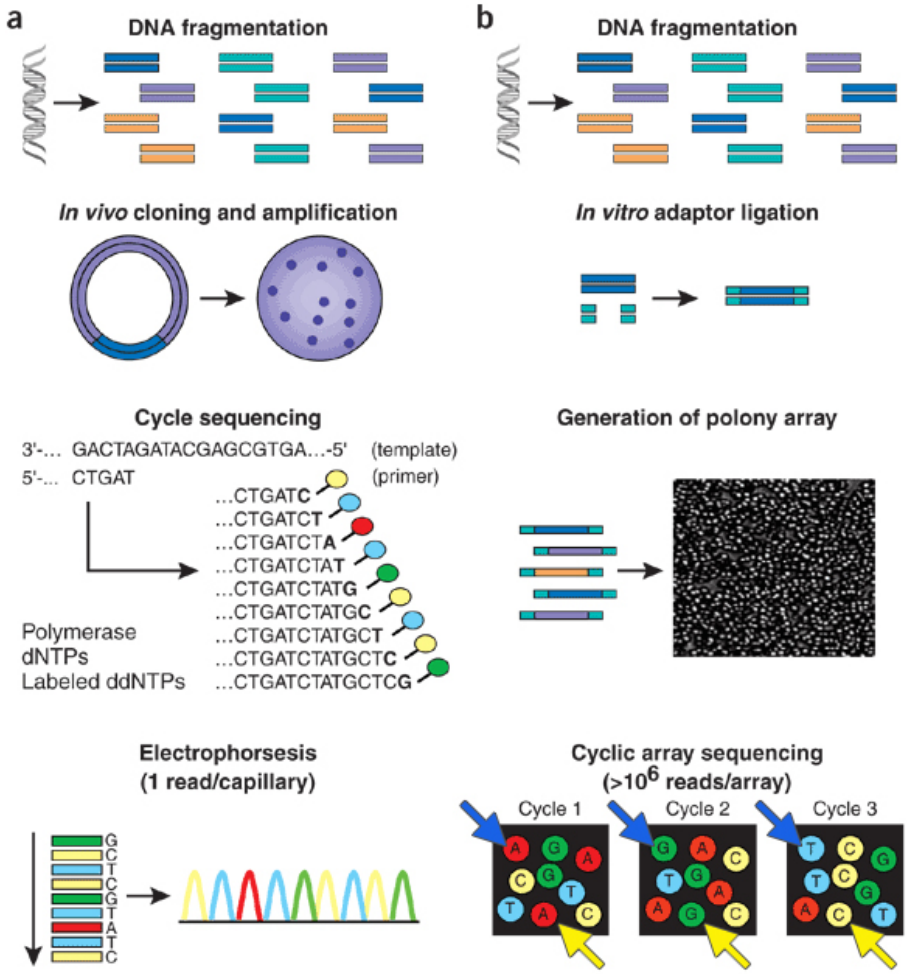



Figure 2.1: **Work-flow: Sanger sequencing and Next Generation Sequencing**(a) Schematic work-flow for Sanger sequencing and (b) for next generation sequencing. Image adapted from [Shendure and Ji, 2008].

| Platform | Library/ template preparation | Read length (bases) | Run time (days) | Gb per run | Machine cost (US$) | Pros | Cons | Biological applications |
|---|---|---|---|---|---|---|---|---|
| Roche/454's GS FLX titanium | frag, MP emPCR | 330<br><br>(245) | 0.35 | 0.45 | 500,000 | Longer reads improve mapping in repetitive regions;fast run times | High reagent cost; high error rates in homo-polymer repeats | Bacterial and insect genome de novo assemblies; medium scale(¡3Mb) exome capture;16S in meta-genomics |
| Illumina/ Solexa's $GA_{II}$ | frag, MP/ solid-phase | 75 or 100<br><br>(36) | 4,9 | 18<br>35 | 540,000 | Currently the most widely used platform in the field | Low multiplexing capability of samples | Variant discovery by whole-genome resequencing or whole -exome capture; gene discovery in metagenomics |
| Life/APG's SOLiD3 | frag, MP/ emPCR | 50<br><br>(35) | 7,14 | 30<br>50 | 595,000 | Two-base encoding provides inherent error correction | Long run times | Variant discovery by whole-genome resequencing or gene discovery in metagenomics |

Table 2.1: **Main features of the 3 most used next generation sequencing platforms:** Illumina (Solexa), SOLiD (Applied biosystems) and 454 (Roche). Adapted from [Metzker, 2009]. The read length into parenthesis is the read length that could be achieved from the first machines in 2004.

"cycle sequencing" reaction. At each round of the reaction the template is stochastically cut and a fluorescent label is incorporated in the terminal position of the truncated template. Sequence is determined by high resolution electrophoretic separation. Laser excitation of fluorescent labels provides the readout that is represented in a Sanger sequence "trace". Software then translates the traces into DNA sequences, and generates error probabilities for each base call as well. The work-flow used is Sanger sequencing is shown in figure 2.1a, taken from [Shendure and Ji, 2008]. For a review on Sanger sequencing see [Metzker, 2005].

**Next Generation Sequencing.** Figure 2.1b shows the work-flow of next generation sequencing. The unique combination of specific protocols distinguishes one technology of each sequencing platform from another. Despite the great differences in template preparation, sequencing process, imaging and in outcomes (i.e. the read length), all NGS sequencing platform share the same work-flow. The library is prepared by random fragmentation, followed by *in vitro* ligation of common adaptor sequences. Then the library is amplified through PCR. The sequencing process itself consists of alternating cycles of enzyme driven biochemistry and imaging based data acquisition. In the three consectutive cycles shown in the bottom part of figure 2.1b, in the locus indicated by the yellow arrow the machine will read A,G,C, while in the locus indicated by the blue arrow the machine will read A, G, T.

**Read Length.** At the moment there are 3 widely used next generation sequencing platforms: 454 (Roche), Solexa (Illumina) and SOLiD (Applied biosystems). There are lots of differences among them and discussing all of them is beyond the purposes of this thesis. Very good reviews of NGS platforms are [Mardis, 2008, Metzker, 2009, Shendure and Ji, 2008]. A summary of the main features of the three most used next generation sequencing platforms is given in table 2.1. There is one key feature that makes a difference among

NGS platforms when studying microsatellites: the read length. Although the read length of each platform is improving with time, here we just focus on typical read lengths at the time the 1000 Genomes Pilot Projects were performed [Durbin et al., 2010]. More details about the read length in the different phases of the next generation sequencing machines is given in column 3 of table 2.1. Solexa and SOLiD at that time had a read length of 36 and 35 basepairs (bps), too short as you need to cover the complete microsatellites (a trinucleotide repeat with 12 repeating units will be 36 nt long), as well as at least a few basepair on both sides of the repeat in order to map correctly the repeat to the genomic region it belongs to. Therefore with Solexa and SOLiD it was impossible to study repeats longer than 10 trinucleotide repeat units. The $CAG$ repeats involved in neuromuscular diseases studied in table 2.3 span from 10 rus to 38 rus [Cummings and Zoghbi, 2000].

On the other hand, when the 1000 Genomes Pilot Project data was sequenced [Durbin et al., 2010], 454 had a read length of 245 bps [Harismendy et al., 2009], which allowed us to study repeats up to tens of repeat units. Unfortunately 454 platform has some problems with homopolymeric runs that are intrinsic to the technology used by this platform, known as pyrosequencing. In pyrosequencing each incorporation of a nucleotide by DNA polymerase results in the release of a pyrophosphate, which initiates a series of downstream reactions which ultimately produce light. The amount of light is proportional to the number of nucleotides incorporated. The 454 base calling software is calibrated to detect the light emitted by each nucleotide incorporation. However the calibrated base calling cannot properly interpret long stretches ($> 6$) of the same nucleotide, so these areas are prone to base insertion and deletions during base calling. By contrast very low rates of substitution errors are reported for 454 sequence reads [Mardis, 2008]. As reported in table 2.1 and in [Metzker, 2009], a Solexa machine with read length up to 100 bps is already available, it can be used to sequence trinucleotide repeats up to almost 30 rus. Still no repeat data obtained with this platform has

been made publicly available, so no comparison with the 454 machine is possible at the moment.

**Sources of errors.** There are three main sources of errors that affect sequences generated with NGS technologies: sequence errors, assembly errors and batch effects [Pool et al., 2010]. The severity of these problems depends in part from the depth of sequencing, with high coverage (> 15x) potentially minimizing many errors [Bentley et al., 2008, Harismendy et al., 2009].

We have already discussed the potential sources of sequencing errors for the repeat data we will analyze later. The assembly errors pose mainly a bioinformatics problem. In fact this is a major difference between NGS and sanger sequences, which has not yet been fully addressed from a bioinformatics point of view. The reads produced with NGS platforms, even with 454, are much shorter than the ones produced by Sanger sequencing. The short read length makes it difficult for the alignment software to align repeat regions, resulting in the lack of such repetitive regions in the final assembly [Pop and Salzberg, 2008]. Moreover it is common practice to discard reads with ambiguous placements. Since repeats are often part of simple sequences, which frequently are ambiguously placed, repeat containing reads are routinely discarded. And this causes missing data.

This is a very important source of error when analyzing repeats, especially to study repeat's variability, because it artificially reduces the variability that can be seen. The batch effect [Leek et al., 2010] is another important source of error, especially when comparing data coming from different laboratories or processed in different days, as in the case of the 1000 Genomes Project. In fact little data processing differences like the reagents used can have a great impact on the derived data. A very instructive picture is shown in figure 2.2, taken from [Leek et al., 2010]. In figure 2.2 it is shown the variation in coverage depth with the day in which data is processed. This difference increases the difficulties in analyzing repeats variability.

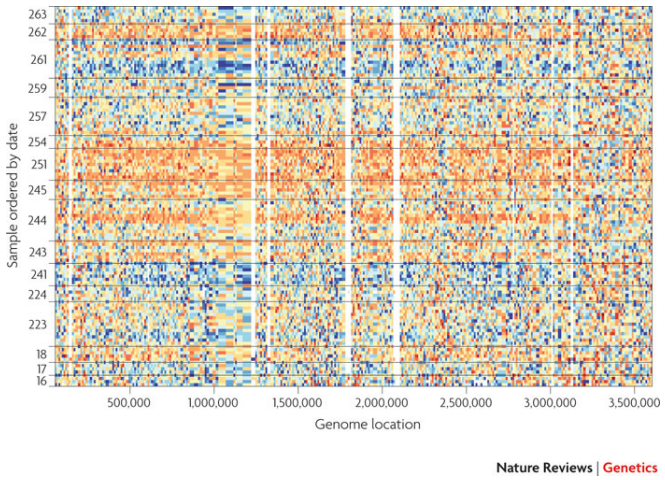



Figure 2.2: **Batch Effect** An example of batch effect in 1000 genomes project: A region of chromosome 16, sequenced in different individuals (each line is a different individual), the black horizontal lines separate the different processing days. Blue represents three standard deviations below average and orange represents three standard deviations above average. There is a clear dependence of the coverage on the processing days. The largest batch effect occurs between days 243 and 251.

## 2.3 Work Plan

Our purpose was to study the variability of tri-nucleotide repeats genome-wide in the human population using data provided from the 1000 Genomes Pilot Projects. Such a study was not even conceivable few years ago, before the advent of next generation sequencing technologies. Nonetheless it still might be a very complex study and very computationally expensive too. Moreover the data from the 1000 Genomes Pilot Projects was a new type of data, obtained with NGS techniques and subject to not well studied errors (see previous section 2.2 and [Pool et al., 2010]). Therefore, there was much uncertainty in the outcome of the study. Furthermore we couldn't use all the data from the 1000 Genomes Pilot Project, because, as explained in the technical introduction 2.2, only the 454 platform has a repeat length suitable for the study of repeats.

For this reason we decided to perform a pilot study on a subset of human repeats. In a recent study from Butland et al [Butland et al., 2007], variability of $CAG$ repeats was studied deeply by sequencing 200 individuals in order to predict possible disease causing repeats. We first decided to perform a study on the variability of $CAG$ repeats in the data from Pilots 1 and 2 of the 1000 Genomes Project, using [Butland et al., 2007] as a comparison.

Therefore we started by performing only a preliminary partial analysis focusing only on one of the best sequenced individuals of the Pilot 2 of the 1000 Genomes Project, NA12878. We first generated a database of $CAG$ encoded glutamine repeats in the human proteome, extracted from the human reference genome NCBI 36. We looked for these repeats in individual $NA12878$, being able to find almost 90% of them. We studied the length distribution of the repeats found in the individual and the repeats variability comparing their length with the length they had in the human reference genome. We also studied in particular the disease causing repeats comparing the variability observed in this individual and in the other individual from the Pilot 2 of the 1000 Genomes Project ($NA19240$) with the variability stated

in Butland's work.

Since the results of this pilot study looked very promising, we decided to perform a study of variability in all the individuals sequenced with the 454 platform in the Pilot 1 of the 1000 Genomes Project. In the Pilot 1 of the 1000 Genomes Project 180 individual of three major geographic groups were sequenced at a coverage of 2x-4x, see 2.1 and [Durbin et al., 2010]. We analyzed the data relative to the 24 individuals sequenced with this platform with the same programs we developed during the first part of our pilot study. The results of this second part were not as good as expected, most of the repeats that we needed were missing, or there were sequencing mistakes. The differences in coverage severely affected the outcome of our analysis.

## 2.4 Methods

In the first phase of the study we focused only on the individual $NA12878$, a female of European origin sequenced at a coverage of 20x.

The 1000 Genomes Project is producing an extraordinary amount of data. To store it they created a new file format, the .bam format. This is a binary format, which needs to be opened with a special informatic tool kit called Samtools [Li et al., 2009]. We downloaded the .bam files of the individual and opened it using Samtools. We isolated all the glutamine repeats located in exons using homemade programs. The repeat had to be completely embedded in a read and at least 4 nucleotides apart from the read end. We had previously retrieved a non redundant list of all the glutamine homopolymeric tracts of at least 4 aminoacids encoded by the $CAG$ codon present in the human reference genome NCBI 36. For each repeat we kept its start and end coordinates as well as its length. We compared the repeats found in the analyzed individual with our database We were able to find 2695 of the 3087 repeats in the database, that were long 4 or more aminoacids in the individual. To avoid sequencing errors, we

applied a filter that kept only events that occurred at least twice in different sequencing reads. For example if a repeat of length 10 in the reference genome appeared only once (one read) in the individual's genome with length 4 we deleted it because we could not be sure it was not a sequencing mistake.

To study the feasibility of the variability study we downloaded the data of the other individual sequenced with 454 platform at a coverage of 20x ($NA19240$). This was a Yoruban female.

In the second phase of the project we downloaded all the data relative to all the individuals sequenced with the 454 platform. The individuals are the HapMap project individuals identified with the following codes: $NA07346$, $NA07347$, $NA11849$, $NA11881$, $NA11894$, $NA11918$, $NA11931$, $NA12043$, $NA12045$, $NA12234$, $NA12249$, $NA12287$, $NA12812$, $NA12814$, $NA12815$, $NA12872$, $NA12873$, $NA12874$, $NA18969$, $NA18970$, $NA19141$, $NA19143$. They were all of European origin except the last four, who are African $NA19xxx$ or Japanese $NA18xxx$. All of them were sequenced at a coverage ranging from 2x to 4x. The data of each one of those individuals was processed with the same programs we processed the data of $NA12878$.

All the statistical analysis was performed using R tools. R was also used to draw all the graphs. To check repeat overlapping we also used the package fjoin [Richardson, 2006].

## 2.5 Results

### 2.5.1 Pilot study: the individual $NA12878$

We first investigated the repeat length distribution of the $CAG$ repeats in the individual $NA12878$. In particular we studied the repeat length variation in the individual with respect to the reference genome length using the plot in figure 2.3.

In the plot is shown the repeat length variation in the individual genome with respect to the repeat length in the reference genome. The colors represent the natural logarithm of the abundance of the

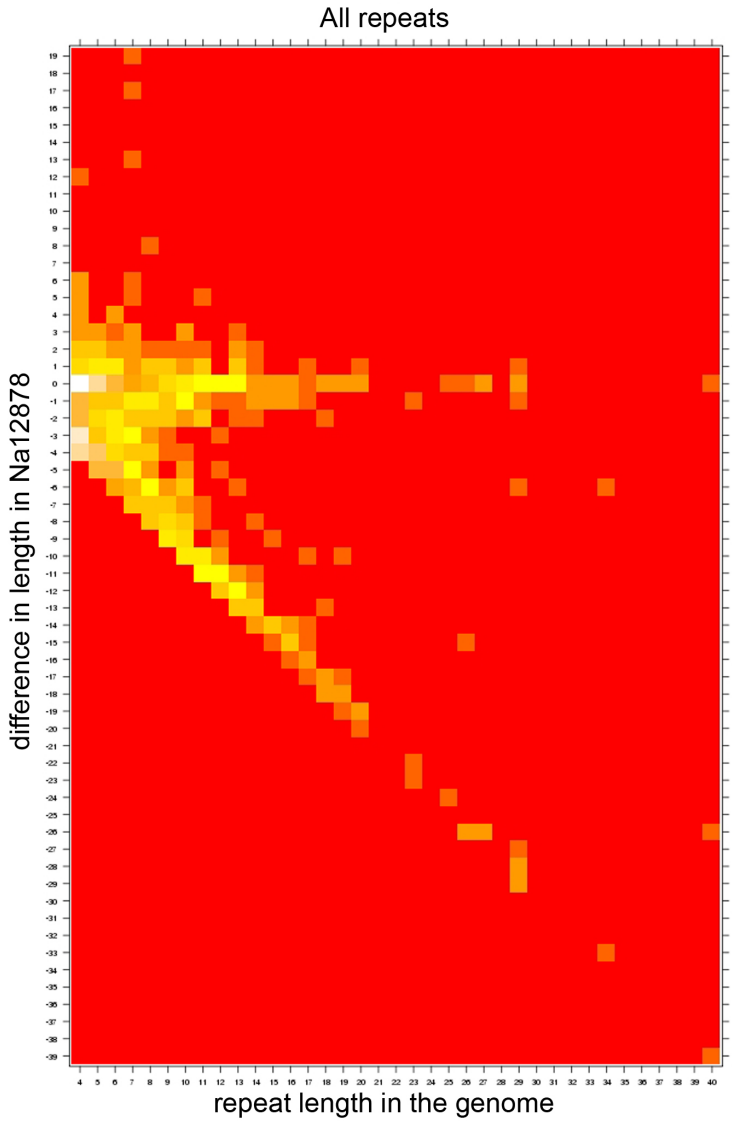



Figure 2.3: **Individual repeat length variation with respect to reference genome repeat length.** We can see clearly the two distributions governing the repeat distribution: the horizontal yellowish line at $y = 0$ is the one with repeats showing the same length as in the reference genome; the diagonal yellowish line is the one in which the repeat length in $NA12878$ is 0. Further details are given in the text.

repeat type. The highest abundance is white, while the lower abundance (0) is red. It is clear from a first look that this graph is the convolution of two distributions: the horizontal yellowish line and the diagonal yellowish line. The first one (horizontal yellowish line) represents the repeats that have the same length in the sequenced individual and in the reference genome. The other distribution (diagonal yellowish line) represents the repeats present in the reference genome but reported to have length 0 in the individual's genome. The first of the two distributions is the one we are interested in, while the latter one is just sequencing noise, due to the intrinsic difficulties in sequencing microsatellites.

We wanted to discriminate between these two distributions in order to study only the repeats that were not sequencing mistakes. For this reason for each repeat length in the human reference genome we plotted the distribution of the differences of the repeat length in the individual's genome relative to the length in the reference genome. The histograms for repeat length from 4 trinucleotide repeats to 11 trinucleotide repeats are shown in figure 2.4. The longer the repeat the better we can distinguish the real distribution, centered in 0, from the sequencing noise, centered on the left. At repeat length 4 it is not so clear the distinction between signal and noise, for this reason we preferred not to use repeats of this length.

We were able to retrieve 2695 of the 3087 repeats (87%) in the database, with a length of 4 or more aminoacids in the individual. To see if the length distribution of the repeats in the individual's genome was likely to be drawn out from the length distribution of the repeats in the reference genome, we performed a kolmogorov smirnov test . As we would have expected from figure (2.5) the test stated that no significative differences were present between the two distributions.

To study in more detail the variability of the repeats we computed for each repeat length in the reference genome the abundance of repeats that had a different length in the individual $NA12878$. It came out that only 2.6% of the repeats had a different length in the studied genome and in the reference genome. We performed the same

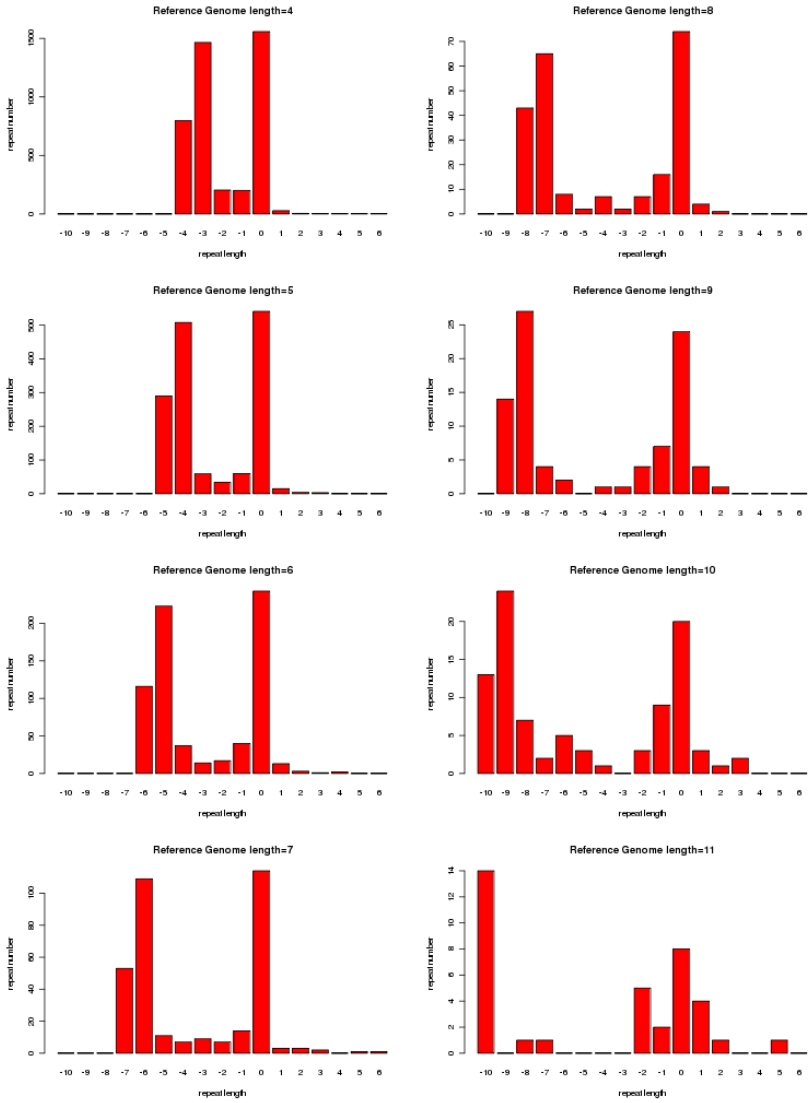



Figure 2.4: **Length distribution in the individual genome for each length in the reference genome.** The longer the repeat in the reference genome, the better we can discriminate between the signal and the noise (see text).

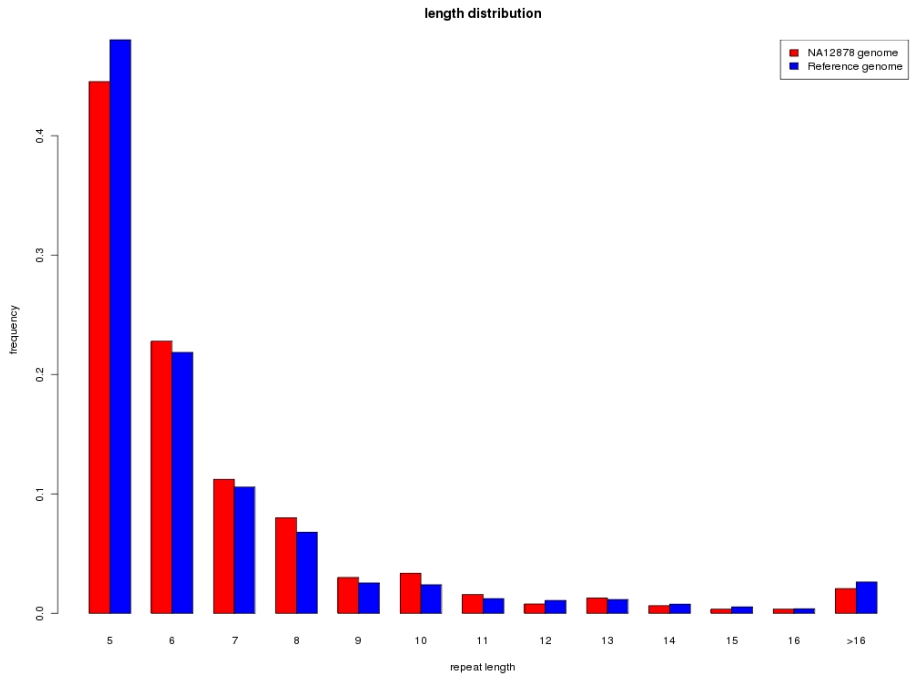



Figure 2.5: **Length distribution of the repeats in reference and in the individual's genome.** Length distribution of the repeats in reference genome in blue, and in the individual's genome in red: there are not statistically significant differences.

study in the individual $NA19240$ and we found that the percentage was 3.5%. These results are consistent with a recent study that compared repeat length in the reference genome with repeat length in Venter's genome [Payseur et al., 2010]. They found that 2.7% of repeats were different in the two genomes. Their result is not directly comparable with our results. They use a database with a minimung length of 3 repeats and they find that the repeats of length 3 are the must abundant but also the least variable. So we can use their result as a lower bound to our estimation.

We then studied in both individuals how many heterozygous sites we were able to find. For the majority of repeats no alleles of different sizes were observed. However for 3.7% of loci in $NA12878$ and for 2.4% in $NA19240$ we could confirm the presence of allelels of different length.

### 2.5.2 Analysis of variability in 22 individuals.

We then moved on to analyze polyQ repeat size in the 22 HapMap samples that had been sequenced using 454 technology. Our aim was to study the variability of human tandem repeats among the various individuals and with respect to the reference genome. For this reason we decided to focus on three main quantities: the repeat size differences, the sensitivity and the number of heterozygous repeats. The repeat size differences (called *Diff* in table 2.2) is the amount of repeats that have a different size in the studied genome and in the reference genome. The percentage is shown in parenthesis. The sensitivity (*Sens.* in table 2.2) is defined as the amount of repeats present in the reference genome that we were able to find also in the studied genome. Finally the number of heterozygous repeats (*Het.* column in table 2.2) is the amount of repeats that had two alleles of different size in the studied genome. The percentage is given in parenthesis. We computed those quantities for the repeats we found in each one of the individual sequenced with 454. Table 2.2 show

| Sample | | Unfiltered Dataset | | | | Filtered Dataset | | | |
|---|---|---|---|---|---|---|---|---|---|
| sample | population | n | sens. | het. | diff. | n | sens. | het. | diff. |
| NA07346 | CEPH-2 | 0 | - | - | - | 0 | - | - | - |
| NA07347 | CEPH-2 | 236 | 34% | 21 (9%) | 28 (11%) | 94 | 13,6% | 5 (5%) | 5 (5%) |
| NA11849 | CEPH-1 | 289 | 42% | 12 (4%) | 23 (7.6%) | 137 | 20% | 3 (2%) | 6 (4%) |
| NA11881 | CEPH-1 | 259 | 37% | 13 (5%) | 24 (9%) | 123 | 17% | 4 (3%) | 6 (5%) |
| NA11894 | CEPH-2 | 230 | 33 % | 10 (4%) | 18 (7.5%) | 97 | 14% | 4 (4%) | 4(4%) |
| NA11918 | CEPH-2 | 226 | 33% | 16(7%) | 27 (11%) | 98 | 14 % | 5 (5%) | 5(5%) |
| NA11931 | CEPH-2 | 224 | 32% | 11(5%) | 22(9%) | 88 | 12% | 2(2%) | 2(2%) |
| NA12043 | CEPH-1 | 230 | 33% | 16(7%) | 27(11%) | 101 | 14.6% | 1(1%) | 4(4%) |
| NA12045 | CEPH-2 | 206 | 30% | 7(3%) | 16(7%) | 91 | 13% | 1(1%) | 4(4%) |
| NA12234 | CEPH-1 | 209 | 30% | 9(4%) | 15(7%) | 101 | 14.6% | 5(5%) | 7(6%) |
| NA12249 | CEPH-1 | 309 | 45 % | 15 (5%) | 33(10%) | 147 | 21% | 4(3%) | 1 (1%) |
| NA12287 | CEPH-2 | 258 | 37 % | 16 (6%) | 31(11%) | 125 | 18% | 6(5%) | 7(5%) |
| NA12812 | CEPH-1 | 439 | 63 % | 31(7%) | 45 (10%) | 296 | 43% | 20(7%) | 23(7%) |
| NA12814 | CEPH-1 | 455 | 66% | 34(7%) | 53 (10%) | 271 | 39% | 14 (5%) | 15(5%) |
| NA12815 | CEPH-1 | 413 | 60% | 91(22%) | 60(12%) | 256 | 37% | 53(20%) | 23(7%) |
| NA12872 | CEPH-1 | 461 | 67 % | 30(6.5%) | 48(10%) | 297 | 43 % | 13 (4%) | 19 (6%) |
| NA12873 | CEPH-1 | 461 | 67% | 43 (9%) | 55(11%) | 305 | 44% | 19 (6%) | 16(5%) |
| NA12874 | CEPH-1 | 463 | 67% | 30(6%) | 44 (9%) | 305 | 44% | 13(4%) | 16 (5%) |
| NA18969 | Japanese1 | 0 | - % | - | - | 0 | -% | - | - |
| NA18970 | Japanese1 | 250 | 36% | 20(8%) | 28 (10%) | 115 | 16% | 9(8%) | 11(9%) |
| NA19141 | Yoruba | 233 | 33% | 17(7%) | 27(11%) | 98 | 14% | 6(6%) | 9(9%) |
| NA19143 | Yoruba | 287 | 41 % | 22(7%) | 37(12%) | 142 | 20% | 8(5%) | 9(6%) |

Table 2.2: **Repeat variability in the 22 individuals sequenced at low coverage.** For each individual it is shown the number of repeats found, the sensitivity, the number of heterozygous sites (both absolute numbers and percentage of the total amount of repeats found) and the repeat size differences with the reference genome (again both in absolute value and as a percentage of the repeats found). Those quantities are computed for both the total dataset (unfiltered) and in the case in which each repeat had to be supported by at least 2 reads (filtered dataset). See text for further details.

the results in two cases: the unfiltered database and the filtered one. The filtering consisted in getting rid of the repeats that were not supported by at least 2 sequencing reads, considering that the repeats supported by only one read could be sequencing mistakes.

The percentage of repeats detected in the filtered dataset ranges from 13% to 44% depending on the sample although two samples had no repeats at all, with an average in the 20 samples of 23.6%. The percentage of repeats displaying two alleles of different size ranges from 1% to 8% and there is an outlier at 20%. The percentage number of repeats that are of different size than in the human reference genome ranges from 1% to 7%. The above data indicates that we have, on average, reliable information on about one quarter of the polyQ sequences in the human genome in each sample. There are more or less 200 polyQ repeats (about 30% of the total) with data from 10 or more samples. About 10% of them are polymorphic, that is, in at least one individual we have a repeat size that is different from the one in the reference genome. This value is very similar to previous estimates based on human expressed sequence tag (EST) sequences [Mularoni et al., 2006a]. The sensitivity is greater for short repeats (size 4 to 6) but some long repeats (size 7 to 40) are still detected.

### 2.5.3 Analysis of the disease related repeats

PolyQ tracts involved in neurological disorders are typically highly polymorphic in human populations [Andrés et al., 2003, Butland et al., 2007]. We got a list of the repeats known to be involved in diseases resulting from aberrant elongation of poly -glutamine repeats from [Butland et al., 2007]. In table 2.3 we show the list of the diseases caused by abnormal elongation of $CAG$ repeats. We also report the gene that harbors the repeat, the normal and pathogenic length as reported in [Butland et al., 2007, Ranum and Cooper, 2006, Gatchel and Zoghbi, 2005] and the length we found in the human reference genome.

| Disease | Gene | Normal length | Pathogenic length | Ref. Gen. length |
|---|---|---|---|---|
| Spinocerebellar Ataxia type1 | ATXN1 | 6-39 | 40-82 | 12 |
| Spinocerebellar Ataxia type2 | ATXN2 | 15-24 | 32-200 | 23 |
| Spinocerebellar Ataxia type3 | ATXN3 | 13-36 | 61-84 | 14 |
| Spinocerebellar Ataxia type6 | CACNAIA | 4-20 | 20-29 | 13 |
| Spinocerebellar Ataxia type7 | ATXN7 | 4-35 | 37-306 | 10 |
| Spinocerebellar Ataxia type17 | TBP | 25-42 | 47-63 | 38 |
| Dentatorubral-pallidoluysian atrophy | ATN1 | 7-34 | 49-88 | 19 |
| Spinobulbar muscular atrophy Kennedy disease | AR | 9-36 | 38-62 | 23 |
| Huntington Disease | HD | 11-34 | 40-121 | 21 |

Table 2.3: **Diseases caused by abnormal eleongation of** $CAG$ **repeats.** The gene in which the repeat lies, the normal repeat length and the pathogenic one (all retrieved from [Butland et al., 2007] and [Usdin and Grabczyk, 2000]) and the length of the repeat in the reference genome.

We wanted to investigate the variability of these loci using data from 1000 Genomes Pilot Projects 1 and 2 and to compare it with the variability obtained in [Butland et al., 2007]. Unfortunately most of the repeats involved in disease, shown in table 2.3, were not present in the data we analyzed. We were able to find the requested repeats in more than 5 of the 24 genomes analyzed only for four of the nine repeats in table 2.3. The results are shown in table 2.4.

Most of the repeats were not present because the sequenced read either started in the middle of the repeat or ended before the repeat end, so we were not able to reconstruct the sequenced repeat length. Some of them have two lengths. As we used unfiltered data, it is unclear whether they correspond to two alleles of different size or to sequencing errors.

| disease | spinocerebellar Ataxia type1 | spinocerebellar Ataxia type3 | Dentatorubral palidoluysian atrophy | Spinobulbar muscular atrophy |
|---|---|---|---|---|
| gene | ATXN1 | ATXN3 | ATN1 | AR |
| Wild Type Repeat Size | 6-39 | 13-36 | 7-34 | 9-36 |
| Pathogenic Repeat Size | 40-82 | 61-84 | 49-88 | 38-62 |
| NCBI36 Repeat Size | 12 | 14 | 19 | 23 |
| NA12878 | 12 | - | 19 | - |
| NA19240 | 17 | 11 | 16-13 | - |
| NA07346 | - | - | - | - |
| NA07347 | - | - | - | - |
| NA11840 | - | - | - | 25 |
| NA11881 | 13 | - | - | - |
| NA11894 | - | - | - | - |
| NA11918 | - | - | - | 25-28 |
| NA11931 | - | 20 | - | - |
| NA12043 | 15-14 | 16-18 | - | 21 |
| NA12045 | - | - | - | - |
| NA12234 | - | 10 | - | - |
| NA12249 | - | - | - | - |
| NA12287 | 17-19 | - | 21 | 19 |
| NA12812 | 16 | - | - | - |
| NA12814 | - | 10 | - | 22 |
| NA12815 | - | 15-10 | 12 | 26 |
| NA12872 | - | - | 20 | - |
| NA12873 | - | 18 | - | 18 |
| NA12874 | - | - | - | - |
| NA18969 | - | - | - | - |
| NA18970 | - | - | - | - |
| NA19141 | - | - | - | - |
| NA19143 | 14-25 | 21 | - | 26-19 |

Table 2.4: **Variability of 4 repeats involved in neuromuscular disorders.** For each repeat it's given the wildtype and pathogenic size as reported in [Butland et al., 2007], the repeat size that we found in the reference genome and the repeat size we found in each studied individual, when we found any.

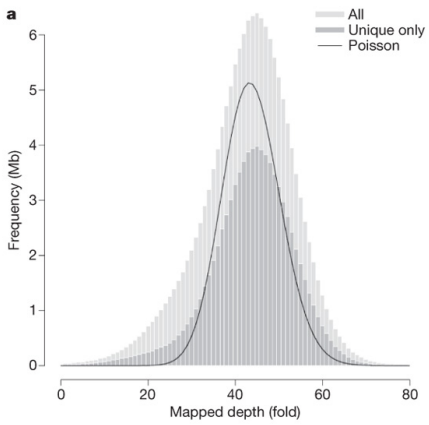



Figure 2.6: **Poisson distribution of the DNA fragments.** The distribution of the DNA fragments follows a Poisson distribution ( taken from [Bentley et al., 2008]).

## 2.6 Discussion

One main source of errors in next generation sequencing resides in the very first step of the sequencing algorithm: the random fragmentation of the DNA and the subsequent amplification of the libraries. If this step was really random as it is supposed to be, the number of reads overlapping on average each genomic position (coverage) obtained from sequencing should be poisson-distributed around the average coverage as shown in figure 2.6 which is taken from [Bentley et al., 2008]. The lower the coverage is, the more probable is that a genomic position does not appear in any read. Using a 454 platform and a coverage of 20x the probability of not finding a trinucleotide repeat longer that 4 nt, given a Poisson distribution is lower than $10^{-6}$. On the other hand with a coverage of 2x in the same platform we won't be able to find 16% of the repeats of length 4. An sketch of the problem is given in figure 2.7. Hence the 13% of repeats we are not able to find in each of the two individuals sequenced 20x cannot

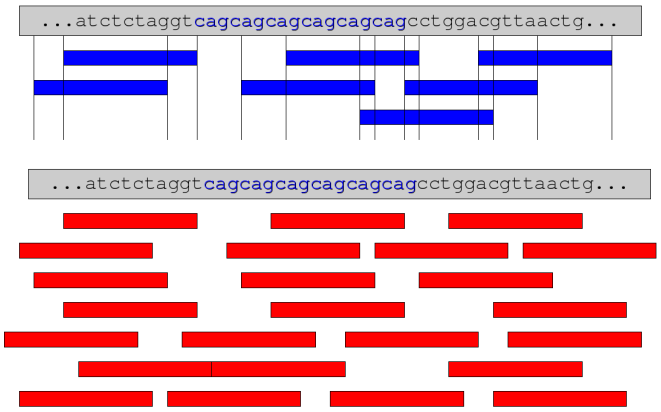



Figure 2.7: **A sketch of how coverage affects the assembly of a genome.** Blue reads show the case with low coverage, red reads show the case of high coverage. The higher the coverage, the least the probability that some part of the genome is not present in any of the sequencing reads.

be due to uneven coverage and is most probably due to mapping errors. As explained in the introduction, if a repeat can be mapped to more than one genomic region it is discarded or marked as low quality. This implies that, since most of the repeats lie in genomic simple regions that occur frequently in the human genome, lots of them are discarded or marked as low quality [Harismendy et al., 2009]. Moreover a repeat whose flanking regions map to the reference genome, but whose length is not the same as in the reference genome is often classified as a mistake (or low quality) [Harismendy et al., 2009]. Very few programs to correctly map repeats are available at the moment.

Similar conclusions were drawn in a very recent paper that performed a microsatellite variation study using data of susequent re-

leases of 1000 Genomes Pilot Projects [McIver et al., 2011].

Nonetheless we think that the study of repeats variability with data from NGS platforms is a very promising field, provided that sequencing is done at an appropriate coverage ($>$ 15x [Bentley et al., 2008, Harismendy et al., 2009]) repeat sequencing mistakes are minimized and that we solve the bioinformatics issues discussed in the chapter.

## 2.7 Conclusions

Next generation sequencing techniques give way to some totally new approaches to questions in molecular biology. We studied $CAG$ repeat variability in the human population using a public database, the 1000 Genomes Pilot Projects. The Pilot Projects were not aiming at studying repeats, and indeed we observed that next generation sequencing technologies have important intrinsic difficulties when dealing with microsatellites. For each one of the individuals we were able to find between 12% and 87% of the repeats we were looking for. In particular we could find almost all the disease related repeats in the high coverage sequenced individuals, but very few of them in the low coverage sequenced individuals. Due to lack of data the results were not fully conclusive. However this study has been useful to develop a bioinformatics pipeline for the study of repeat variability from sequence reads, which should become instrumental in the near future.

# Chapter 3

# NATURAL SELECTION DRIVES THE ACCUMULATION OF AMINO ACID TANDEM REPEATS IN HUMAN PROTEINS

Mularoni L, Ledda A, Toll-Riera M, Albà MM. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. Genome Res. 2010; 20(6): 745-54.

# Chapter 4

# PHYLOGENETIC CONSERVATION OF HUMAN MICROSATELLITES

Microsatellites are perfect repeats of short DNA sequence motifs, usually 1 to 4 bp in length, which are extremely abundant in animal genomes. They can rapidly expand or contract by replication slippage, potentially modifying gene function. Most microsatellites in coding regions are trinucleotide repeats, as changes in repeat length do not alter the open reading frame. They encode for homopolymeric amino acid runs, some of which, such as those composed of proline, glutamine or alanine, have been shown to play roles in modulating transcription regulation complexes. In accordance with a role of selection in preserving them, coding repeats are more evolutionary conserved than repeats located in other genomic regions. However, specific microsatellites located in non-coding gene regions have also been shown to modulate cellular functions, most notably acting as transcriptional or translational regulatory elements whose activity changes depending on microsatellite length variation. Here

we inquiry if this is reflected in a stronger than expected microsatellite conservation in some gene regions such as promoters or 5'UTRs, which would imply that functional roles are common. Using genomic alignments of 8 mammalian species, we have quantified the depth of conservation of microsatellites located in different genomic regions and having different repeating unit length and composition. We find that there is an excess of trinucleotide repeats not only in coding exons, but also in promoters and, in the case of repeats of GCC, in 5'UTRs. Whereas mononucleotide repeats and tetranucleotide repeats are poorly conserved in general, the level of conservation of dinucleotide repeats, and especially trinucleotide repeats, is much higher in specific gene regions, such as promoter, 5'UTR and 3'UTR, than in intergenic regions. The conservation is not random but, after considering microsatellite distribution in each gene region, and the variations in microsatellite stability depending on the nature of the microsatellite, we find that some microsatellite types, such as GC in promoter regions or AT in 3'UTRs, are more conserved than they should be if there was no selective pressure to maintain them. This extends previous findings on functional roles of non-coding microsatellites.

## 4.1 Introduction

Microsatellites are short sequence motif tandemly repeated DNA tracts [Ellegren, 2004] . The main mutational mechanism leading to changes in microsatellite length is replication slippage, a process by which misalignment of the repeat units during DNA replication leads to microsatellite expansion or contraction [Levinson and Gutman, 1987]. Changes in repeat length are estimated to occur very frequently, on the order of $10^{-4}$ to $10^{-3}$ per locus per generation [Weber and Wong, 1993], compared to typical estimates of $10^{-9}$ per nucleotide per generation for point mutations [Crow, 1993]. As a consequence of their high mutation rates, microsatellites are often polymorphic within populations [Wren et al., 2000], and show length variability in inter-species comparisons [Webster et al., 2002, Albà and Guigó, 2004].

Microsatellites account for about 3% of the human genome [Lander et al., 2001]. Mononucleotide repeats (monoNRs), followed by dinucleotide repeats (diNRs), are the most abundant microsatellites in human non-coding regions [Tóth et al., 2000, Subramanian et al., 2003b]. Trinucleotide repeats (triNRs), which do not disrupt the frame when they expand or contract, predominate in coding sequences. Microsatellite frequencies also depend on their composition. In vertebrates, the most frequent diNRs are $(CA)_n$ repeats, whereas $(CG)_n$ repeats are extremely rare [Tóth et al., 2000]. In mammalian coding sequences, the most abundant triNRs in frame are $(CAG)_n$ and $(GAG)_n$, encoding for poly-glutamine and poly-glutamic acid stretches [Albà and Guigó, 2004]. Coding sequences encoding homopolymeric amino acid repeats are much better conserved than similar sequences in non-coding regions, indicating that certain repeats in coding sequences have been maintained by selection [Mularoni et al., 2010]. This is likely to be related to functional roles of coding repeats, it has been observed that poly-glutamine, poly-proline or poly-alanine stretches may modulate gene transcriptional activity [Brown et al., 2005, Gerber et al., 1994, Lanz et al., 1995], and that histidine-rich

tracts are nuclear speckles targeting signal [Salichs et al., 2009].

Long considered junk sequences, evidence has been accumulating that microsatellites, even if located in non-coding genomic regions, may play functional roles [Li et al., 2004]. This is especially true for microsatellites located in gene expression regulatory sequences, where changes in microsatellite length can affect transcription factor binding or nucleosome positioning. For example, the promoter of heat shock protein gene hsp26 contains a CT repeat segment, bound by the GAGA factor, that is required for full heat shock inducibility [Leibovitch et al., 2002]. Another well-known example is the modulation of epidermal growth factor receptor gene transcription by a polymorphic $CA$ repeat tract in intron 1 [Gebhardt et al., 1999]. A recent large-scale study in S.cerevisiae addressed the role of microsatellites as possible modulators of gene transcription [Vinces et al., 2009]. It was found that as many as 25% of the yeast promoters contain microsatellites, and the genes containing the microsatellites showed greater expression divergence than the rest of genes. In addition, experiments performed in three different genes showed that changes in microsatellite length altered gene expression. Microsatellites in 5'UTRs also have the potential to regulate gene expression. For example deleting a $(CAG)_7$ in the 5'UTR of human calmodulin-1 gene decreases the expression of the gene by 45% [Toutenhoofd et al., 1998]. Experiments using $CTG$ repeated tracts fused to a reporter gene showed that the formation of more stable hairpins with longer $(CTG)_n$ progressively inhibited the scanning step of translation initiation [Raca et al., 2000].

The expansion of non-coding microsatellites can also result in disease. One of the first triplet expansion diseases to be discovered, myotonic dystrophy type 1 ($DM1$), is caused by a $CTG$ expansion in the 3' untranslated region of the dystrophica myotonin protein kinase gene [Brook et al., 1992]. Almost ten years later, expansion of the tretranucleotide repeat $(CCTG)_n$ in the first intron of zinc finger gene 9 was also found to cause myotonic dystrophy ($DM2$), termed DM2 [Liquori et al., 2001]. Other neuromuscular diseases

have been associated with the uncontrolled expansion of non-coding repeats, including fragile X syndrome ($FRAXA$), caused by expansions of a $CGG$ repeat in the 5'UTR of FMR-1 [Kenneson et al., 2001], spinocerebellar ataxia of type 8, due to expansions of $(CTG)_n$ in the 3'UTR of the SCA8 gene [Koob et al., 1999], spinocerebellar ataxia of type 10 associated with intronic $(ATTCT)_n$ expansions in SCA10 [Matsuura et al., 2000], and Friedereich's ataxia, caused by $GAA$ repeats in the intron of the FRDA-encoding gene [Ohshima et al., 1998]. The expanded alleles caused misregulation of gene expression and other gain-of-function pathogenic effects [Li et al., 2004, Mirkin, 2007, Ranum and Cooper, 2006].

Functional sequences are subject to evolutionary constraints and thus are expected to be more conserved across species than neutrally evolving sequences. This principle, called phylogenetic footprinting, has been the basis of several methods for the identification of functional regulatory sequences [Duret and Bucher, 1997, Prakash and Tompa, 2005]. A recent comparison of the level of conservation of sequences encoding human amino acid tandem repeats, with similar sequences located in non-coding genomic regions, has shown that the former are much more conserved, which is consistent with the stronger purifying selection acting on proteins and the functional roles of some amino acid tandem repeats [Mularoni et al., 2010]. In agreement, analysis of the conservation of human microsatellites in other vertebrate species has shown that microsatellites in coding exons are more conserved than microsatellites located in other genomic regions [Buschiazzo and Gemmell, 2010]. Intriguingly, this study has also reported that, among non-coding microsatellites, those located in untranslated regions (UTRs) are more conserved than those found in introns and intergenic regions. This suggests that microsatellites may frequently play functional roles in the 5'UTR.

Which is the nature of this putatively functional microsatellites? Are microsatellites located in other genomic regions, such as promoters and 3'UTRs, also more conserved than microsatellites located in intergenic regions? Do they have the same composition? To address

these questions we have performed a systematic study of the distribution and conservation of human mono-, di-, tri- and tetra-nucleotide repeats located in different genomic regions - intergenic, promoter, 5'UTR, coding exon, intron, 3'UTR and RNA gene. The results bring new light into the evolution and function of microsatellites in mammals.

## 4.2 Methods

**Identification and localization of microsatellites in the human genome** We downloaded human genome NCBI build 36 (March 2006), assembly hg18, from the University of California Santa Cruz (UCSC) database [Fujita et al., 2011]. Using a perl script, and pattern matching perl built-in functions, we identified all mono-, di-, tri- and tetranucleotide perfect repeats in the human genome (abbreviated monoRNs, diNRs, triNRs and tetraNRs). The minimum number of repeating units considered was 5 except for monoNRs, which had to be at least 10 repeating units in length (10 bp long).

In the case of di-, tri- and tetraNRs one motif was taken as representative of shifted motif repeats or equivalent repeats in the complementary strand (Table 4.1 ). For example $AC$ repeats (or $(AC)_n$) represented $AC$, $CA$, $GT$ and $TG$ repeats. For repeats found in different frames (for example $AC$ and $CA$) the longest one was kept.

Gene annotations for the hg18 version of the human genome were downloaded from Ensembl [Flicek et al., 2011]. We considered promoter (1 Kb upstream from the transcription start site), 5'UTR, coding exon, intron, 3'UTR, RNA gene and intergenic region. Using the genomic locations, we mapped all microsatellites to genomic regions, using the phyton based program Fjoin [Richardson, 2006]. In some cases a microsatellite was classified in several genomic regions (for example in an intron and in a coding exon if the region mapped to an alternative exon). More than 90% of the microsatellites had a unique classification.

**Microsatellite phylogenetic conservation** We obtained genomic alignments from Ensembl [Flicek et al., 2011] corresponding to syntenic regions from 9 eutherian species: *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Macaca mulatta* (macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog), *Bos Taurus* (cow) and *Equus caballus* (horse). We only considered those alignments that included information on at least one non-human primate species, one rodent species (mouse or rat) and an additional mammalian species (dog, cow or horse).

Human microsatellites were initially classified into four phylogenetic groups based on the existence of an overlapping repeat of size 4 or longer in a subset, or in the complete set, of aligned genomic sequences: human-specific (only found in human), primate-specific (found in humanand at least one other primate), *euarchontoglires*-specific (found in primates and rodents but not in more distant mammals) and, eutheria (found in at least one more distant mammal as well). As the number of repeats declined very rapidly with phylogenetic depth, making comparisons between groups difficulet, we decided to concentrate on just two groups: non-conserved and conserved human microsatellites. In the first group we clustered human-specific and primate-specific microsatellites. In the second group we clustered the euarchontoglires-specific and the eutheria classes. The first class consisted of 10,690 microsatellites (2,588 monoNRs, 7,462 diNRs, 377 triNRs, 263 tetraNRs) and the second class of 6,598 microsatellites (233 monoNRs, 5,594 diNRs, 735 triNRs, 36 tetraNRs).

**Statistical analyses** Statistical analyses was performed with the R statistical package [Team, 2008]. Comparisons between different distributions were performed with a chi-square test unless otherwise stated.

| | Repeat shown | Stands for |
|---|---|---|
| monoNR | A | A, T |
| | C | C, G |
| diNR | AC | AC, CA, GT, TG |
| | AG | AG, GA, CT, TC |
| | AT | AT, TA |
| | GC | CG, GC |
| triNR | AAC | AAC, ACA, CAA, GTT, TGT, TTG |
| | AAG | AAG, AGA, GAA, CTT, TCT, TTC |
| | AAT | AAT, ATA, TAA, ATT, ATA, TTA |
| | ACC | ACC, CAC, CCA, GGT, GTG, TGG |
| | ACG | ACG, CGA, GAC, CGT, GTC, TCG |
| | ACT | ACT, CTA, TAC, AGT, GTA, TAG |
| | AGC | AGC, GCA, CAG, GCT, CTG, TGC |
| | AGG | AGG, GGA, GAG, CCT, CTC, TCC |
| | ATG | ATG, TGA, GAT, CAT, ATC, TCA |
| | CCG | CCG, CGC, GCC, CGG, GCG, GGC |
| tetraNR | AAAC | AAAC, AACA, ACAA, CAAA, TTTG, TTGT, TGTT, GTTT |
| | AAAG | AAAG, AAGA, AGAA, GAAA, TTTC, TTCT, TCTT, CTTT |
| | AAAT | AAAT, AATA, ATAA, TAAA, TTTA, TTAT, TATT, ATTT |
| | AACC | AACC, ACCA, CCAA, CAAC, TTGG, TGGT, GGTT, GTTG |
| | AACG | AACG, ACGA, CGAA, GAAC, TTCG, TCGT, CGTT, GTTC |
| | AACT | AACT, ACTA, CTAA, TAAC, AGTT, GTTA, TTAG, TAGT |
| | AAGC | AAGC, AGCA, GCAA, CAAG, TTGC, TGCT, GCTT, CTTG |
| | AAGG | AAGG, AGGA, GGAA, GAAG, TTCC, TCCT, CCTT, CTTC |
| | AAGT | AAGT, AGTA, GTAA, TAAG, ACTT, CTTA, TTAC, TACT |
| | AATC | AATC, ATCA, TCAA, CAAT, ATTG, TTGA, TGAT, GATT |
| | AATG | AATG, ATGA, TGAA, GAAT, ATTC, TTCA, TCAT, CATT |
| | AATT | AATT, ATTA, TTAA, TAAT |
| | ACAC | ACAC, CACA, TGTG, GTGT |
| | ACCC | ACCC, CCCA, CCAC, CACC, TGGG, GGGT, GGTG, GTGG |
| | ACCG | ACCG, CCGA, CGAC, GACC, CGGT, GGTC, GTCG, TCGG |
| | ACCT | ACCT, CCTA, CTAC, TACC, AGGT, GGTA, GTAG, TAGG |
| | ACGC | ACGC, CGCA, GCAC, CACG, GCGT, CGTG, GTGC, TGCG |
| | ACGG | ACGG, CGGA, GGAC, GACG, CCGT, CGTC, GTCC, TCCG |
| | ACGT | ACGT, CGTA, GTAC, TACG |
| | ACTC | ACTC, CTCA, TCAC, CACT, GAGT, AGTG, GTGA, TGAG |
| | ACTG | ACTG, CTGA, TGAC, GACT, CAGT, AGTC, GTCA, TCAG |
| | AGAC | AGAC, GACA, ACAG, CAGA, GTCT, TCTG, CTGT, TGTC |
| | AGAG | AGAG, GAGA, CTCT, TCTC |
| | AGCC | AGCC, GCCA, CCAG, CAGC, GGCT, GCTG, CTGG, TGGC |
| | AGCG | GCGA, CGAG, GAGC, AGCG, CGCT, GCTC, CTCG, TCGC |
| | AGCT | AGCT, GCTA, CTAG, TAGC |
| | AGGC | AGGC, GGCA, GCAG, CAGG, GCCT, CCTG, CTGC, TGCC |
| | AGGG | AGGG, GGGA, GGAG, GAGG, CCCT, CCTC, CTCC, TCCC |
| | ATAC | ATAC, TACA, ACAT, CATA, GTAT, TATG, ATGT, TGTA |
| | ATAG | ATAG, TAGA, AGAT, GATA, CTAT, TATC, ATCT, TCTA |
| | ATAT | ATAT, TATA |
| | ATCC | ATCC, TCCA, CCAT, CATC, GGAT, GATG, ATGG, TGGA |
| | ATCG | ATCG, TCGA, CGAT, GATC |
| | ATGC | ATGC, TGCA, GCAT, CATG |
| | GCCC | GCCC, CCCG, CCGC, CGCC, GGGC, GGCG, GCGG, CGGG |
| | GCGC | GCGC, CGCG |
| | GGCC | GGCC, GCCG, CCGG, CGGC |

Table 4.1: **Index of microsatellite abbreviations**

| | Repeat | Average Length | Maximum Length | N of repeats |
|---|---|---|---|---|
| monoNR | A | 14.89687 | 90 | 1083872 |
| | C | 11.60620 | 81 | 10254 |
| | total | 14.86603 | 90 | 1094126 |
| diNR | AC | 9.560622 | 61 | 215268 |
| | AG | 7.365849 | 67 | 102198 |
| | AT | 7.886096 | 300 | 125299 |
| | GC | 6.239022 | 14 | 1799 |
| | total | 8.570678 | 300 | 444564 |
| triNR | AAC | 6.424260 | 18 | 18656 |
| | AAG | 8.197691 | 62 | 4851 |
| | AAT | 7.554911 | 20 | 20697 |
| | AGG | 6.473695 | 29 | 5094 |
| | ATC | 7.406951 | 123 | 3165 |
| | GCC | 6.791358 | 23 | 3240 |
| | others | 6.995060 | 210 | 6073 |
| | total | 7.072115 | 210 | 61776 |
| tetraNR | AAAC | 5.970444 | 13 | 12654 |
| | AAAG | 10.139881 | 83 | 11903 |
| | AAAT | 7.464180 | 17 | 22711 |
| | AAGG | 9.553437 | 39 | 7813 |
| | ATAG | 10.040596 | 17 | 6577 |
| | ATCC | 8.234513 | 133 | 3842 |
| | others | 6.862834 | 18 | 9062 |
| | total | 8.050616 | 133 | 74562 |

Table 4.2: **Average length, maximum length and total number of microsatellites repeats.** In the category others we grouped all the repeats that accounted for less than 5% of the total amount of repeats of each repeat length type. Mononucleotide repeats had a minimum trheshold of 10 nts, while all the others had a minimum threshold of 5 repeat units.

## 4.3 Results

**Microsatellite quantification in the human genome** We searched for all microsatellites of repeat unit size 1 to 4 in the human genome (NCBI build 36). The minimum microsatellite size we considered was 5 tandem repeat units, with the exception of mononucleotide repeats, which, because of their short size, were required to have at least 10 repeat units. Equivalent microsatellites in the complementary strand were clustered (for example $AC$ and $GT$ were both classified as $AC$), and overlapping microsatellites (for example $AC$ and $CA$) were counted only once ( Table 4.1). We obtained 1,675,028 microsatellites, of which about 65% were mononucleotide repeats (monoRNs), 27% dinucleotide repeats (diNRs), 4% trinucleotide repeats (triNRs) and 4% tetranucleotide repeats (tetraNRs) (Table 4.2).

On average, triNRs were shorter, in terms of number of repeat units, than diNRs and tetraNRs. We also observed that, in general, CG-rich repeats were less abundant, and tended to be shorter, than AT-rich repeats. This was exacerbated in the case of $(GC)_n$ repeats, which only constituted 0.4% of all diNRs. The relative frequencies of microsatellites of different composition were similar to those found in previous studies [Kelkar et al., 2008, Subirana and Messeguer, 2010, Tóth et al., 2000].

**Depletion of CG/GC and excess of TG/CA containing microsatellites** We noted that not only $(GC)_n$ (representing $(GC)_n$ and $(CG)_n$), but also $(ACG)_n$ was very rare (Table 4.4). In fact, $(AGC)_n$, with the same composition but lacking CG dinucleotides, was about 60 times more abundant than $(ACG)_n$. This is likely to be related to the known $CG$ restrictions in the human genome. In the context of the CpG dinucleotide, most cytosines are methylated and mutate to thymine ($CG \rightarrow TG$), and this is believed to cause a depletion of $CG$ in animal genomes [Bird, 1980].

To better understand the effect of CpG depletion, we quantified dinucleotide and trinucleotide abundance in the human genome (table

| | motif | type | N motif | frequency type | N type |
|---|---|---|---|---|---|
| DiNT | AC | AC, GT | 708847846 | 0.41 | 290443348 |
| | | CA, TG | | 0.59 | 418404498 |
| | AG | AG, CT | 745197717 | 0.54 | 403139895 |
| | | GA, TC | | 0.46 | 342057822 |
| | AT | AT, TA | 411480241 | | 411480241 |
| | GC | CG, GC | 151492657 | | 151492657 |
| TriNT | AAC | AAC, GTT | 309314309 | 0.27 | 83985862 |
| | | ACA, TGT | | 0.38 | 116195615 |
| | | CAA,TTG | | 0.35 | 109132832 |
| | AAG | AAG, CTT | 355972227 | 0.32 | 114941554 |
| | | AGA, TCT | | 0.36 | 127442590 |
| | | GAA, TTC | | 0.32 | 113588083 |
| | AAT | AAT, ATT | 381952735 | 0.38 | 143529981 |
| | | ATA, ATA | | 0.31 | 118686198 |
| | | TAA, TTA | | 0.31 | 119736556 |
| | ACC | ACC, GGT | 260054989 | 0.26 | 67055294 |
| | | CAC, GTG | | 0.33 | 86656569 |
| | | CCA, TGG | | 0.41 | 106343126 |
| | ACG | ACG, CGT | 81709067 | 0.18 | 14504770 |
| | | CGA, TCG | | 0.16 | 12743964 |
| | | GAC, GTC | | 0.67 | 54460333 |
| | ACT | ACT,AGT | 232304731 | 0.40 | 92686985 |
| | | CTA,TAG | | 0.32 | 74277650 |
| | | TAC GTA | | 0.28 | 65340096 |
| | AGC | AGC, GCT | 280441608 | 0.29 | 80594366 |
| | | GCA, TGC | | 0.30 | 83012357 |
| | | CAG, CTG | | 0.42 | 116834885 |
| | AGG | AGG, CCT | 288517621 | 0.35 | 102415086 |
| | | GGA, TTC | | 0.31 | 89017456 |
| | | GAG, CTC | | 0.34 | 97085079 |
| | ATG | ATG,CAT | 295555441 | 0.36 | 105779225 |
| | | TGA, TCA | | 0.38 | 112852619 |
| | | GAT, ATC | | 0.26 | 76923597 |
| | CCG | CCG, CGG | 98368593 | 0.16 | 15961290 |
| | | CGC, GCG | | 0.14 | 13760304 |
| | | GCC, GGC | | 0.70 | 68646999 |

Table 4.3: **Dinucleotide and trinucleotide abundance in the human genome.** for each motif the motif types belonging to it are give. The total amount of the motif occurrencies is given, along with the fraction and amount relative to each motif type.

4.3). $GC$ was 4.32 times more abundant than $CG$, denoting CpG depletion. Importantly the same effect could be observed at the level of trinucleotides: $CAG$ was about 4 times more abundant than $ACG$ or $CGA$, and $GCC$ about 4.65 times more abundant than $CCG$ or $CGC$.

We also found that $(ATC)_n$ was about 6.6 times more abundant than the alternative microsatellite sharing composition, $(ACT)_n$. The former includes $(TGA)_n$, which may potentially result from $CG$ mutations. In addition, $TG$ repeats (represented by $(CA)_n$ in table 4.4) were by far the most abundant diNRs. Therefore, one possibility was that previously reported increased $CG \rightarrow TG$ mutability was related to these observations. The dinucleotide and trinucleotide frequency data supported this hypothesis. There was an excess of $TG$ over $GT$ (1.44 times), as well as an excess of $TGA$ and $ATG$ over $GAT$ (1.46 and 1.37 times respectively).These differences become amplified when we consider repeating units (microsatellites).

**Excess of trinucleotide repeats in transcribed genomic regions** Using human genome annotations from Ensembl [Flicek et al., 2011] we classified the microsatellites according to genomic location: gene promoter (1 Kb upstream from the transcription start site), 5'UTR, coding exon, intron, 3'UTR, intergenic and RNA gene (figure 4.1). In introns and 3'UTR the distribution of microsatellites of different repeat unit size was very similar to the distribution in intergenic regions. Compared to these regions, promoters showed a slight increase in triNRs, and RNA genes in di- and triNRs. Not surprisingly, given the open reading frame restrictions, in coding exons triNRs were, by far, the most abundant among all classes. TriNRs were also the most abundant microsatellite types in 5'UTRs, strongly contrasting with the findings for 3'UTRs, in which triNRs were almost negligible.

For all microsatellite types analyzed (monoNRs, diNRs, triNRs and tetraNRs), the distribution in different genomic regions was non-random ($p < 10^{-5}$, Table 4.4). In intergenic and introns there were no

| | 3UTR | 5UTR | EXON | INTERGENIC | INTRON | NCgene | promoter |
|---|---|---|---|---|---|---|---|
| A | 12303 | 725 | 450 | 680819 | 326682 | 622 | 11495 |
| C | 275 | 68 | 21 | 5808 | 2941 | 26 | 470 |
| monoNR | | | | | | | |
| total | 12578 | 793 | 471 | 686627 | 329623 | 648 | 11965 |
| fraction | 0.0115 | 0.0008 | 0.0005 | 0.3161 | 0.0121 | 0.0006 | 0.6585 |
| AC | 2689 | 240 | 241 | 152035 | 57225 | 301 | 2537 |
| AG | 801 | 248 | 334 | 75418 | 24112 | 81 | 1204 |
| AT | 1309 | 24 | 19 | 93034 | 29844 | 91 | 978 |
| GC | 36 | 108 | 58 | 868 | 449 | 13 | 267 |
| diNR | | | | | | | |
| total | 4835 | 620 | 652 | 321355 | 111630 | 486 | 4986 |
| fraction | 0.0112 | 0.0014 | 0.0015 | 0.2511 | 0.0109 | 0.0011 | 0.7229 |
| AAC | 144 | 7 | 23 | 13580 | 4699 | 765 | 6 |
| AAG | 24 | 8 | 80 | 3619 | 1044 | 198 | 3 |
| AAT | 121 | 3 | 4 | 15282 | 5066 | 853 | 6 |
| ACC | 40 | 10 | 151 | 1768 | 746 | 148 | 2 |
| ACG | 1 | 1 | 10 | 15 | 12 | 3 | 4 |
| ACT | 0 | 0 | 1 | 335 | 139 | 21 | 1 |
| AGC | 72 | 79 | 697 | 1093 | 749 | 129 | 9 |
| AGG | 59 | 109 | 508 | 2928 | 1142 | 202 | 17 |
| ATC | 30 | 5 | 87 | 2154 | 861 | 130 | 4 |
| GCC | 58 | 672 | 448 | 914 | 448 | 334 | 66 |
| triNR | | | | | | | |
| total | 549 | 894 | 2009 | 41688 | 14906 | 118 | 1612 |
| fraction | 0.0261 | 0.0145 | 0.0325 | 0.2413 | 0.0089 | 0.0019 | 0.6748 |
| AAAC | 60 | 4 | 4 | 8502 | 3433 | 5 | 115 |
| AAAG | 27 | 9 | 8 | 9049 | 2264 | 6 | 126 |
| AAAT | 68 | 2 | 5 | 15318 | 6057 | 5 | 239 |
| AAGG | 11 | 0 | 3 | 6001 | 1455 | 2 | 66 |
| ATAG | 28 | 0 | 1 | 5061 | 1205 | 1 | 47 |
| ATCC | 8 | 0 | 1 | 2366 | 1213 | 3 | 43 |
| others | 51 | 31 | 4 | 5868 | 2529 | 8 | 130 |
| tetraNR | | | | | | | |
| total | 253 | 46 | 26 | 52165 | 18156 | 30 | 766 |
| fraction | 0.0107 | 0.0006 | 0.0004 | 0.2541 | 0.0035 | 0.0004 | 0.7302 |
| Genome (bp) | | | | | | | |
| total | 41,711,000 | 69,721,260 | 8,303,475 | 843,716,048 | 43,968,570 | 3,249,342 | 2,547,178,616 |
| fraction | 0.0117 | 0.0196 | 0.0023 | 0.2371 | 0.0124 | 0.0009 | 0.7159 |

Table 4.4: **Distribution of microsatellites in different genomic regions** The observed distribution was significantly different from the expected distribution when consiudering the genome fraction occupied by each region, for all microsatellite types (monoNRs, diNRs, triNRs,tetraNRs) using a $\chi^2$-test ($p < 10^{-5}$). In others we refer to the sum of microsatellites types found at a frequency less than 5%.

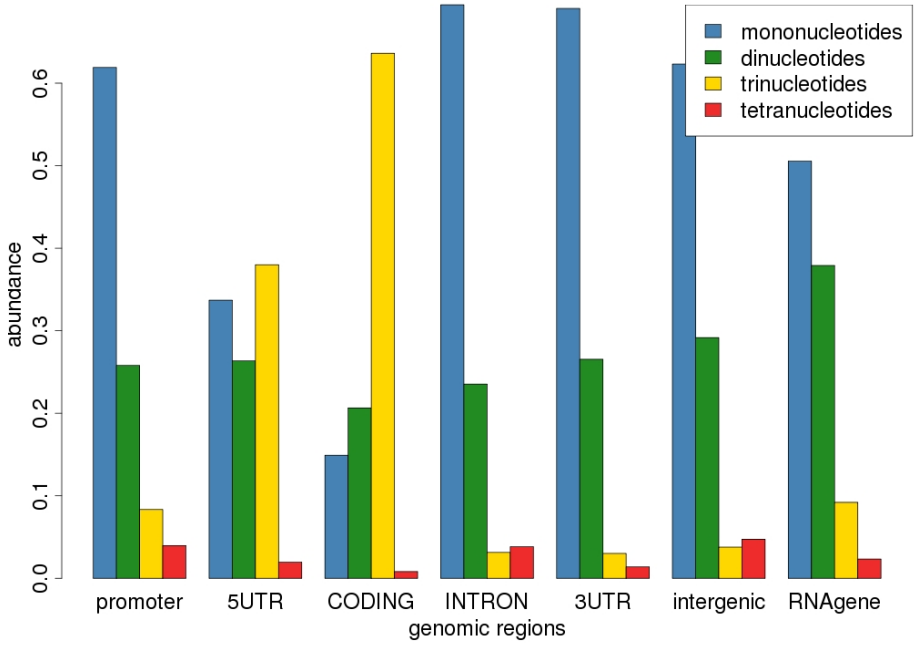



Figure 4.1: **Relative microsatellite abundance depending on genomic location** Relative abundance of human microsatellites of different repeat unit size and located in different genomic regions is shown.

large differences in the distribution of microsatellites of different repeating unit size with respect to the expectation, given the fraction of the genome these regions occupied and the general microsatellite abundance. Coding exons, as expected, showed a depletion of monoNRs, diNRs and tetraNRs, for the deleterious effects of frame-disrupting mutations, as discussed earlier. For reasons that remain to be explained, this depletion extended to 5'UTRs. TriNRs were over-represented in coding exons, and, to a lesser extent, promoters and RNA genes. In 5'UTRs there was not an excess of triNRs per se, but we observed a very strong bias towards $(GCC)_n$, which constituted 75% of all trinucleotide repeats (versus 22% in coding exons or 3% in introns). As we do not expect triNRs to be banned from regions that are evolving, essentially, in a neutral manner, such as intergenic regions, this excess in some gene regions can only be explained in terms of selection favoring their formation and/or maintenance. Selection favoring amino acid tandem repeats, originated by triNR expansion, in coding regions, has been recently reported [Mularoni et al., 2010]. The results presented here indicate that the effect of selection could extend to sequences located upstream from coding regions with the potential to regulate transcription and/or translation.

Not surprisingly, intergenic regions and introns contained higher frequencies of AT-rich microsatellites than other genomic regions with a higher GC content, such as promoters, 5'UTRs and coding exons (Table 4.4). For example, the most abundant triNRs in intergenic sequences are $(AAT)_n$ (37%) and $(AAC)_n$ (33%), whereas in coding exons they are $(AGC)_n$ (35%), $(AGG)_n$ (25%) and $(GCC)_n$ (22%). Promoters were enriched in different GC-rich microsatellites, such as $(GCC)_n$ (56%) but also $(GC)_n$ (5%, a relatively high fraction compared to 0.4% in the whole genome). As already mentioned, the majority of triNRs in 5'UTRs were $(GCC)_n$. In contrast, in 3'UTRs, as well as in introns, $(AAC)_n$ and $(AAT)_n$ predominated.

**Highly conserved microsatellites** The level of conservation of microsatellites can provide interesting clues about their functional-

| | | | intergenic | promoter | 5'UTR | coding | intron | 3'UTR | RNA gene | total |
|---|---|---|---|---|---|---|---|---|---|---|
| diNRs | ALL | N | 321,355 | 4,986 | 620 | 652 | 111,630 | 4,835 | 486 | 444,564 |
| | CONS | N | 8,552 | 230 | 63 | 81 | 3,615 | 483 | 32 | 13,056 |
| | | % | 2.66% | 4.61% | 10.16% | 12.42% | 3.24% | 9.99% | 6.58% | 2.94% |
| triNRs | ALL | N | 41,688 | 1,612 | 894 | 2,009 | 14,906 | 549 | 118 | 61,776 |
| | CONS | N | 355 | 102 | 75 | 309 | 234 | 27 | 10 | 1,112 |
| | | % | 0.85% | 6.33% | 8.39% | 15.38% | 1.57% | 4.92% | 8.47% | 1.8% |

Table 4.5: **Conservation of di- and triNRs in different genomic regions.** ALL: complete dataset; CONS: subset of microsatellites conserved in at least one non-primate mammal.

ity in different genomic regions. We quantified how many of the human microsatellites were conserved in other mammals, using genomic multiple alignments from Ensembl Compara containing data from 9 mammalian species. We selected alignments that contained syntenic sequences from at least one other primate species (chimpanzee and/or macaque), one rodent species (mouse and/or rat) and one additional mammalian species (out of dog, cow and horse). We recovered alignments for nearly 40% of the microsatellites in the initial dataset (646,783 out of 1,675,028).

Using the alignments, we determined the level of conservation of microsatellites formed by different repeat units in different genomic regions. A human microsatellite was defined as conserved if we could identify an equivalent microsatellite in at least one non-human primate and one rodent species. In 38% of the cases microsatellite conservation extended to more distant mammalian species, but we did not consider these cases separately in order to achieve sufficient statistical power in our comparisons.

MonoNRs and tetraNRs were, in general, more poorly conserved than diNRs and triNRs ($0.14 - 0.27\%$ versus $1.8 - 2.94\%$ conserved). We compared diNR and triNR conservation for different genomic regions (Table 4.5 ). Intergenic regions showed the lowest level of diNR and triNR conservation (2.66% and 0.85%, respectively). Relative to intergenic conservation, conservation of triNRs was higher than con-

servation of diNRs in all gene regions. The highest triNR conservation was observed in coding exons, with 15.4% conserved microsatellites, and 5'UTRs, 8.4% conserved microsatellites.

**Microsatellite stability** Some microsatellites may be more stable than others due to their different composition. This can be best measured in intergenic regions, where microsatellites are expected to evolve free from selection. We defined microsatellite stability as the ratio between conserved microsatellite and total microsatellite in intergenic regions (Table 4.6). Indeed, we observed that repeat stability strongly depended on the nature of the repeat. For example, among diNRs, $(AC)_n$ was 6 times better conserved than $(AT)_n$. Among triNRs, the most stable, although very rare in genomes, was $(ACG)_n$, followed by $(GCC)_n$, $(AGG)_n$ and $(AGC)_n$.

**Selective biases in microsatellite conservation** We used the repeat stability values, calculated as explained above using intergenic regions, to calculate the number of conserved repeats, of each type, that we expected in each gene region, given its specific microsatellite distribution. For example the stability of $(AC)_n$ is 0.043, which means that, using our collection of genomic alignments, 4.3% of $(AC)_n$ microsatellites in human are conserved, above the length threshold, in several mammals (at leats one primate and one non-primate mammal). In promoters we have identified $2,537$ $(AC)_n$ microsatellites, so, just by mutational processes alone, we will expect to find 109 conserved at the mammalian level (2,537 x 0.043). As we observe 151, there is a 1.38 increase in the conservation of $(AC)_n$ in promoters with respect to intergenic regions. Summing for all diNRs we expect 144 conserved microsatellites. As we observe 230, there is 1.6 times more conservation than expected.

Using this approach, different gene regions continued to show an increased number of conserved repeats over the neutral expectation (Table 4.7). For example, the number of conserved triNRs was about 4.3 times larger than expected. More moderate but still significant

| | intergenic | all | conserved | stability | rel. stability |
|---|---|---|---|---|---|
| monoNRs | A | 680,819 | 1,718 | 0.00252 | 1.007 |
| | C | 5,808 | 2 | 0.00034 | 0.137 |
| | total | 686,627 | 1,720 | 0.00250 | |
| diNRs | AC | 152,035 | 6,533 | 0.04297 | 1.615 |
| | AG | 75,418 | 1,326 | 0.01758 | 0.661 |
| | AT | 93,034 | 669 | 0.00719 | 0.270 |
| | GC | 868 | 24 | 0.02765 | 1.039 |
| | total | 321,355 | 8,552 | 0.02661 | |
| triNRs | AAC | 13,580 | 37 | 0.00272 | 0.319 |
| | AAG | 3,619 | 32 | 0.00884 | 1.038 |
| | AAT | 15,282 | 38 | 0.00248 | 0.292 |
| | ACC | 1,768 | 10 | 0.00565 | 0.664 |
| | ACG | 15 | 6 | 0.40000 | 46.972 |
| | ACT | 335 | 2 | 0.00597 | 0.701 |
| | AGC | 1,093 | 24 | 0.02195 | 2.578 |
| | AGG | 2,928 | 133 | 0.04542 | 5.334 |
| | ATC | 2,154 | 20 | 0.00928 | 1.090 |
| | GCC | 914 | 53 | 0.05798 | 6.809 |
| | total | 41,688 | 355 | 0.00851 | |
| tetraNRs | AAAC | 8,502 | 25 | 0.00294 | 0.649 |
| | AAAG | 9,049 | 7 | 0.00077 | 0.171 |
| | AAAT | 15,318 | 38 | 0.00248 | 0.548 |
| | AAGG | 6,001 | 18 | 0.00299 | 0.663 |
| | ATAG | 5,061 | 64 | 0.01264 | 2.795 |
| | ATCC | 2,366 | 36 | 0.01521 | 3.363 |
| | others | 5,868 | 24 | 0.00408 | 0.904 |
| | total | 52,165 | 236 | 0.00452 | |

Table 4.6: **Stability values of diffrent intergenic microsatellites.** Stability: fraction of conserved versus all microsatellites. Rel. stability: relative stability; stability of the microsatellite subtype divided by total stability. others: sum of microsatellite types found at a frequency of less than 5%.

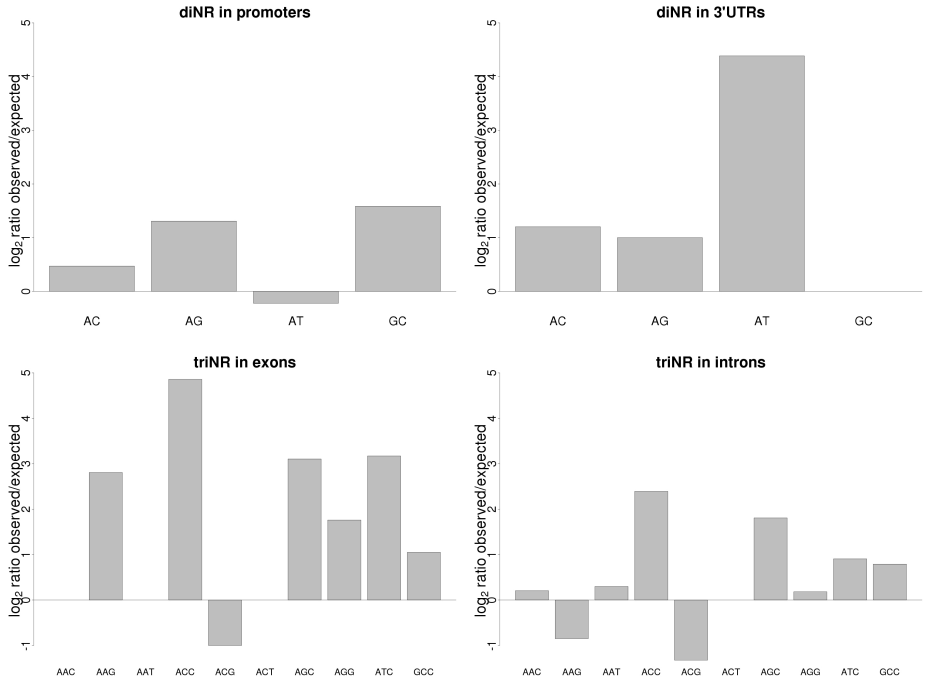



Figure 4.2: **Conserved microsatellites in different genomic regions.** The $log_2$ ratio of the observed versus expected number of microsatellites conserved in non-primate mammals is shown. The expected number is obtained from observations in intergenic regions and assumes no natural selection. In the four cases shown the distribution of observed microsatellites of different composition was significantly different from the expected distribution ($\chi^2$test, $p < 10^{-3}$).

| | promoter | | 5UTR | | coding | | intron | | 3'UTR | | RNAgene | | intergenic | | total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | obs | exp | obs | exp | obs | exp | obs | exp | obs | exp | obs | exp | obs | exp | obs | exp |
| A | 46 | 29 | 21 | 2 | 5 | 1 | 958 | 824 | 66 | 31 | 1718 | 2 | 4 | 1718 | 2818 | 2607 |
| C | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 2 | 3 | 3 |
| total | 47 | 29 | 21 | 2 | 5 | 1 | 95 | 825 | 66 | 31 | 1720 | 2 | 4 | 1720 | 2821 | 2610 |
| fract | 1.62 | | 10.5 | | 5 | | 1.16 | | 2.12 | | 860 | | 0.002 | | 1.08 | |
| AC | 151 | 109 | 26 | 10 | 25 | 10 | 2811 | 2459 | 267 | 116 | 6533 | 13 | 19 | 6533 | 9832 | 9250 |
| AG | 52 | 21 | 26 | 4 | 54 | 6 | 521 | 424 | 28 | 14 | 1326 | 1 | 13 | 1326 | 2020 | 1797 |
| AT | 6 | 7 | 1 | 0 | 0 | 0 | 272 | 215 | 188 | 9 | 669 | 1 | 0 | 669 | 1136 | 901 |
| GC | 21 | 7 | 10 | 3 | 2 | 2 | 11 | 12 | 0 | 1 | 24 | 0 | 0 | 24 | 68 | 50 |
| total | 230 | 144 | 63 | 17 | 81 | 18 | 3615 | 3110 | 483 | 140 | 8552 | 15 | 32 | 8552 | 13056 | 11998 |
| fract | 1.6 | - | 3.70 | - | 4.5 | 0 | 1.16 | - | 3.45 | - | 570.13 | - | 0.004 | - | 1.09 | 0 |
| AAC | 2 | 0 | 0 | 0 | 0 | 0 | 6 | 8 | 0 | 0 | 23 | 1 | 0 | 23 | 31 | 0 |
| AAG | 0 | 0 | 1 | 0 | 5 | 1 | 4 | 9 | 0 | 0 | 32 | 2 | 0 | 32 | 42 | 1 |
| AAT | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 9 | 6 | 0 | 27 | 2 | 0 | 27 | 44 | 0 |
| ACC | 0 | 0 | 1 | 0 | 14 | 1 | 11 | 3 | 2 | 0 | 7 | 1 | 0 | 7 | 35 | 0 |
| ACG | 3 | 2 | 0 | 0 | 1 | 4 | 2 | 5 | 0 | 0 | 6 | 1 | 4 | 6 | 16 | 2 |
| ACT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AGC | 2 | 0 | 2 | 2 | 112 | 15 | 54 | 16 | 2 | 2 | 24 | 3 | 0 | 24 | 196 | 2 |
| AGG | 9 | 0 | 5 | 1 | 31 | 3 | 24 | 6 | 3 | 0 | 15 | 1 | 0 | 15 | 87 | 2 |
| ATC | 0 | 0 | 0 | 0 | 6 | 1 | 6 | 5 | 0 | 0 | 13 | 1 | 0 | 13 | 25 | 0 |
| GCC | 25 | 2 | 31 | 22 | 38 | 15 | 28 | 15 | 3 | 2 | 30 | 11 | 6 | 30 | 161 | 21 |
| total | 41 | 4 | 40 | 25 | 207 | 40 | 146 | 76 | 16 | 4 | 177 | 23 | 10 | 177 | 637 | 28 |
| fract | 10.25 | 0 | 1.6 | 0 | 5.175 | 0 | 1.92 | 0 | 4 | 0 | 7.69 | 0 | 0.06 | 0 | 22.75 | 0 |
| AAAC | 2 | 0 | 0 | 0 | 0 | 0 | 9 | 10 | 0 | 0 | 25 | 0 | 0 | 25 | 36 | 36 |
| AAAG | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 7 | 0 | 0 | 7 | 9 | 9 |
| AAAT | 1 | 1 | 0 | 0 | 0 | 0 | 13 | 15 | 2 | 0 | 38 | 0 | 0 | 38 | 54 | 54 |
| AACC | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| AAGG | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 18 | 0 | 0 | 18 | 22 | 23 |
| AATC | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 3 | 2 |
| AATG | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 10 | 2 | 0 | 19 | 0 | 0 | 19 | 32 | 29 |
| AGGG | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 2 | 0 | 0 | 2 | 4 | 3 |
| ATAC | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 2 | 1 |
| ATAG | 0 | 1 | 0 | 0 | 0 | 0 | 13 | 15 | 0 | 0 | 64 | 0 | 0 | 64 | 77 | 80 |
| ATCC | 0 | 1 | 0 | 0 | 0 | 0 | 20 | 18 | 1 | 0 | 36 | 0 | 0 | 36 | 57 | 55 |
| GCCC | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| AATT | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| total | 3 | 3 | 0 | 0 | 0 | 0 | 79 | 76 | 5 | 0 | 212 | 0 | 0 | 212 | 299 | 293 |
| fract | 1 | 0 | 0 | 0 | 0 | 0 | 1.039 | 0 | - | 0 | - | 0 | 0 | 0 | 1.02 | 0 |

Table 4.7: **Observed and expected conserved microsatellites.** For each genomic region the amount of observed and expected number of conserved microsatellite is given. To see how we inferred the expected number of conserved microsatellite see text. For tetranucleotide repeats only the non zero repeats are shown.

increases in triNR conservation were detected in other gene regions, such as promoters (1.73 times) and 5'UTRs (1.63 times). Further, we could identify several cases in which the relative distribution of different repeat types was significantly different from the expected ($p < 10^{-3}$, Figure 2). For example, in coding exons $(ACC)_n$, $(ATC)_n$ and $(AGC)_n$ were more conserved than average. The most surprising result, however, was the about 20 fold increased in conserved $(AT)_n$ in 3'UTRs.

## 4.4 Discussion

Genomic studies have shown that some microsatellite types are much more abundant than others, and that this strongly depends on the species or taxonomic group under study [Katti et al., 2001, Subirana and Messeguer, 2010, Tóth et al., 2000]. The predominance of $(CA)_n$ over other dinucleotide repeats is widespread in vertebrates and arthropods, although in plants and fungi $(AT)_n$ is more common. $(CG)_n$ is very rare in all animal genomes, plants and fungi. This agrees with the known CpG depletion in genomes, due to cytosine methylation and subsequent deamination of 5-methylcytosine leading to TpG/CpA dinucleotides. An excess of TpG dinucleotides has actually been correlated with CpG depletion [Simmen, 2008]. We found that these biases were amplified in microsatellites: whereas CG dinucleotides in the human genomes were approximately about 9 times less frequent than expected if all dinucleotides had equal frequencies, $(CG)_n$ were about 62 times less frequent than expected. We found that stability of $(CG)_n$ was average, so this strong is likely to result from deleterious effects of $(CG)_n$. CpG dinucleotides are rare everywhere in the genome except in CpG islands, regions near the transcription start site characterized by high CpG frequency. CpG islands are associated with housekeeping gene transcription [Farré et al., 2007, Yamashita et al., 2005], and, contrary to other genomic locations, the CpGs are in normal conditions unmethylated. The ob-

served abundance of $(CG)_n$ in promoters was in accordance with the frequent presence of CpG islands in these regions. We noted that $(CG)_n$ located in promoters were exceptionally well-conserved, perhaps due to decreased mutability of unmethylated CpG and/or to selective pressure to maintain them.

The distribution of triNRs was also very uneven. 5'UTRs contained many more triNRs than 3'UTRs, suggesting that some triNRs may play regulatory roles in the context of 5'UTRs. The majority of these triNRs were $(GCC)_n$. These repeats have been linked to neurodegenerative diseases: the uncontrolled expansion of $(GCC)_n$ in the 5'UTRs of the FMR-1 and FMR-2 genes is associated with different forms of mental retardation and ataxia [Cummings and Zoghbi, 2000, Gatchel and Zoghbi, 2005, Kenneson et al., 2001]. When $(GCC)_n$ expands over a certain limit, transcriptional silencing and loss of the protein products results. In the case of FMR-2, it has been reported that $(GCC)_n$ expansion leads to hypermethylation of a CpG island, which would be the cause of transcriptional silencing [Knight et al., 1993]. In coding sequences, the most overrepresented triNRs were $(AGC)_n$, $(AGG)_n$ and $(GCC)_n$. These triNRs correspond to some of the most common amino acid homopolymeric repeats in the human genome. For example, $(AGC)_n$ includes AGC, CAG, GCT and CTG repeats, encoding poly-serine, poly-glutamine, poly-alanine and poly-leucine, which are among the most frequent ones [Albà and Guigó, 2004]. Interestingly, the triNRs overrepresented in coding exons are those that can more easily form hairpins [Kozlowski et al., 2010]. Induction of RNA structures by the presence of triNRs in transcripts could potentially play roles in translational regulation.

The evolutionary stability of microsatellites formed by different nucleotides was assessed using conservation data from intergenic regions. The repeat motif composition of microsatellites has been shown to affect microsatellite mutability [Ellegren, 2004, Kelkar et al., 2008], so this was important to discern between the influence of mutation and selection in the conservation of microsatellites in different

gene regions. We found that the nature of the microsatellite had a strong influence on its conservation. First, monoNRs and tetraNRs repeats were less conserved than diNRs and triNRs. For mononucleotide repeats this is likely to be related to the longer size threshold used for microsatellite detection, which was at least 10 repeating units instead of 5 repeating units. It is however unclear why, in intergenic regions, tetraNRs were about 6 times less conserved than diNRs and triNRs. The percent conservation at the mammalian level of human diNR types ranged from 0.7%, in the case of $(AT)_n$ to 4.3%, in the case of $(AC)_n$. Although we may have missed some repeats due to incomplete sequencing of the syntenic regions included in the alignments, this should affect all diNRs equally, so mutational biases are likely to be responsible. Among triNRs, $(ACG)_n$, $(GCC)_n$, and $(AGG)_n$ were $> 5$ times more conserved than average, and $(AGC)_n$, was about 2.5 times more conserved than average. These triNRs, except $(ACG)_n$, which is very rare, happen to be the most common in coding exons, 5'UTRs and promoters (4.4). As triNRs in coding regions are subject to negative selection [Mularoni et al., 2010], one possible explanation is the presence of yet unannotated genes in intergenic regions. If this were the case, our estimates on the increased conservation of triNRs in coding exons with respect to intergenic regions would be conservative.

Sequence conservation has been widely used to identify putatively functional sequences in non-coding genomic regions, typically involved in gene expression regulation [Duret and Bucher, 1997, Prakash and Tompa, 2005, Xie et al., 2005]. For example, using alignments of mammalian syntenic sequences, 174 candidate motifs in mammalian promoters, and 106 motifs in 3'UTRs were discovered [Xie et al., 2005]. Most motifs in promoters corresponded to known transcription factor binding sites, whereas some motifs in 3'UTRs were putative miRNAs, and, in general, proposed to play roles in post-transcriptional regulation. Here we used sequence conservation measurements to investigate if microsatellites located in particular gene regions, such as promoters and UTRs, showed abnormally high con-

servation levels, indicative of selection acting to preserve their function.

As microsatellites, due to their repetitive nature, are difficult to align with precision, we used conservation of a minimum sequence length instead of conservation of the sequence per se, similarly to previous studies on the conservation of coding sequence repeats [Mularoni et al., 2007, Mularoni et al., 2010]. Microsatellites in coding exons (mostly triNRs) showed the highest conservation, as expected given recent reports on a role of selection in preserving amino acid tandem repeats [Haerty and Golding, 2010b, Mularoni et al., 2010, Haerty and Golding, 2010b]. The best conserved, although not the most common, was $(ACC)_n$. This microsatellite can encode for histidine repeats (frame CAC), which have been recently discovered to be important for targeting proteins to nuclear speckles [Salichs et al., 2009]. Therefore it is not surprising that, due to selective constraints to maintain its function, it is especially well preserved. it is highly conserved. The second region with the highest conservation of microsatellites was 5'UTR, with a strong over-representation of $(GCC)_n$ and, to a lesser degree, $(AG)_n$ and $(GC_n$. The presence of RNA secondary structures in the 5'UTRs requires additional energy during RNA scanning for translation [Jackson and Linsley, 2010]. Thus, one can expect that changes in the length of $(GCC)_n$ or of other micorsatellites able to to form hairpins, will have a direct effect on the efficiency of translation. Microsatellites in promoters were also remarkably well conserved in comparison to the results for intergenic regions. Here GC-rich microsatellites also predominate, probably in many cases as part of CpG islands.

The most unexpected result was the strong conservation of $(AT)_n$ in 3'UTRs. Thre are numerous known examples of control of translation initiation by 3'UTR-protein interactions, especially in developmental genes, and 3'UTRs are also the target of miRNA-mediated translational repression [Jackson and Linsley, 2010]. It has been reported that miRNA have a preference for AT-rich 3'UTRs [Robins and Press, 2005]. AT-rich elements in the 3'UTR od insulin-like

growth factor binding protein 3 ($IGFBP-3$) gene have been shown to mediate translation repression of the protein[Subramaniam et al., 2010]. A repeat unit length polymorphism in the $(AT)_n$ microsatellite located at the 3'UTR of cytotoxic T-lymphocite antigen 4 ($CTLA4$) has been associated with different susceptibility to rheumatoid arthritis [Rodriguez et al., 2002]. Although the putative functions of $(AT)_n$ in 3'UTR remain to be determined, it is tempting to speculate that they may act or influence the control of translational regulation in many genes.

## 4.5 Conclusions

We have performed the first study yo date aimed at comparing the level of conservation of microsatellites of different composition and located in different genomic regions. By using intergenic regions as a control, we have been able to distinguish between differential conservation due to mutational bias, and to selection. We find increased microsatellite conservation in differnt regions that play roles in gene expression control, such as promoters and UTRs, which provides support for a role of microsatellites in modulating cellular functions.

## 4.6 Aknowledgements

The data for table 4.3 was provided by Nicolás Bellora, Ph.D.

We received financial support from Ministerio de Ciencia e Innovación (BIO2009-08160), Regione Autonoma della Sardegna (A.L.) and Fundació Institució Catalana de recerca i Estudis Avancats (M.M.A)

# Part IV

# Discussion

# Chapter 5

# DISCUSSION

Microsatellites are repetitive genomic regions ubiquitous in eukaryotic genomes. Since their discovery, in the late '80s, microsatellites have been widely used in biology. In the rest of this chapter we will give a brief summary of the topics discussed in this thesis and place them in the broader perspective of the study of microsatellites.

## 5.1 Thesis overview

This thesis analyses some aspects of microsatellites variability and conservation.

The first part of this thesis gives an introduction to microsatellites, to their evolutionary mechanism, to their function and role inside and outside proteins, to the main bioinformatics method we used to infer their functionality and to their role in some of the diseases they are involved in.

In the second part the objectives of the thesis are stated and in part three the research carried out during this thesis is described. One of the studies has already been published as a research article, while the other two are still in preparation. Another study we took part in, which was published as a research article, is reported in the appendix.This article was in the thesis of Eulalia Salichs (UPF, 2009).

Each chapter of this thesis includes a discussion of the corresponding results. This section does not attempt to go over those points again but to provide some general remarks on our work in the general context of studies on microsatellites.

In chapter 2 we focused our attention on polyglutamine polymorphism in the human genome. Polymorphic polyglutamine repeats are known to be involved in neurodegenerative diseases [Butland et al., 2007]. We wanted to develop a pipeline to study the polymorphism of human polyglutamine repeats, and in particular of disease-related polyglutamine repeats, with data from next generation sequencing technologies [Metzker, 2009]. For this reason we used data from two of the three pilots of the 1000 Genomes Project [Durbin et al., 2010].

The objectives of this work were accomplished, within the limitations of the data available in the Pilots.

In chapter 3 we studied the role of selection in the evolution of amino acid tandem repeats. To do this we compared the conservation in 12 vertebrate genomes of DNA sequences encoding aminoacid tandem repeats with sequences with the same trinucleotide repeat composition located in syntenic non-coding regions. Sequences located in non-coding regions showed much less conservation than sequences located in coding regions. This result pointed to a strong role of negative selection in shaping the evolution of amino acid tandem repeats.

The work presented in chapter 4 was a natural follow up of the work presented in chapter 3. In this work we analyzed in more detail the role of evolution in shaping the properties of microsatellites in different genomic regions such as promoters, 5' and 3' UTRs, introns, noncoding genes and intergenic regions. For this reason we studied the abundance and length distribution of microsatellites in each region. Moreover we studied the differences in the conservation pattern of the microsatellites belonging to the different genomic regions. we observed that in some regions, such as promoters and 5'UTRs, microsatellites were relatively well conserved in general. we also identified some microsatellite types that showed an unusually high conservation and which are strongly candidate to play functional roles.

In summary, the work presented in this thesis represents an effort to apply the most recent technological innovations and the most up to date questions to a subject, like microsatellites, that is considered old-fashioned by a certain part of the scientific community.

## 5.2 Microsatellites variability

Many studies have been undertaken on microsatellite variability in the human population [Andrés et al., 2003, Becker et al., 2007, But-

land et al., 2007, Lavoie et al., 2003]. The most used approach has been genotyping a small amount of microsatellites in a representative group of individuals (usually few hundreds of individuals). This approach is costly and time spending, so the analysis has focused on the study of specific problems. A study of this kind was for example the one done by Butland et al. [Butland et al., 2007] to investigate the variability of CAG encoded polyglutamine repeats to identify possible candidate disease genes [Butland et al., 2007]. Microsatellite loci were studied in different populations to obtain accurate allele frequency data and other parameters of forensic interest [Pérez-Lezaun et al., 2000, Becker et al., 2007]. Another study used data available from sequence databases to infer variability of human repeats [Mularoni et al., 2006b]. Recontruction of the entire picture of microsatellite variability has only become possible with the advent of NGS technologies.

.

Recently the microsatellites in all the genomic regions from the human reference genome and from Craig Venter's genome were used to reconstruct a profile of human microsatellite variability [Payseur et al., 2010]. In this panorama our approach to study microsatellites was a natural follow up of the previously used methods. Next Generation Sequencing is nowadays the most popular way to sequence genomes. There are several reasons for this popularity, among them that is a cheap and quick technique (a more general review of NGS sequencing technology is given in chapter 2.2). The amount of data, especially resequencing data, produced by NGS platforms is so enormous that they had to invent new ways to store and access data [Li et al., 2009] because using text files would have been impossible. We think that in the near future most of the bioinformatics groups will have to deal with this type of data. We dealt with it at the dawn of the NGS era, with the first data release of one of the first projects [Durbin et al., 2010]. Many factors were against us. First we were dealing with the Pilot Projects which were aiming at deciding how to run the final Project. For example they sequenced the data of the

1st Pilot Project at coverage of 2x-4x which resulted to be too low. In the final project they will sequence at a coverage no lower than 6x.

Moreover, we were using data from a public database, which was not completely suited to our purposes. Most of the data was sequenced with Illumina Solexa platform which had too short reads to study microsatellites (see chapter 2.2).

Nonetheless we think that NGS is a very promising tool in the study of microsatellites. In principle it is the right technology to provide us enough data to discriminate between the various proposed models of microsatellite evolution. Moreover it will allow us to study the entire spectrum of microsatellites variability at different loci. This would allow us to identify previously unknown disease related alleles.

## 5.3 Inferring microsatellite functionality

### 5.3.1 Inferring microsatellite functionality experimentally

In the literature there are many experimental articles that describe microsatellite functionality. However these cases are not always easy to find, as often microsatellite functionality is a by product of the research on the functionality of a particular protein in a particular pathway. One clear example is the article of Alvarez et al [Alvarez et al., 2003] and the following article form Salichs et al [Salichs et al., 2009]. In the first article, while studying the functionality of the protein $DYRK1A$, Alvarez et al found that the histidine tract in this protein was responsible for targeting the protein to nuclear speckles. Nuclear speckles are nuclear compartments of which very little is known [Lamond and Spector, 2003].

Only later the same group decided to address the question of whether the histidine tract was a general signal to target proteins containing it to nuclear speckles. In a collaboration with our group,

all histidine containing proteins in the human genome were identified. For many of them , they checked if the wild-type proteins were able to localize to nuclear speckles. They checked as well if the same proteins were still able to localize to nuclear speckles once the histidine tract had been depleted. They found out that the lack of the histidine track impeded the protein to localize into nuclear speckles. So the histidine tract was both necessary and sufficient to target proteins containing it to nuclear speckles. Moreover they cloned a chimeric protein made by GFP and an histidine tract of variable length. They found that an histidine tract of length 6 or more is sufficient to target such a chimeric protein to nuclear speckles.

This is one of the many examples of how a study focused on a different subject resulted in a new discovery on microsatellites. We participated in the second study investigating the histidine containing proteins from a bioinformatics point of view. Moreover we compared the features of histidine repeats in proteins with the features of compositionally similar repeats located in non coding regions, concluding that those in proteins were much better conserved in other species [Salichs et al., 2009]. Since this study implied a big experimental effort we decided not to include it in the corpus of the thesis and we put it in the appendix.

### 5.3.2 Inferring microsatellite functionality from a bioinformatics point of view

Microsatellite conservation is an important indication of functionality and can be studied by bioinformatics means. It relies on the idea that microsatellites are very variable and in the absence of selective forces preserving them we would expect to see very little conservation when we look for human microsatellites in other mammalia or vertebrate genomes. On the contrary, if we see much more conservation than expected it means that negative selection is acting as to preserve microsatellites. So we can infer that highly conserved microsatellites are probably functional. Of course we need a null model to compare

with.

We tested this hypothesis in chapter 3 . There we compared the rate of conservation in vertebrate lineages of repeats that were located in protein coding regions and of compositionally similar sequences that were located in non-coding regions [Mularoni et al., 2010]. The latter were considered as the null hypothesis of evolution free of constraints. We found that repeats located in protein coding regions were more conserved in vertebrate lineages than similar sequences located in non coding regions. This confirmed our idea that negative selection was acting on potentially functional repeats. Moreover we retrieved a list of the most conserved repeats. Most of them had already been experimentally proven to be functional.

We followed this research line in chapter 4. There we studied the features of microsatellites in different non-coding genomic regions. Many experimental articles describe the functionality of microsatellites in regions such as promoters [Shimajiri et al., 1999, Vinces et al., 2009, Xie et al., 2005], introns [Gebhardt et al., 1999, Ejima et al., 2000] and UTRs [Lawson and Zhang, 2008, Raca et al., 2000, Robins and Press, 2005, Xie et al., 2005]. In these examples changes in microsatellite length alter the expression of the protein. We wanted to study the possibility that negative selection acted with different strength in the different genomics regions. We retrieved the microsatellites present in each genomic region. We looked for those repeats in the alignments with 9 eutherian species. We kept as conserved only the ones that were conserved in at least one of the primates and one of the rodents. For each genomic region we obtained the expected amount of conserved repeats if there was no negative selection acting. This expectation value was based on the repeat conservation we found in intergenic regions, where we expect negative selection not to act. We compared the distribution we found with the expected one, and we found that in some regions microsatellites were significantly more conserved than expected, strongly suggesting that some of them may be functional.

# Part V

# Conclusions

The main points of the work presented in this thesis can be summarized as follows:

1. Next generation sequencing technology, with a sufficiently high sequencing coverage and sufficiently long reads, represents a great opportunity to deepen our knowledge on intraspecific variability of microsatellites.

2. Amino acid tandem repeats in coding region tend to be relatively well conserved in vertebrate lineages with respect to similar sequences located in non coding regions.

3. Several of the well-most conserved repeats in vertebrate lineages has already been experimentally proven to be functional.

4. We found a significant depletion in CpG containing repeats with respect to compositionally similar sequences without the CpG motif. This can be explained by the conversion of methylated cytosine to thymine in the CpG context.

5. Promoters and 5'UTR gene regions are enriched in conserved trinucleotide repeats.

6. $(AT)_n$ repeats in 3'UTRs have been found to be 20 times more conserved than expected in mammalian genomes, suggesting that they may play a functional role.

# Part VI

# Appendix

# Chapter 6

# GENOME-WIDE ANALYSIS OF HISTIDINE REPEATS REVEALS THEIR ROLE IN THE LOCALIZATION OF HUMAN PROTEINS TO THE NUCLEAR SPECKLES COMPARTMENT

Salichs E, Ledda A, Mularoni L, Albà MM, de la Luna S. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. PloS Genet. 2009; 5(3): e1000397.

# Bibliography

[Albà and Guigó, 2004] Albà, M. and Guigó, R. (2004). Comparative analysis of amino acid repeats in rodents and humans. *Genome research*, 14(4):549.

[Alba et al., 1999] Alba, M., Santibanez-Koref, M., and Hancock, J. (1999). Conservation of polyglutamine tract size between mice and humans depends on codon interruption. *Mol Biol Evol*, 16(11):1641–1644.

[Albanese et al., 2001] Albanese, V., Biguet, N., Kiefer, H., Bayard, E., Mallet, J., and Meloni, R. (2001). Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite. *Human molecular genetics*, 10(17):1785.

[Alvarez et al., 2003] Alvarez, M., Estivill, X., and De La Luna, S. (2003). DYRK1A accumulates in splicing speckles through a novel targeting signal and induces speckle disassembly. *Journal of cell science*, 116(Pt 15):3099.

[Andrés et al., 2003] Andrés, A., Lao, O., Soldevila, M., Calafell, F., and Bertranpetit, J. (2003). Dynamics of CAG repeat loci revealed by the analysis of their variability. *Human mutation*, 21(1):61–70.

[Bacolla et al., 2008] Bacolla, A., Larson, J., Collins, J., Li, J., Milosavljevic, A., Stenson, P., Cooper, D., and Wells, R. (2008). Abundance and length of simple repeats in vertebrate genomes

are determined by their structural properties. *Genome research*, 18(10):1545.

[Becker et al., 2007] Becker, D., Vogelsang, D., and Brabetz, W. (2007). Population data on the seven short tandem repeat loci D4S2366, D6S474, D14S608, D19S246, D20S480, D21S226 and D22S689 in a German population. *International journal of legal medicine*, 121(1):78–81.

[Bentley et al., 2008] Bentley, D., Balasubramanian, S., Swerdlow, H., Smith, G., Milton, J., Brown, C., Hall, K., Evers, D., Barnes, C., Bignell, H., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

[Bhargava and Fuentes, 2010] Bhargava, A. and Fuentes, F. (2010). Mutational dynamics of microsatellites. *Molecular biotechnology*, 44(3):250–266.

[Bird, 1980] Bird, A. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499.

[Bovine et al., 2009] Bovine, G. et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science (New York, NY)*, 324(5926):522.

[Brook et al., 1992] Brook, J., McCurrach, M., Harley, H., Buckler, A., Church, D., Aburatani, H., Hunter, K., Stanton, V., Thirion, J., Hudson, T., et al. (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3'end of a transcript encoding a protein kinase family member. *Cell*, 68(4):799–808.

[Brown et al., 2005] Brown, L., Paraso, M., Arkell, R., and Brown, S. (2005). In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: alanine tract expansion modulates DNA binding and transactivation. *Human molecular genetics*, 14(3):411.

[Buchanan et al., 2004] Buchanan, G., Yang, M., Cheong, A., Harris, J., Irvine, R., Lambert, P., Moore, N., Raynor, M., Neufing, P., Coetzee, G., et al. (2004). Structural and functional consequences of glutamine tract variation in the androgen receptor. *Human molecular genetics*, 13(16):1677.

[Buschiazzo and Gemmell, 2006] Buschiazzo, E. and Gemmell, N. (2006). The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays*, 28(10):1040–1050.

[Buschiazzo and Gemmell, 2010] Buschiazzo, E. and Gemmell, N. (2010). Conservation of human microsatellites across 450 million years of evolution. *Genome Biology and Evolution*, 2010(0):153.

[Butland et al., 2007] Butland, S., Devon, R., Huang, Y., Mead, C., Meynert, A., Neal, S., Lee, S., Wilkinson, A., Yang, G., Yuen, M., et al. (2007). CAG-encoded polyglutamine length polymorphism in the human genome. *BMC genomics*, 8(1):126.

[Butler, 2006] Butler, J. (2006). Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of forensic sciences*, 51(2):253–265.

[Calkhoven et al., 1994] Calkhoven, C., Bouwman, P., Snippe, L., and Ab, G. (1994). Translation start site multiplicity of the CCAAT/enhancer binding protein $\alpha$ mRNA is dictated by a small 5' open reading frame. *Nucleic acids research*, 22(25):5540.

[Carbonell et al., 2005] Carbonell, S., Masi, L., Marini, F., Del Monte, F., Falchetti, A., Franceschelli, F., and Brandi, M. (2005). Genetics and pharmacogenetics of osteoporosis. *Journal of endocrinological investigation*, 28(10 Suppl):2.

[Chakraborty et al., 1997] Chakraborty, R., Kimmel, M., Stivers, D., Davison, L., and Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the*

*National Academy of Sciences of the United States of America*, 94(3):1041.

[Crow, 1993] Crow, J. (1993). How much do we know about spontaneous human mutation rates? *Environmental and molecular mutagenesis*, 21(2):122–129.

[Cummings and Zoghbi, 2000] Cummings, C. and Zoghbi, H. (2000). TRINUCLEOTIDE REPEATS: Mechanisms and Pathophysiology. *Annual review of genomics and human genetics*, 1(1):281–328.

[Durbin et al., 2010] Durbin, R., Altshuler, D., Abecasis, G., Bentley, D., Chakravarti, A., Clark, A., Collins, F., De La Vega, F., Donnelly, P., Egholm, M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

[Duret and Bucher, 1997] Duret, L. and Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. *Current Opinion in Structural Biology*, 7(3):399–406.

[Ejima et al., 2000] Ejima, Y., Yang, L., and Sasaki, M. (2000). Aberrant splicing of the ATM gene associated with shortening of the intronic mononucleotide tract in human colon tumor cell lines: a novel mutation target of microsatellite instability. *International Journal of Cancer*, 86(2):262–268.

[Ellegren, 2000] Ellegren, H. (2000). Microsatellite mutations in the germline::: implications for evolutionary inference. *Trends in Genetics*, 16(12):551–558.

[Ellegren, 2004] Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, 5(6):435–445.

[Farré et al., 2007] Farré, D., Bellora, N., Mularoni, L., Messeguer, X., and Albà, M. (2007). Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biology*, 8(7):R140.

[Faux et al., 2007] Faux, N., Huttley, G., Mahmood, K., Webb, G., Garcia de la Banda, M., and Whisstock, J. (2007). RCPdb: An evolutionary classification and codon usage database for repeat-containing proteins. *Genome research*, 17(7):1118.

[Flicek et al., 2011] Flicek, P., Amode, M., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2011). Ensembl 2011. *Nucleic acids research*, 39(suppl 1):D800.

[Fondon and Garner, 2004] Fondon, J. and Garner, H. (2004). Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(52):18058.

[Fujita et al., 2011] Fujita, P., Rhead, B., Zweig, A., Hinrichs, A., Karolchik, D., Cline, M., Goldman, M., Barber, G., Clawson, H., Coelho, A., et al. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*, 39(suppl 1):D876.

[Galant and Carroll, 2002] Galant, R. and Carroll, S. (2002). Evolution of a transcriptional repression domain in an insect Hox protein. *Nature*, 415(6874):910–913.

[Gatchel and Zoghbi, 2005] Gatchel, J. and Zoghbi, H. (2005). Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews Genetics*, 6(10):743–755.

[Gebhardt et al., 1999] Gebhardt, F., Zanker, K., and Brandt, B. (1999). Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *Journal of Biological Chemistry*, 274(19):13176.

[Gemayel et al., 2010] Gemayel, R., Vinces, M., Legendre, M., and Verstrepen, K. (2010). Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Genetics*, 44.

[Gerber et al., 1994] Gerber, H., Seipel, K., Georgiev, O., Hofferer, M., Hug, M., Rusconi, S., and Schaffner, W. (1994). Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science*, 263(5148):808.

[Haberman et al., 2008] Haberman, Y., Amariglio, N., Rechavi, G., and Eisenberg, E. (2008). Trinucleotide repeats are prevalent among cancer-related genes. *Trends in Genetics*, 24(1):14–18.

[Haerty and Golding, 2010a] Haerty, W. and Golding, G. (2010a). Genome-wide evidence for selection acting on single amino acid repeats. *Genome research*, 20(6):755.

[Haerty and Golding, 2010b] Haerty, W. and Golding, G. (2010b). Low-complexity sequences and single amino acid repeats: not just junk peptide sequences. *Genome*, 53(10):753–762.

[Hancock et al., 2001] Hancock, J., Worthey, E., and Santibáñez-Koref, M. (2001). A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice. *Molecular Biology and Evolution*, 18(6):1014.

[Harismendy et al., 2009] Harismendy, O., Ng, P., Strausberg, R., Wang, X., Stockwell, T., Beeson, K., Schork, N., Murray, S., Topol, E., Levy, S., et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*, 10(3):R32.

[Huntley and Clark, 2007] Huntley, M. and Clark, A. (2007). Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. *Molecular biology and evolution*, 24(12):2598.

[Imai and Yamamoto, 2008] Imai, K. and Yamamoto, H. (2008). Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. *Carcinogenesis*, 29(4):673.

[Jackson and Linsley, 2010] Jackson, A. and Linsley, P. (2010). Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nature Reviews Drug Discovery*, 9(1):57–67.

[Janody et al., 2001] Janody, F., Sturny, R., Schaeffer, V., Azou, Y., and Dostatni, N. (2001). Two distinct domains of Bicoid mediate its transcriptional downregulation by the Torso pathway. *Development*, 128(12):2281.

[Karlin et al., 2002] Karlin, S., Brocchieri, L., Bergman, A., Mrázek, J., and Gentles, A. (2002). Amino acid runs in eukaryotic proteomes and disease associations. *Proceedings of the National Academy of Sciences of the United States of America*, 99(1):333.

[Kashi and King, 2006] Kashi, Y. and King, D. (2006). Simple sequence repeats as advantageous mutators in evolution. *TRENDS in Genetics*, 22(5):253–259.

[Katti et al., 2001] Katti, M., Ranjekar, P., and Gupta, V. (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular Biology and Evolution*, 18(7):1161.

[Kelkar et al., 2010] Kelkar, Y., Strubczewski, N., Hile, S., Chiaromonte, F., Eckert, K., and Makova, K. (2010). What Is a Microsatellite: A Computational and Experimental Definition Based upon Repeat Mutational Behavior at A/T and GT/AC Repeats. *Genome Biology and Evolution*, 2(0):620.

[Kelkar et al., 2008] Kelkar, Y., Tyekucheva, S., Chiaromonte, F., and Makova, K. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome research*, 18(1):30.

[Kenneson et al., 2001] Kenneson, A., Zhang, F., Hagedorn, C., and Warren, S. (2001). Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in

intermediate-length and premutation carriers. *Human Molecular Genetics*, 10(14):1449.

[Knight et al., 1993] Knight, S., Flannery, A., Hirst, M., Campbell, L., Christodoulou, Z., Phelps, S., and Pointon, J. (1993). Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. *Cell*, 74(1):127–134.

[Koob et al., 1999] Koob, M., Moseley, M., Schut, L., Benzow, K., Bird, T., Day, J., and Ranum, L. (1999). An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nature genetics*, 21:379–384.

[Kozlowski et al., 2010] Kozlowski, P., De Mezer, M., and Krzyzosiak, W. (2010). Trinucleotide repeats in human genome and exome. *Nucleic Acids Research*.

[Lai and Sun, 2003] Lai, Y. and Sun, F. (2003). The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular biology and evolution*, 20(12):2123.

[Lamond and Spector, 2003] Lamond, A. and Spector, D. (2003). Nuclear speckles: a model for nuclear organelles. *Nature Reviews Molecular Cell Biology*, 4(8):605–612.

[Lander et al., 2001] Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Lanz et al., 1995] Lanz, R., Wieland, S., Hug, M., and Rusconi, S. (1995). A transcriptional repressor obtained by alternative translation of a trinucleotide repeat. *Nucleic acids research*, 23(1):138.

[Lavoie et al., 2003] Lavoie, H., Debeane, F., Trinh, Q., Turcotte, J., Corbeil-Girard, L., Dicaire, M., Saint-Denis, A., Page, M.,

Rouleau, G., and Brais, B. (2003). Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. *Human molecular genetics*, 12(22):2967.

[Lawson and Zhang, 2008] Lawson, M. and Zhang, L. (2008). Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5'-UTR region. *Gene*, 407(1-2):54–62.

[Leclercq et al., 2010] Leclercq, S., Rivals, E., and Jarne, P. (2010). DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome Biology and Evolution*, 2(0):325.

[Leek et al., 2010] Leek, J., Scharpf, R., Bravo, H., Simcha, D., Langmead, B., Johnson, W., Geman, D., Baggerly, K., and Irizarry, R. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739.

[Leibovitch et al., 2002] Leibovitch, B., Lu, Q., Benjamin, L., Liu, Y., Gilmour, D., and Elgin, S. (2002). GAGA factor and the TFIID complex collaborate in generating an open chromatin structure at the Drosophila melanogaster hsp26 promoter. *Molecular and cellular biology*, 22(17):6148.

[Levinson and Gutman, 1987] Levinson, G. and Gutman, G. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular biology and evolution*, 4(3):203.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078.

[Li et al., 2002a] Li, Y., Korol, A., Fahima, T., Beiles, A., and Nevo, E. (2002a). Microsatellites: genomic distribution, putative func-

tions and mutational mechanisms: a review. *Molecular Ecology*, 11(12):2453–2465.

[Li et al., 2002b] Li, Y., Korol, A., Fahima, T., Beiles, A., and Nevo, E. (2002b). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, 11(12):2453–2465.

[Li et al., 2004] Li, Y., Korol, A., Fahima, T., and Nevo, E. (2004). Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution*, 21(6):991.

[Liquori et al., 2001] Liquori, C., Ricker, K., Moseley, M., Jacobsen, J., Kress, W., Naylor, S., Day, J., and Ranum, L. (2001). Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science*, 293(5531):864.

[Madsen et al., 2008] Madsen, B., Villesen, P., and Wiuf, C. (2008). Short tandem repeats in human exons: a target for disease mutations. *BMC genomics*, 9(1):410.

[Mardis, 2008] Mardis, E. (2008). Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, 9(1):387.

[Matsuura et al., 2000] Matsuura, T., Yamagata, T., Burgess, D., Rasmussen, A., Grewal, R., Watase, K., Khajavi, M., McCall, A., Davis, C., Zu, L., et al. (2000). Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nature genetics*, 26(2):191–194.

[McIver et al., 2011] McIver, L., Fondon III, J., Skinner, M., and Garner, H. (2011). Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics*, Epub ahead of print.

[Messier et al., 1996] Messier, W., Li, S., and Stewart, C. (1996). The birth of microsatellites. *Nature*, 381:483.

[Metzker, 2005] Metzker, M. (2005). Emerging technologies in DNA sequencing. *Genome research*, 15(12):1767.

[Metzker, 2009] Metzker, M. (2009). Sequencing technologiesâthe next generation. *Nature Reviews Genetics*, 11(1):31–46.

[Mirkin, 2007] Mirkin, S. (2007). Expandable DNA repeats and human disease. *Nature*, 447(7147):932–940.

[Mularoni et al., 2006a] Mularoni, L., Guigó, R., and Albà, M. (2006a). Mutation patterns of amino acid tandem repeats in the human proteome. *Genome Biology*, 7(4):R33.

[Mularoni et al., 2006b] Mularoni, L., Guigó, R., and Albá, M. (2006b). Mutation patterns of amino acid tandem repeats in the human proteome. *Genome Biology*, 7(4):R33.

[Mularoni et al., 2010] Mularoni, L., Ledda, A., Toll-Riera, M., and Albà, M. (2010). Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome research*, 20(6):745.

[Mularoni et al., 2008] Mularoni, L., Toll-Riera, M., and Albà, M. (2008). Comparative genetics of trinucleotide repeats in the human and ape genomes. *Encyclopedia of life sciences*.

[Mularoni et al., 2007] Mularoni, L., Veitia, R., and Albà, M. (2007). Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics*, 89(3):316–325.

[Ohno, 1972] Ohno, S. (1972). So much" junk" DNA in our genome. In *Brookhaven symposia in biology*, volume 23, page 366.

[Ohshima et al., 1998] Ohshima, K., Montermini, L., Wells, R., and Pandolfo, M. (1998). Inhibitory Effects of Expanded GAA· TTC Triplet Repeats from Intron I of the Friedreich Ataxia Gene on Transcription and Replicationin Vivo. *Journal of Biological Chemistry*, 273(23):14588.

[Orgel and Crick, 1980] Orgel, L. and Crick, F. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284(5757):604–607.

[Payseur et al., 2010] Payseur, B., Jing, P., and Haasl, R. (2010). A Genomic Portrait of Human Microsatellite Variation. *Molecular Biology and Evolution*, 28(1):303–12.

[Pérez-Lezaun et al., 2000] Pérez-Lezaun, A., Calafell, F., Clarimón, J., Bosch, E., Mateu, E., Gusmao, L., Amorim, A., Benchemsi, N., and Bertranpetit, J. (2000). Allele frequencies of 13 short tandem repeats in population samples from the Iberian Peninsula and Northern Africa. *International Journal of Legal Medicine*, 113(4):208–214.

[Pool et al., 2010] Pool, J., Hellmann, I., Jensen, J., and Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome research*, 20(3):291.

[Pop and Salzberg, 2008] Pop, M. and Salzberg, S. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, 24(3):142–149.

[Prakash and Tompa, 2005] Prakash, A. and Tompa, M. (2005). Discovery of regulatory elements in vertebrates through comparative genomics. *Nature biotechnology*, 23(10):1249–1256.

[Pupko and Graur, 1999] Pupko, T. and Graur, D. (1999). Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units. *Journal of molecular evolution*, 48(3):313–316.

[Raca et al., 2000] Raca, G., Siyanova, E., McMurray, C., and Mirkin, S. (2000). Expansion of the (CTG) n repeat in the 5'-UTR of a reporter gene impedes translation. *Nucleic Acids Research*, 28(20):3943.

[Ranum and Cooper, 2006] Ranum, L. and Cooper, T. (2006). RNA-mediated neuromuscular disorders. *Neuroscience*, 29.

[Richardson, 2006] Richardson, J. (2006). Fjoin: simple and efficient computation of feature overlaps. *Journal of Computational Biology*, 13(8):1457–1464.

[Robins and Press, 2005] Robins, H. and Press, W. (2005). Human microRNAs target a functionally distinct population of genes with AT-rich 3' UTRs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15557.

[Rodriguez et al., 2002] Rodriguez, M., Núŋez-Roldán, A., Aguilar, F., Valenzuela, A., Garcia, A., and Gonzalez-Escribano, M. (2002). Association of the CTLA4 3'untranslated region polymorphism with the susceptibility to rheumatoid arthritis. *Human immunology*, 63(1):76–81.

[Rose and Falush, 1998] Rose, O. and Falush, D. (1998). A threshold size for microsatellite expansion. *Molecular biology and evolution*, 15(5):613.

[Rubinsztein et al., 1999] Rubinsztein, D., Amos, B., and Cooper, G. (1999). Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 354(1386):1095.

[Salichs et al., 2009] Salichs, E., Ledda, A., Mularoni, L., Albà, M., and De La Luna, S. (2009). Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet*, 5(3):e1000397.

[Schlotterer, 2000] Schlotterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma*, 109(6):365–371.

[Shah et al., 2010] Shah, S., Hile, S., and Eckert, K. (2010). Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. *Cancer Research*, 70(2):431.

[Sharma et al., 2007a] Sharma, P., Grover, A., and Kahl, G. (2007a). Mining microsatellites in eukaryotic genomes. *TRENDS in Biotechnology*, 25(11):490–498.

[Sharma et al., 2007b] Sharma, V., Kumar, N., Brahmachari, S., and Ramachandran, S. (2007b). Abundance of dinucleotide repeats and gene expression are inversely correlated: a role for gene function in addition to intron length. *Physiological genomics*, 31(1):96.

[Shendure and Ji, 2008] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–1145.

[Shimajiri et al., 1999] Shimajiri, S., Arima, N., Tanimoto, A., Murata, Y., Hamada, T., Wang, K., and Sasaguri, Y. (1999). Shortened microsatellite d (CA) 21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS letters*, 455(1-2):70.

[Simmen, 2008] Simmen, M. (2008). Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. *Genomics*, 92(1):33–40.

[Simon and Hancock, 2009] Simon, M. and Hancock, J. (2009). Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol*, 10(6):R59.

[Strand et al., 1993] Strand, M., Prolla, T., Liskay, R., and Petes, T. (1993). Destabilization of tracts of simple repetitive DNA

in yeast by mutations affecting DNA mismatch repair. *Nature*, 365(6443):274.

[Subirana and Messeguer, 2010] Subirana, J. and Messeguer, X. (2010). The most frequent short sequences in non-coding DNA. *Nucleic Acids Research*, 38(4):1172.

[Subramaniam et al., 2010] Subramaniam, K., Ooi, L., and Hui, K. (2010). Transcriptional down-regulation of IGFBP-3 in human hepatocellular carcinoma cells is mediated by the binding of TIA-1 to its AT-rich element in the 3'-untranslated region. *Cancer Letters*, 297(2):259–268.

[Subramanian et al., 2003a] Subramanian, S., Madgula, V., George, R., Mishra, R., Pandit, M., Kumar, C., and Singh, L. (2003a). Triplet repeats in human genome: distribution and their association with genes and other genomic regions. *Bioinformatics*, 19(5):549.

[Subramanian et al., 2003b] Subramanian, S., Mishra, R., and Singh, L. (2003b). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol*, 4(2):R13.

[Sumiyama et al., 1996] Sumiyama, K., Washio-Watanabe, K., Saitou, N., Hayakawa, T., and Ueda, S. (1996). Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals. *Journal of molecular evolution*, 43(3):170–178.

[Taylor et al., 1999] Taylor, J., Durkin, J., and Breden, F. (1999). The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. *Molecular biology and evolution*, 16(4):567.

[Team, 2008] Team, R. (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing Vienna Austria ISBN*, 3(10).

[Timchenko et al., 1999] Timchenko, N., Lu, A., and Timchenko, L. (1999). CUG repeat binding protein (CUGBP1) interacts with the 5'region of C/EBP$\beta$ mRNA and regulates translation of C/EBP$\beta$ isoforms. *Nucleic acids research*, 27(22):4517.

[Tóth et al., 2000] Tóth, G., Gáspári, Z., and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, 10(7):967.

[Toutenhoofd et al., 1998] Toutenhoofd, S., Garcia, F., Zacharias, D., Wilson, R., and Strehler, E. (1998). Minimum CAG repeat in the human calmodulin-1 gene 5'untranslated region is required for full expression. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1398(3):315–320.

[Usdin and Grabczyk, 2000] Usdin, K. and Grabczyk, E. (2000). DNA repeat expansions and human disease. *Cellular and Molecular Life Sciences*, 57(6):914–931.

[Venter et al., 2001] Venter, J., Adams, M., Myers, E., Li, P., Mural, R., Sutton, G., Smith, H., Yandell, M., Evans, C., Holt, R., et al. (2001). The sequence of the human genome. *science*, 291(5507):1304.

[Vinces et al., 2009] Vinces, M., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, 324(5931):1213.

[Weber and Wong, 1993] Weber, J. and Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics*, 2(8):1123.

[Webster et al., 2002] Webster, M., Smith, N., and Ellegren, H. (2002). Microsatellite evolution inferred from human–chimpanzee genomic sequence alignments. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13):8748.

[Woodside et al., 2006] Woodside, M., Behnke-Parks, W., Larizadeh, K., Travers, K., Herschlag, D., and Block, S. (2006). Nanomechanical Measurements of the Sequence-Dependent Folding Landscapes of Single Nucleic Acid Hairpins. *Proceedings of the National Academy of Sciences of the United States of America*, 103(16):6190–6195.

[Wren et al., 2000] Wren, J., Forgacs, E., Fondon III, J., Pertsemlidis, A., Cheng, S., Gallardo, T., Williams, R., Shohet, R., Minna, J., and Garner, H. (2000). Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *The American Journal of Human Genetics*, 67(2):345–356.

[Xie et al., 2005] Xie, X., Lu, J., Kulbokas, E., Golub, T., Mootha, V., Lindblad-Toh, K., Lander, E., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals. *Nature*, 434(7031):338–345.

[Yamashita et al., 2005] Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. (2005). Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene*, 350(2):129–136.