Tandem repeat variation in human and great ape populations and its impact on gene expression divergence

Tiago Loureiro de Carvalho

TESI DOCTORAL UPF / ANY 2015

DIRECTOR DE LA TESI

Dr. Tomàs Marquès-Bonet

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

upf. Universitat Pompeu Fabra Barcelona

Acknowledgments

The past three years have been incredible in all possible senses, and have allowed me to grow both at the scientific and personal level. This was in part only possible thanks to those of you with whom I had the privilege to interact during the duration of my PhD.

Em primeiro lugar gostaria de agradecer à minha família por todo o amor e encorajamento que mostraram, e também por serem sempre tão compreensivos comigo. Acima de tudo obrigado por me darem liberdade e me apoiarem em todas as escolhas que faço.

Também gostaria de agradecer a todos os meus amigos em Portugal que mesmo estando longe sempre me pareceram estar perto. Da mesma maneira agradeço aos amigos portugueses em Barcelona, não só pela amizade, mas também por me fazerem sentir não estar tão longe de casa.

I also want to thank the people from the old 407 office, now 421, for their friendship, patience and help. My work days would not be the same without your companionship, and it was truly a great pleasure to work surrounded by you. This goes to the ones present, and the ones who already departed, Marcos, Irene, Jéssica, Raquel, Javier, Quilez, Marta, Bélen, Dabad, Lóbon, Marc, Lukas, Claudia and Aitor.

I am too thankful for the support and friendship from all the friends at PRBB, too many to mention. Thank you for all the good moments, the football and volley matches, the climbing sessions, the many “chocolate now”, the dinners, trips, and parties and the occasional beer or coffee. I will really miss these times!

I am further obliged to thank the people with whom I collaborated in the work presented in this thesis, in particular to Tugce Bilgin and Andreas Wagner, but also to Andrew Sharp, David Mittelman, Michael Krutzen Gareth Highnam, David Comas, Maja Greminger, and Mark Robinson.

Last but not least, I would also like to personally thank Tomàs Marques-Bonet, for accepting to be my supervisor, and granting me the incredible opportunity to come and do my PhD in Barcelona. Without you none of this would be possible. Coming here was the beginning of a fantastic adventure, and I must really thank you not only for believing in me, but also and for all the support and guidance provided.

This thesis is dedicated to all of you.

Abstract

Genetic variation in humans and the great apes has been amply explored using a wide variety of markers, among them tandem repeats (TRs). Because of the nature of TRs, highly variable in length due to its high mutation rate, they are an important source of genetic variation, and thus especially informative in fields such as population and conservation genetics. Particularly, they are still often used to illuminate natural populations complex evolutionary histories and structure.

TR variation is also associated with several pathological conditions, and hypothesized to have an important role in the evolution of gene regulation.

In this work a recently developed TR genotyping algorithm was applied on human and nonhuman great apes whole-genome sequencing data. The analysis of the TR variation indicate that this information is useful to describe fine scale population variation, and hints at a substantial contribution of TRs to gene expression divergence during great apes evolution.

Resumen

La variación genética en los seres humanos y grandes simios ha sido amplamente explorada usando una grande variedad de marcadores, entre ellos repeticiones en tándem (RT). Debido a la naturaleza de las RT, muy variables en longitud debido a su alta tasa de mutación, estas constituen una importante fuente de variación genética, y por lo tanto altamente informativas en áreas como la genética de poblaciones y de la conservación. En particular, a menudo aún se utilizan para elucidar las complejas historias evolutivas de las poblaciones naturales y su estructura genética.

La variación de RT está también asociada con varias enfermedades, y se cree que desempeña un papel importante en la evolución de la regulación génica.

En este trabajo un algoritmo desarrollado recientemente que genotipa RT a nivel de todo el genoma, se aplicó sobre datos de secuenciación de genomas humanos y de grandes simios. La analisis de la variacion de RT sugiere que esta información es útil para describir la variación en populaciones, y alude a una aportación sustancial de las RT a la divergencia de expresión génica durante la evolución de los grandes simios.

Preface

Traditionally, tandem repeats have been thought mostly of as sequences which mutate neutrally, with some propensity to cause disease if extreme variation takes place within or nearby protein-coding regions. Because of their mutational behaviour, they have been mostly used as molecular markers in population and conservation genetics.

However, this classical view is starting to be overturned by studies which show that this abundant source of genetic variation can also have a role as a gene expression modulator. Importantly, because the changes introduced are frequent and occur in a gradual and readily reversible manner, they can allow for precise attunement of the genome, with minimal impact on the genetic load.

Until very recently, genotyping these repetitive elements was a laborious and costly effort, and hence, the repeat landscape has remained largely unexplored. This has changed with the development of technologies that allow entire genomes to be sequenced, and algorithms that can use this information to accurately infer repeat genotypes.

The recent availability of whole-genome data from humans and our closest relatives, allows for the first time to describe the genome-wide tandem repeat variation in these populations. This information will help determine the impact of tandem repeats on expression divergence, and clarify their role on the great apes evolution. In addition, it will also be an useful resource for conservation efforts, since it provides many new molecular markers which are informative for determining subspecies and even geographical origin of great apes.

Table of Contents

1. INTRODUCTION

1.1 Tandem repeats

The evolution of eukaryotes is an ongoing process that results from the accumulation of genomic changes that have been occurring for over 1.6–2.1 billion years. These changes provide the substratum upon which natural selection and genetic drift act on, shaping the genomes of all extant eukaryotes, and range from single nucleotide mutations to structural variation that involves the duplication of entire genomes. In-between these two extremes of DNA's sequence variation spectrum, there is a type of variation that consists of repetitive regions which can range from a few nucleotides to several megabases in size. These regions are an abundant and important source of variation in eukaryote genomes, and make up to half of the human genome (Treangen and Salzberg 2012).

A subset of these DNA elements are termed tandem repeats, small stretches of DNA in which a repeat unit is repeated several times side by side. These are also overrepresented in many genomes, and in humans account for up to 3% of the genome (Lander et al. 2001). Since many evolve in a neutral fashion, and without any recognizable function, they have been mostly regarded as “selfish DNA” and used as neutral molecular markers. However, these simple repetitive regions are increasingly being recognized as having the potential to play an important role in genome evolution.

1.1.1 Simple sequences with remarkable properties

Tandem repeats (TRs) are commonly described as DNA sequences in which a nucleotide motif, usually taken to mean a DNA sequence as small as one base pair (bp) and up to 60 bps, is repeated several times in an head-to-tail pattern. An example of a TR is the sequence TGTGTGTGTGTG in which the motif TG is repeated numerous times. If the biological mechanisms dictating genome composition were governed by completely random processes one would expect this type of sequence to occur very seldom in any given genome. However, this and other variations of repeated nucleotide motifs can be found across eukaryote genomes, mainly in intergenic regions and introns, but also in coding regions. The reason for their ubiquity remains elusive, but the finding that many are often highly conserved suggests that these simple DNA elements may also have a functional role. This suggestion is supported by the observation that TR variation is associated not only with several human pathologies, such as cancer and neurodegenerative diseases (Pearson et al. 2005; Usdin 2008), but also with striking morphological and behavioral phenotypical changes in a wide range of organisms (Hammock and Young 2005; Fondon et al. 2008).

These occur as a result of mutations that alter the number of repeat unit copies, either by expanding or contracting the TR, and are often mediated by a process called replication slippage (Figure 1).

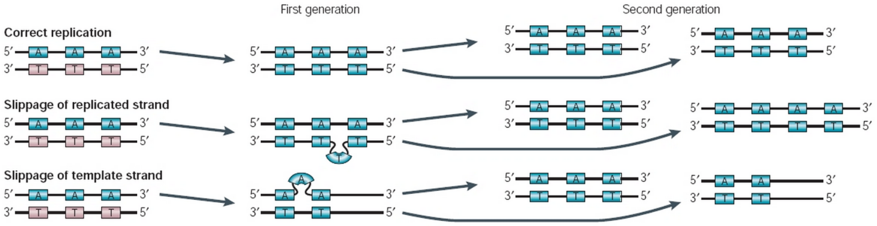

Figure 1. - Replication slippage. During replication the two DNA strands might detach, and a hairpin loop may form in either strand so that these strands will realign out of register (First generation). If the loop forms in the replicate strand (bottom strand), such as portrayed by the example in the middle, once the replication is over this strand will have gained an extra repeat copy and thus increased its length (Second generation). Otherwise if the loop is formed in the template strand (upper strand), such as the example at the bottom, upon the replication is finished the replicated strand will have decreased in length relatively to the template (Second generation) (Thomson et al. 2003).

The high frequency that characterizes this type of replication error underlies the rapid accumulation of mutations across TRs. Due to this high mutation rate, reported to be 1 x 10^{-4} to 1 x 10^{-3} mutations per locus per generation (Sun et al. 2012), many TRs are highly polymorphic and thus often multiallelic, making TRs one of the largest source of genomic variation.

Previous research seems to indicate that the factors which determine how stable a TR is, i.e. its mutability, are the number of repeat units, the size of each repeat unit, and the purity degree of the repeat stretch (Figure 2) (Legendre et al. 2007). Those TRs with more repeat units are generally more unstable, and as are those whose purity is higher, i.e. those TR stretches where few to none point mutations or indels have accumulated.

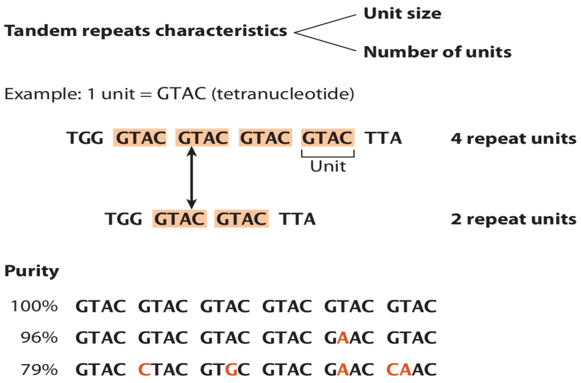
Tandem repeats characteristics
Unit size
Number of units
Example: 1 unit = GTAC (tetranucleotide)
TGG GTAC GTAC GTAC GTAC TTA
4 repeat units
Unit
TGG GTAC GTAC TTA
2 repeat units
Purity
100% GTAC GTAC GTAC GTAC GTAC GTAC
96% GTAC GTAC GTAC GTAC GAAC GTAC
79% GTAC CTAC GTGC GTAC GAAC CAAC

Figure 2. - TRs are usually defined based both on their repeat unit size, and the number of times these units are repeated. In the example the repeat motif GTAC, whose unit size is four base pairs, is repeated four and two times. TRs can also be classified based on their purity. A repeat tract consisting uniquely of the same repeat motif is considered 100% pure. As it accumulates point mutations, its degree of purity decreases. This measure is commonly computed by dividing the number of changes introduced into the repeat tract by its entire length. (Figure adapted from Gemayel et al. 2010).

Since TRs mutate up to ten thousand times faster than point mutations (Weber and Wong 1993; Brinkmann et al. 1998; Li et al. 2002; Legendre et al. 2007), they are more bound to have multiple alleles per locus and a high heterozygosity. For this reason, with very few molecular markers it is possible to use TRs to perform

DNA fingerprinting, as well as genetic mapping, calculate kinship coefficients, or describe population diversity (Ellegren 2004).

The combination of the two properties described above, i.e. functional potential and high mutation rate, make TRs a rich source of novel genetic variation that can fuel new adaptative breakthroughs. Importantly, because many TR mutations occur in small steps, they often translate into minor and often tolerable changes in protein function or gene expression. This is in contrast to point mutations which often have a strong and deleterious effect, and thus evolve slower than TRs. As a consequence, TRs may provide an efficient mechanism by which populations can deal with new biological and environmental challenges.

1.1.2 Importance of tandem repeats

a) Molecular markers with an influential role in primate genomics

Genetic markers have been around ever since Alfred Sturtevant used phenotypic markers to develop the first genetic map of *Drosophila Melanogaster* over a century ago (Sturtevant 1913). Since then many types of molecular markers have been developed, with TRs, discovered in the 1980's, occupying a prominent role and widespread use in forensic and population genetic fields (Schlötterer 2004). Their popularity stems from their abundance and high genetic polymorphism. Because of TRs high polymorphism, characterized by high heterozygosity and multiple allele per loci, TRs can leave a trace in genomes even at short time scales. As a result, tasks that may be difficult to perform even with a reasonable number of SNPs, such as detecting recent shifts in genetic diversity, or identity-by-descent analyses, are feasible even with few TR markers. Consequently, TRs contribution has been particularly relevant in the study of nonhuman primates genetic diversity and their conservation.

Indeed, much of our knowledge of nonhuman great apes genetic diversity stems from studies which focused on the variation not only of mitochondrial DNA (Ferris et al. 1981; Garner and Ryder 1996; Vigilant and Bradley 2004; Stone et al. 2010; Zsurka et al. 2010; Bjork et al. 2011; Fischer et al. 2011; Hvilsom et al. 2013, 2014), but also of nuclear microsatellite loci. The latter have been useful to understand the relationships between fragmented great apes populations, by assessing their population structure and/or migration patterns (Reinartz et al. 2000; Warren et al. 2000; Zhang et al. 2001; Becquet et al. 2007; Bergl and Vigilant 2007; Arora et al. 2010; Fünfstück et al. 2014; Nater et al. 2013; Roy et al. 2014), as well as to estimate the decline of these natural populations (Goossens et al. 2006), and evolutionary history (Wegmann and Excoffier 2010). This view has only recently been complemented with a range of publications including now a comprehensive catalog of great ape diversity using nuclear SNP datasets (Locke et al. 2011; Vallender 2011; Prado-Martinez et al. 2013; Scally et al. 2013; Greminger et al. 2014; McManus et al. 2015).

Because different marker types are characterized by different evolutionary rates and modes, they can give us complementary insights into the evolutionary history and present diversity patterns within and between closely related species. In this regard, analyses of the full spectrum of genomic variation in human and nonhuman great ape populations are critical, and in that regard have already proved very informative (Prado-Martinez et al. 2013; Sudmant et al. 2013; Hormozdiari et al. 2013). In light of these facts, a more complete description of the repeat landscape in great apes can help us further understand how humans differ from their closest relatives, and how human populations have been shaped by processes such as natural selection and demographic history. This information will also be particularly valuable for conservation efforts of nonhuman primates, since proper population management is greatly enhanced by the availability of molecular markers that allow for efficient diversity assessment from, for example, non-invasive samples. In particular, since all great ape species have been classified either as endangered or critically endangered (IUCN 2015), and are increasingly threatened by poaching, deforestation and disease, efforts that aim to preserve their diversity in the wild, and prevent inbreeding depression, are crucial and can be guided by the use of this genomic information about the populations.

b) Role in disease

A major argument in favour of the thorough characterization of TR variation within and across populations and species lies in their pathological potential. Pathological phenotypes can be mediated both by variation within exons as well as outside of the open reading frame (ORF), and associated with the disruption of several distinct molecular processes (Figure 3). Namely, TR variation has been associated with gene expression modulation and alteration of the structure and function of RNAs and proteins (Hannan 2010).

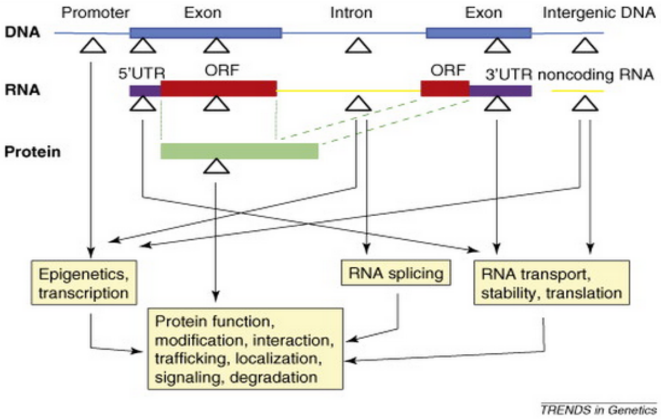
Promoter
Exon
Intron
Exon
Intergenic DNA
DNA
5'UTR
ORF
ORF
3'UTR noncoding RNA
RNA
Protein
Epigenetics, transcription
RNA splicing
RNA transport, stability, translation
Protein function, modification, interaction, trafficking, localization, signaling, degradation
TRENDS in Genetics

Figure 3. - Examples of how TR variation may affect several molecular processes, depending on its genomic location, and at what stage these may occur. Here TRs (triangles), are found to have molecular consequences at several levels independently if they are in coding or in non-coding regions. (Hannan 2010).

Among the more than 40 neurodegenerative diseases reported to be associated with TRs, two of the most widely described are two which occur as a result of a polyglutamine repeat expansion, Huntington's Disease (HD) and Fragile X Syndrome (FXS), (Pearson et al. 2005).

These two conditions illustrate the disparate ways by which abnormal TR variation can have pathological consequences depending on where on the genome it occurs.

On the one hand, HD is part of a set of disorders in which the TR is located on the ORF. Disorders of this type are typically associated with mutation of triplet repeats, since repeat motif lengths which are not multiple of three are more prone to induce frame-shift and thus are largely absent from ORF's. However, even if no frame-shift mutation occurs, such as in HD, when TRs surpass a given threshold length, protein conformational changes may occur which translate into the acquisition of toxic properties, or lead to protein malfunction. For example, individuals possessing 7-34 CAG repeats in the IT15 (interesting transcript 15) of the huntingtin gene show a normal phenotype, high risk of developing HD if they have between 36 and 39 repeats, and when the repeat count is beyond 39 several protein associated processes, such as folding, cleavage, interactions with other proteins, trafficking and degradation, are affected, leading to HD development (Gemayel et al. 2010). In addition, in this and many other repeat associated diseases, increased juvenile onset and higher severity is positively correlated with the number of repeats.

On the other hand, FXS typifies a type of disorder where the TR is located in a non-coding region, and its abnormal length leads to epigenetic changes and differential gene transcriptional regulation. In the case of FXS, the *FMR1* gene, which contains a TR on its 5′-untranslated region (UTR), exhibits a normal phenotype when the TR has between 6 and 53 repeats, and leads to disease by transcription silencing if it is beyond 200 repeats. This occurs because of increased methylation of the CpG island, which results in the inhibition of transcription factor binding. Moreover, if the repeat number is between 55-200 repeats, these TRs become so-called «premutation alleles», given that the probability of pathological allele expansion increases, and may be associated with fragile X tremor/ataxia syndrome and autism spectrum disorder, as a result of increased gene transcription (Usdin 2008).

The examples presented portray typical mechanisms by which TRs can inflict disease, but many others exist. These include, among several others, induction of chromosomal fragility, which can result in chromosome breakage and translocation, generation of a more open chromatin architecture, which can alter the transcription rate, or by serving as target for the ribonuclease Dicer, producing CUG repeats that are involved in RNA interference (Usdin 2008).

While TRs have been implicated in several diseases, there is currently an undeniable prevalence of TR variation associated with neurodegenerative conditions. Although this observation might represent some kind of bias, one intriguing possibility is that during recent human evolution, selective forces might have favored variable TRs in genes involved in neurological functions. However, due to the inherent instability of these TRs many were eventually pushed into “premutation” boundaries. In this scenario, what could have started as an evolutionarily advantage, resulting from a mechanism that enhances phenotypical plasticity, might have ended having possible pathological consequences.

c) Role in evolutionary adaptation

One of the least known features of TRs is their adaptative potential. Because many TRs appear to evolve in a selectively neutral fashion, as evidenced by the ubiquity of polymorphic loci at the population level in many taxa, for some time it seemed implausible that they could have any significant functional impact. In addition, the association between TR variation and pathological phenotypes may seem at variance with a potential role in adaptation. However, studies reporting on TRs high abundance and phylogenetic conservation suggest that many might be of functional importance (Schaper et al. 2015). Furthermore, evidence collected over the past thirty years advocate for a role of TRs in the fine attunement of gene expression and function in the genome. In addition, the evidence seems to suggest that TRs occasional pathogenicity represents a rare event which lies on the extreme of the phenotypical variation spectrum.
Indeed, past studies have shown that TRs may underlie phenotypical variation of several traits ranging from vole behaviour to sporulation in yeast (Hammock and Young 2005; Vinces et al. 2009), and have also highlighted the large selection of molecular processes that TRs can affect. These processes mainly take place at the RNA and protein level, and include the regulation of

transcription rates and stability, as well as the way proteins fold, interact or even degrade (Fondon and Garner 2004).

i) Mechanisms

Notably, TR variation exhibits two main modes of exerting an effect at the phenotypical level. If the outcome is binary, i.e. a phenotype is expressed or not, TR variation is said to function like a genetic switch (ON or OFF), on the other hand, if changes in repeat number translate into quantifiable phenotypical changes, TR variation is said to function like an "evolutionary tuning knob".

One way by which TRs can act as genetic switches (ON or OFF) is whenever, for example, their mutation happens to induce frame-shifting (Figure 4). One example which clearly shows how such genetic switch might be useful in nature, is exemplified by the mechanism some bacteria have developed to evade its host defense system during infection. Through a mechanism called phase variation, the random and reversible TR mutation that leads to gain and loss of a particular phenotype, some variants may quickly arise in a bacterial population so that at least some of its constituents survive its host (Kita et al. 1991).

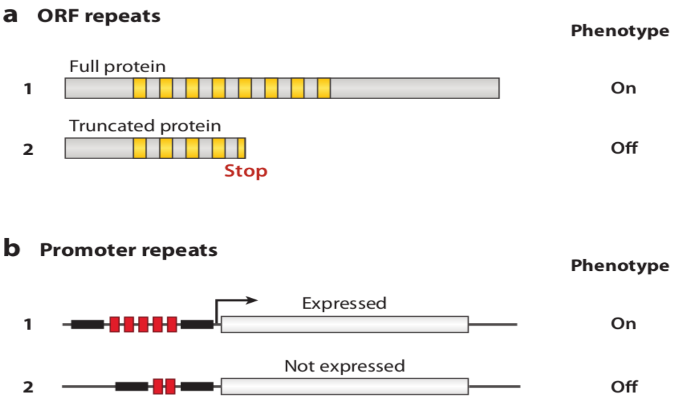

Figure 4. - Mechanisms by which TRs may function as genetic switches. a) Variation of TRs inside a gene coding region can generate nonfunctional or mistranslated proteins due to frameshift induction. b) TR variation at the promoter can also affect RNA polymerase binding sites, and thus determine if a gene is expressed or not. (Gemayel et al. 2010).

When TRs behave as ”evolutionary tuning knobs”, the mode in which TR variation affects the phenotype might vary, so it is important to first disaggregate the term.

Specifically, this term can be broken down into three other terms that better illustrate the different ways in which TR variation modulates a given phenotypical output. In particular, TRs can function as a “volume knobs” if the repeat copy number is correlated with the phenotype, as “tuning knobs” if this relationship is not linear so that its behaviour is more akin to that of a radio tuning dial, and finally as “optimality knobs” if the correlation is reversed beyond some threshold copy number TRs (Figure 5) (Elmore et al. 2012).

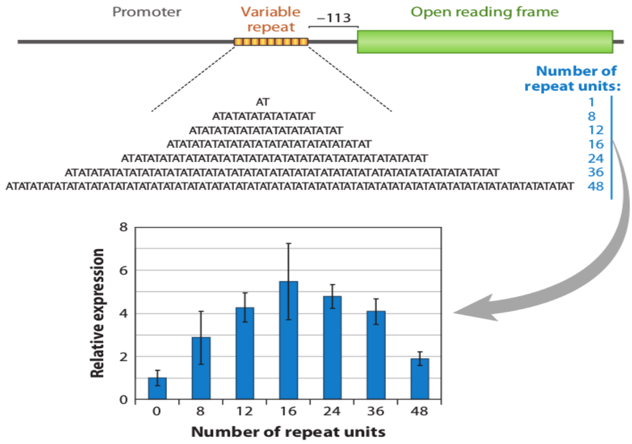

Figure 5. - Gradual changes in the length of a TR tract located inside a promoter translate into gradual differences in gene expression. This association is not necessarily monotonic, as the example shows. (Gemayel et al. 2010)

Among the many ways in which this variation may lead to expression changes, some of the currently described in the scientific literature include induction of structural modifications at the DNA and RNA level, and alteration of the regulatory portions of genes. In particular, TRs are responsible by formation of Z-DNA structures

(Naylor and Clark 1990; Rothenburg et al. 2001), known to be involved in gene regulation, and of secondary structures in RNA which can affect the processing, stability and translation of mRNA (Galvão et al. 2001; Tian et al. 2000). In addition, TRs can modulate the extent to which regulatory proteins access transcription factor binding sites (Martin et al. 2005; Vinces et al. 2009), and other regulatory regions, by varying the size of the former sites, and by either altering the chromatin landscape (Godde and Wolffe 1996; Sandman and Reeve 1999; Tomita et al. 2002; Vinces et al. 2009), which becomes nucleosome-free and thus accessible to these proteins, or by modifying the spacing between regulatory sites in the promoter (Willems et al. 1990; van Ham et al. 1993). Lastly, TRs can also affect transcription rates by affecting splicing efficiency (Hefferon et al. 2004; Hui et al. 2005).

Because TR variation is often characterized by frequent mutations which occur in small steps and can be readily reversed, these genomic elements then have the potential to efficiently and steadily introduce novel genetic variation in a quantitative manner and with minimal genetic load.

ii) Evolutionary potential

One particular striking example that clearly highlights the potential of TRs to mediate rapid phenotypic changes is found in bull terriers (Fondon et al. 2008). This breed of dogs has been artificially selected by human action over the last 150 years to have long midfaces and a snout markedly bent downwards (Figure 5).

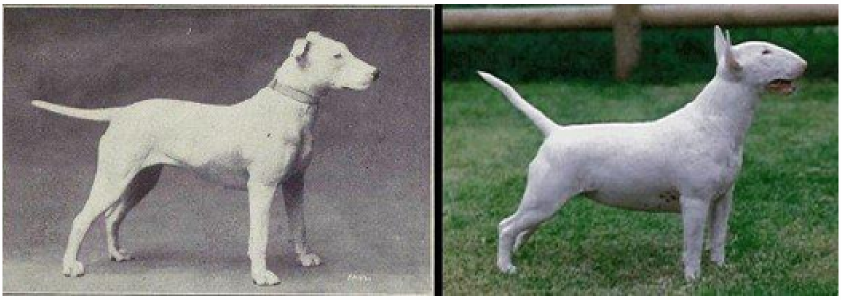

Figure 6. - The bull terrier, then (1915) and now. Due to intense selective breeding, the bull terrier has suffered dramatic skeletal morphology changes. These seem to be correlated with the ratio of two TRs located on a gene related to bone formation. (Source: https://dogbehaviorscience.files.wordpress.com/2012/09/01.jpg)

Interestingly, changes in these two craniofacial phenotypes have been found to be strongly correlated with the length ratio of two polymorphic TRs, encoding respectively for polyglutamine and polyalanines amino acid stretches, located on the coding region of *Runx-2* (runt-related transcription factor 2), a gene involved in bone formation. Since the presence of polyglutamnine and polyalanine stretches in genes have been previously found to respectively increase and repress transcription of regulated genes, changes in their length ratio in *Runx-2,* which in vertebrates encodes a transcription factor involved in osteoblast differentiation, could potentially explain the remarkably swift phenotypical evolution observed in these dogs.

These and other marked phenotypical changes can be found across several dog breeds due to the intensive domestic breeding they have been subjected over the last century. Because of the strong selection against genetic diversity imposed by an intense breeding process, the emergence of such acute changes is unlikely to be uniquely explained by DNA variation which accumulates slowly, such as SNPs, and point to TRs as the potential instigators of these changes. The enrichment of TRs in vertebrate genes related to body morphology also suggests that these repetitive elements may underlie the plasticity that characterizes vertebrates' anatomical evolution (Legendre et al. 2007).

Another very convincing example for the role of TRs as evolutionarily tuning knobs was observed in laboratory populations of yeast. Specifically, the size of TRs in the promoters of some genes of these yeast was found not only modulate gene expression level, and accelerate the transcriptional divergence between different strains and species, but also to promote evolutionary adaptation which was advantageous for the species in question (Vinces et al. 2009).

The previous example illustrates the adaptative potential of TRs, and it is not unreasonable to suggest that many other TRs may be under selection.

A particularly interesting piece of evidence that highlights TRs adaptative role, is found in the clock gene *period (per)* in *Drosophila melanogaster.* This gene has two common alleles, differentiated by the size of its TR. Their frequency show a latitudinal cline across Europe and North Africa, as well as in Australia. The geographical distribution pattern of these alleles points to differential selective advantage related to the capacity to maintain a circadian period at different temperatures, and thus implicate this TR in the evolution of circadian rhythms. In particular, the longer allele, more prevalent in colder regions, allows for a better response to temperature variation, thus minimizing its effect on the circadian cycle, while the shorter allele, which shows a circadian period of ~24 hours, is more often found in warmer temperate climate (Sawyer et al. 1997, 2006). Interestingly, differences in the circadian behaviour of rat moles have also been observed to be enacted by TR variation in the *per* homolog gene in these taxa (Ben-Shlomo et al. 1996).

Furthermore, one aspect that sets TRs apart from other types of DNA variation is its high mutability, which depends on the purity, repeat motif length, and repeat copy number of any given TR. For this reason, when selection favors a beneficial TR allele, it may also be selecting for its mutability. As a result, there is extra layer of fine-tuning acting in TRs which might drive the mutation rate of a particular TR to an optimal level. This type of selection might help explain why the TRs found in some genes display highly conserved flanking regionsv across human and nonhuman vertebrate species, but had their repeat motifs replaced during their evolutionary histories (Riley and Krieger 2009a, 2009b).

Lastly, TRs can facilitate adaptation by promoting genomic rearrangements and evolution at the level of chromosome structure. In particular, in addition to their capacity to determine recombination sites and rates, they can also induce chromosomal fragile sites. The latter can drive rapid phenotypic evolution, such as the loss of pelvic spines in stickleback fishes (Chan et al. 2010), and evidence suggests that it could also potentially explain the large-scale genomic rearrangements that have occurred in great apes evolution (Ruiz-Herrera et al. 2006).

1.1.3 Genotyping

The interest in the study of microsatellite variation, which for a long time involved performing capillary gel electrophoresis, a costly and time-consuming task, drastically subdued with the development of techniques that could genotype in parallel and in a cost-efficient manner up to thousands of SNPs. As a result, until very recently, variation databases such as dbSNP, were vastly depleted in terms of microsatellite polymorphism data. Furthermore, in part for this reason, until very recently most studies of the human and/or nonhuman great apes repeat landscape were either restrained to comparison of genome references (Webster et al. 2002; Kelkar et al. 2008; Payseur et al. 2011; Kelkar et al. 2011; Loire et al. 2013), or to small-scale genotyping efforts (Rosenberg et al. 2002; Molla et al. 2009; Pemberton et al. 2009; Tishkoff et al. 2009; Sun et al. 2012; Pemberton et al. 2013).

A significant shift in this trend coincided with the advent of high-throughput sequencing, and the development of tools that allow repeat genetic variation to be genotyped. These technological advances meant that sequence information could be efficiently retrieved at the genome-wide level, and that this data could be used to infer repeat genotypes. However, genotyping TRs remained a challenging task mainly for two reasons.

In the first place, since the content of many genomes is often highly repetitive and the sequencing reads generated are often short, when mapped, the reads may not always be unambiguously placed in the reference genome. Secondly, the construction of the sequencing library typically involves an amplification step, during which the polymerase may experience slippage and introduce stutter noise into the TRs.

Because the first tools used to genotype TRs failed to explicitly address these issues, the accurate genotyping of TRs only became possible once more repeat-aware algorithms which took these issues into consideration, such as lobSTR (Gymrek et al. 2012) and Repeatseq (Highnam et al. 2013), were developed. Specifically, Repeatseq, in addition to ignoring sequencing data from any reads which do not overlap the TR in its totality and contain some unique flanking sequence, uses a probabilistic model to assess the reliability of each genotype assigned (Figure 7).

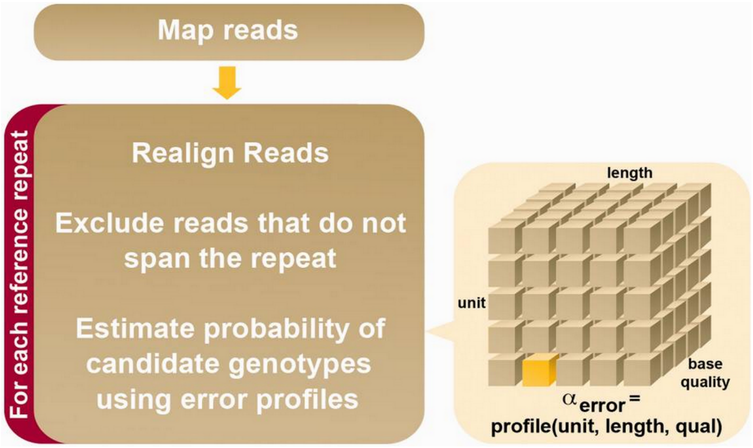
Map reads
For each reference repeat
Realign Reads
Exclude reads that do not span the repeat
Estimate probability of candidate genotypes using error profiles
length
unit
base quality
α_{error} = profile(unit, length, qual)

Figure 7. - Outline of Repeatseq's TR genotyping pipeline. After the sequencing reads are mapped to the genome reference and realigned, only those that overlap TRs in their totality are kept. This set of genome reference TRs must have been previously identified using an appropriate tool for that effect such as Tandem Repeat Finder. Repeatseq then estimates the probability of each genotype given the data using a Bayesian model. The most probable genotype in each locus is then chosen as the true one. (Highnam et al. 2013)

In essence, the algorithm is based on a Bayesian framework in which the error rates that characterize different TRs, and the quality of the sequencing reads, are explicitly accounted for when assigning a probability to the genotype at each TR locus. To produce these error profiles, several TR loci from a population of exclusively homozygous flies were genotyped, and the number of genotyping errors computed. Specifically, this data was generated as part of the Drosophila Genetics Reference Panel (DRGP), and consists of >100 inbred isolates that were generated by full-sib mating for over 20 generations, so that the individuals were homozygous in every TR locus (Figure 8) (Fondon et al. 2012; Mackay et al. 2012).

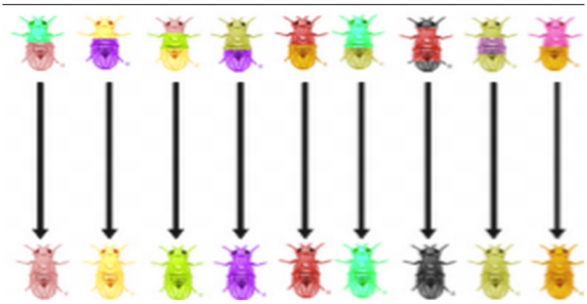

Figure 8. - In the DRGP, flies were subjected to full-sib mating for 20 generations so that by the end of the process there was only one repeat allele per TR locus (Source: http://compgen.cshl.edu/INSIGHT/figs/DGRPfig.png)

As a result, when genotyped, those TR loci called in this population as heterozygotes most likely represented errors. The number of errors were tabulated according to the repeat unit size, repeat length in the genome reference, and mean base quality of the reads, so that the uncertainty measure associated with these properties could be generated. Guided by these error profiles Repeatseq can then better assess the uncertainty associated with a given TR genotype, and thus is able to more accurately genotype TRs.

1.2 Evolution of gene expression

1.2.1 General patterns

Determining how and what mechanisms underlie the evolution of species-specific phenotypes is one of biology's oldest pursuits. In particular, the relative contribution to phenotypic evolution of changes that affect gene products at the functional and regulatory level is currently unknown (Necsulea and Kaessmann 2014). While functional changes primarily arise as a result of mutations in coding sequences which produce alterations at the RNA and protein level, regulatory changes mainly occur as a result of mutations in regions such as promoters and enhancers, and influence a wide range of processes such as transcription, translation and even degradation of gene products. Since changes at the protein-coding level are more likely to be deleterious, it is expected that regulatory mutations will not only occur more often, but also carry a signifcant weight in driving species-specific differences. This is supported, for example, by the finding that despite the many anatomical and behavioral differences that separate humans from our closest relatives, chimpanzees, very few differences exist at the protein-coding level between the two species (King and Wilson 1975). This observation hints at a more prominent role of modifications that affect gene regulation at facilitating evolutionary adaptation.

However, for the most part, evidence shows that much of the evolution of protein-coding gene expression is also constrained. In support of this argument it has been shown that unlike what would be expected under a neutral evolution scenario, gene expression divergence in amniotes has not accumulated in a linear fashion (Jordan et al. 2005). In addition, as expected, given the vital role some organs play, gene expression patterns cluster better by organ type than by species, implying the action of selective pressures at this level to conserve their function (Brawand et al. 2011).

Furthermore, the strength of evolutionary constraints is not uniform across all organs. Perhaps counter-intuitively, given the striking differences in brain complexity observed across vertebrates, neural tissues show the lowest rates of expression evolution of all tissues (Khaitovich et al. 2005a; Chan et al. 2009; Brawand et al. 2011), and brain-specific genes show low rates of protein sequence divergence (Warnefors and Kaessmann 2013; Khaitovich et al. 2005b). These observations suggest that the brain is highly fine-tuned, and that only few but precise changes are allowed to take place. On the other hand, lying at the other end of the gene expression and sequence divergence spectrum are testis (Khaitovich et al. 2005; Brawand et al. 2011). Presumably due to exceptional conditions during spermatogenesis, which cause chromatin conformation to be more open and transcription rates to occur in a less restricted manner, testis are subject to relaxed purifying selection which contributes to for the evolution expression divergence to accelerate.

Rates of expression divergence also vary across lineages. Interestingly, primates seem to evolve faster than rodents at this level (Brawand et al. 2011). Because mutation rates are much higher in the latter (Li et al. 1996), it is difficult to reconciliate this observation with the divergence trend, unless the much lower population effective size of primates compared to rodents is taken into account (Kaessmann et al. 1999a, 1999b; Yu et al. 2002; Keightley et al. 2005). Small effective population size would lead to selection being less efficient in primates, and consequently accelerate the rate of expression evolution due to an increase the number of slightly deleterious mutations which cannot be easily purged from the genome of these taxa (Keightley et al. 2005).

1.2.2 A role for tandem repeats

One the many mechanisms that may underlie gene expression divergence patterns, is TR variation in genic or nearby regions. Vinces and colleagues (Vinces et. al 2009) showed very convincingly that not only do genes with repeat-containing promoters exhibit more expression divergence across several strains and species of yeast, compared to its repeat-less counterparts, but also that this divergence is increased for highly variable TRs. This association remained even when accounting for other factors that could explain such gene expression divergence, such as the presence of TATA boxes. In addition, consistent with the proposed role for TRs as “evolutionary tuning knobs”, some of the expression changes mediated by TRs conferred higher fitness under a scenario where there was selective pressure. This example together with other studies that show how TR variation can affect gene expression levels (reviewed in Gemayel et al. 2010), and the observation that at least in *Drosophila* TRs were found to be general enhancer features (Yáñez-Cuna et al. 2014), make a plausible case for the role of TRs as drivers of gene expression divergence.

Finally, the finding that transcription factors seem to be over-represented in scans for selectively driven lineage-specific expression changes in humans (Gilad et al. 2006; Blekhman et al. 2008), argues in favor of a prominent regulatory role in human evolution. In that sense, previous observations that associate TR variation to the modulation of the binding of transcription factors, and their abundance in human promoters (Sawaya et al. 2013), supports the notion that these may have been important on the course of human evolution.

2. OBJECTIVES

1. Assay population diversity and structure in human and nonhuman great ape populations using short tandem repeats

2. Study the impact of tandem repeat presence in genic regions in primate gene expression divergence and evolution

3. Infer the effect of tandem repeat polymorphism in the promoters of human and chimpanzee individuals to gene expression divergence

3. RESULTS

Chapter 1

3.1. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence

Tiago Carvalho[+], Tugce Bilgin Sonay[+], Mark D. Robinson, Maja P. Greminger, Michael Krützen, David Comas , Gareth Highnam, David Mittelman, Andrew Sharp, Tomàs Marques-Bonet[+], Andreas Wagner[+]

+Equal contribution

Published

Bilgin Sonay T, Carvalho T, Robinson MD, Greminger M, Krutzen M, Comas D, Highnam G, Mittelman D, Sharp A, Marques Bonet T, Wagner A. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. **Genome Research.** 2015;118(4):437-43.

doi: 10.1101/gr.190868.115

4. DISCUSSION

Current sequencing technologies have opened many doors in the genomics field. It is now possible to sequence entire genomes at a reduced cost and in an efficient way. Due to this so-called genomic revolution it is now possible to explore the genomic variation of species at an unprecedented level.

The analysis of a comprehensive great ape genomic dataset, comprised of >80 human and non-human great ape genomes, has already yielded important insights into the diversity and evolution of these species through the analysis of single nucleotide polymorphisms (SNPs), mobile element insertions, and large-scale copy number variations (Prado-Martinez et al. 2013; Sudmant et al. 2013; Hormozdiari et al. 2013). However, despite the importance of tandem repeats (TRs) in population and conservation genetics, as well as their potential functional and regulatory role, until the work performed in this thesis was performed, the repeat landscape of great apes had remained largely unexplored at the genotype level.

Taking advantage of the fact that tools to accurately genotype TRs from genome-wide sequencing data are now available, a comprehensive catalog of TR variation in great apes was compiled.

As part of the work presented in this thesis, this catalog was used to find if repeat information could be used to study the genetic diversity and population structure of great apes. Reassuringly, the results obtained showed that this set of TRs could be used to reliably assess within-species population structure and define the number of distinct genetic units within a species, recapitulating those results obtained in previous studies using millions of SNPs (Li et al. 2008; Prado-Martinez et al. 2013). In particular, these markers were informative not only to distinguish between individuals from different subspecies, but even according to their geographical location of origin.

In this thesis, the contribution of TRs to gene expression divergence in primates was also examined. Following a comparison of genes with and without TRs in their promoters at the level of expression divergence in human, chimpanzee, and macaque, it was found that across several tissues the genes with TRs exhibit higher expression divergence. These observations are in line with previous studies that also looked at expression divergence patterns between different yeast strains and species, and that conclusively showed that repeat presence was associated with higher expression divergence (Vinces et al. 2009).

Notably, in the work presented in this thesis, the pattern of association of TRs presence with gene expression divergence

seemed to hold independently of the genic region considered to contain TRs.
Interestingly, those genic regions and features associated with higher expression divergence, such as 3' UTRs, first introns relative to other introns, and distance of TR from the the transcription start site (TSS), have previously been shown to be involved in gene regulation (Jonsson et al. 1992; Rohrer and Conley 1998; Wray et al. 2003; Charron et al. 2007; Spitz and Furlong 2012; Yoon et al. 2012).

Furthermore, due to the availability of the TR variation information it was also possible to classify TRs according to their polymorphism in human and chimpanzee populations, and ask if genes with polymorphic TRs showed more or less expression divergence than those where the same TR was found to be fixed in both populations. The results also show that across all analyzed tissues genes whose promoters contain polymorphic TRs exhibit higher expression divergence than those where the same TR is fixed across the two primate populations, or where no TRs are present. These findings seem to further reinforce the idea that TR variability may contribute to gene expression divergence, and suggest that TRs may indeed accelerate expression evolution, and are again concordant with the observations by Vinces and colleagues, which found that higher expression divergence was more prominent for those genes whose promoters contained highly variable repeats (Vinces et al. 2009).

Even though no study of associations can prove causation, confounding factors that could better explain these findings were analyzed. One of them was relaxed selection. Earlier work had detected that an increase in expression divergence for genes associated with species-specific transposable elements was caused by relaxed selection on those genes, rather than by the transposable elements themselves (Warnefors et al. 2010). To exclude this factor, several analyses were performed to show that repeat-containing genes were not subject to relaxed selection. These included performing comparisons in all groups analyzed regarding measures that are indicative of relaxed selection such as d_N/d_S, divergence between human and chimpanzee at the promoter level, or enrichment of TRs in recombination spots. Since no statistical significant differences were encountered between any of the groups compared, it is improbable that relaxed selection could explain the findings reported in this work.

To date this work represents the most comprehensive effort to catalog TR variation in great apes at a genome-wide scale. This information can be used to distinguish among different taxonomical groups, and assess the degree of diversity present in natural populations.

In particular, the compilation of a list of 3,521 AIMs that allow subspecies of great apes to be distinguished will be very useful for conservation and breeding programs. Importantly, since TRs have an high mutation rate, few molecular markers can be used to gather a great deal of information from any given population, a definite advantage for great apes conservation genetics which heavily relies on non-invasive samples. In addition, because TRs also exhibit a wide range of mutation rates, they can be used to capture both old and more recent shifts at the genetic diversity level in natural populations. This is particularly useful since other popular molecular markers, such as SNPs are characterized by lower mutations rates, and for this reason are may be unsuitable for the latter task.

Nonetheless, in the future, it will be important not only to add more individuals from populations already sampled, but also from those populations that could not be included in the first large initiative to characterize great ape diversity. This effort will help further expand the power of TRs to differentiate between individuals from different populations at an even higher resolution, as well as remove AIMs falsely identified as such due to limited sampling.

Given the success of this approach in assessing within-species population structure it will also be interesting to apply it to other species and populations that are not so well described.
However, since TR properties have been shown to affect their own mutability, it will also be important to consider this fact in future analyses, unlike what was done in this work.

Furthermore, future research should also focus on understanding the precise mechanisms by which TRs accelerate gene expression divergence. In this sense, it will be important not only to identify which features are influenced by TRs, apart from some of those already described such as DNA, RNA and chromatin structure and the size of transcription factor binding sites, but also to decipher the many ways in which TR variability modulates phenotypical evolution.

Acquiring a deeper insight into these mechanisms will be particularly important for the study of human evolution given the abundance of these repetitive elements in human promoters (Sawaya et al. 2013), and their association with regulatory elements, an observation suggestive of a significant role in gene regulation. In addition, given the instability that characterizes TRs, it is likely that they may have underlied substantial changes in gene expression, and thus have driven the evolution of species-specific phenotypes.

In this sense, extending the analysis of TRs to other species may help to elucidate if, for example, TRs can also partially explain why primates seem to have evolved much faster than rodents (Brawand et al. 2011) .

Addressing these questions will then help understand the role of TRs, during human and even primate evolution, in generating adaptative traits, and possibly aid in identifying new disease-associated TRs.

Probably one of major limitations of the work presented in this thesis is the fact that the length of all genotyped TRs had to be under 100 base pairs, the maximum size of the sequencing reads used for this study. This not only represents a severe limitation of the TR catalog presented here in terms of size, since a wide range of TRs cannot be genotyped, but also in accuracy terms, since shorter reads are more likely to be ambiguously mapped to the reference genome. In addition, it is possible that longer TRs show more interesting mutational properties which might confer them particularly special adaptative value.

In this sense, future studies will greatly benefit not only of the use of methods that can simultaneously and in a more accurate fashion target and genotype many TRs (Guilmatre et al. 2013; Duitama et al. 2014; Carlson et al. 2015), but also from using a sequencing technology that produces longer sequencing reads.

Notwithstanding, repeat-aware software such as Repeatseq, which make use of empirical data to better assess the accuracy of each repeat genotype, and the availability of abundant sequencing data, already permit repeat variation to be probed, and as shown in the work present in this thesis, for meaningful information about natural populations to be produced.

Finally, since the correlation between gene and protein expression levels is often far from perfect (Battle et al. 2015; Bauernfeind et al. 2015), future efforts that study the functional impact of TRs will be enhanced if protein expression levels are also assessed. Otherwise cases such as the one in longer TR alleles in the 5'UTR of the thymidylate synthase human gene are associated with higher protein levels, but not gene expression changes, will be missed (Kawakami et al. 2001). In addition, since in the case of protein-coding genes what ultimately determines the phenotype are the translated proteins, assessing their levels will be important to understand the true impact of TRs in phenotypical evolution.

CONCLUDING REMARKS

In this thesis I explored for the first time the repeat landscape of TR variation in recent human evolution. This information proved very useful to study the population structure and diversity of humans and nonhuman great apes, and is expected to be a valuable resource for conservationists interested in improving the management of these great ape species both in the wild and in captivity.

In addition, much of work presented here attempts to examine the effect of TRs on primate gene expression divergence across several tissues. The results suggest that genes with TRs show more expression divergence than those without, independently of the presence in the promoter or any other genic region, and that there is an association between polymorphic TRs in promoters and greater expression divergence of those genes. These findings are supported by previous work which has similarly implicated TRs in promoters of yeast genes to be involved in accelerated divergence of gene expression between different at the strain and even species level (Vinces et al. 2009).

Given the potential of such mechanism to facilitate evolutionary adaptation, it will be interesting to explore how it might have impacted the evolutionary history of humans and great apes, and if it underlies particularly interesting phenotypical innovations in any of the great apes lineages.

Finally, given the abundance of traits and diseases which common variation such as SNPs cannot appropriately explain at the present, and TRs functional potential, it will be worthwhile for future GWAS studies to try to genotype these repetitive elements to see if they can account for the so-called “missing heritability”.

List of communications

Articles published in peer-reviewed journals

Carvalho T*, Bilgin Sonay T*, Robinson MD, Krutzen M, Lorente-Galdos B, Comas D, Highnam D, Sharp A, Marques Bonet T, Wagner A.. *Tandem repeat variation in human and great ape populations and its impact on gene expression divergence.* **Genome Research**. 2015;118(4):437-43 (featured on the Genome Research cover)
* Equal contribution

Prado-Martinez J., Sudmant PH., Kidd JM., Li, H, Kelley JL., Lorente-Galdos B, Veeramah K,
Woerner A,O'Connor TD., Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE., Malig M, Hernandez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C,Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML., Stevison L,Camprubí C, **Carvalho T**, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE., Pusey A.,Lankester F., Kiyang JA., Bergl RA., Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D,

Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA., Mullikin JC., Wilson RK., Gut IG., Gonder MK., Ryder OA., Hahn BH., Navarro A., Akey JM., Bertranpetit J, Reich D, Mailund T, Schierup MH., Hvilsom C., Andrés AM., Wall J, Bustamante CD., Hammer M, Eichler EE., Marques-Bonet T. *Great ape genetic diversity and population history*. **Nature**. 2013 Jul 25;499(7459):471-5. doi:10.1038/nature12228.

In preparation

André M. M. Sousa, Ying Zhu, Mary Ann Raghanti, Andrew T. Tebbenkamp, Robert R. Kitchen, Kyle A. Meyer, Mingfeng Li, Yuka Imamura Kawasawa, Marta Mele, Raquel Garcia Perez, **Tiago Carvalho**, Andrew T. N. Tebbenkamp, Mihovil Pletikos, John J. Ely, Patrick R. Hof, Mark Reimers, Richard P. Lifton, Shrikant M. Mane, James P. Noonan, Matthew W. State, Ed S. Lein, James A. Knowles, Tomas Marques-Bonet, Chet C. Sherwood, Mark B. Gerstein, Nenad Sestan *Evolutionary Adaptations of Cerebral Dopaminergic Neurons in the Human Lineage*

Belen Lorente-Galdos, Arturo Silveyra, Gerard Serra, Himla Soodyall, Pierre Zalloua, Karima Fadhlaoui-Zid, **Tiago Carvalho**, Javier Prado-Martinez, Marcos Fernandez, David Reich, Tomas

Marques-Bonet, David Comas. *Demographic inferences from a diverse panel of African human genomes*

Katja Nowick, Rui Faria, Jan Aerts, Walter Salzberg, Mehmet Sohmel, Giovanni Dall'olio, Deborah Triant, Sofia Robb, Lydia Muller, **Tiago Carvalho**, Rik Verdonck, Vladimir Jovanovic, Jan Engelhardt, Rohit Kolora, Alvaro Sabogal, Henrike Indrischek, Sonja Grath, Katja Liebal, Christoph Bleidorn, Bert Overduin, Mario Fasold, Johannes Engelken. *Ten Simple Rules for running a >2 weeks long introductory bioinformatics training course*

Bibliography

Arora N, Nater A, van Schaik CP, Willems EP, van Noordwijk MA, Goossens B, Morf N, Bastian M, Knott C, Morrogh-Bernard H, et al. 2010. Effects of Pleistocene glaciations and rivers on the population structure of Bornean orangutans (Pongo pygmaeus). *Proc Natl Acad Sci U S A* **107**: 21376–21381.

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. 2015. Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**: 664–7.

Bauernfeind AL, Soderblom EJ, Turner ME, Moseley MA, Ely JJ, Hof PR, Sherwood CC, Wray GA, Babbitt CC. 2015. Evolutionary Divergence of Gene and Protein Expression in the Brains of Humans and Chimpanzees. *Genome Biol Evol* **7**: 2276–2288.

Becquet C, Patterson N, Stone AC, Przeworski M, Reich D. 2007. Genetic structure of chimpanzee populations. *PLoS Genet* **3**: e66.

Ben-Shlomo R, Ritte U, Nevo E. 1996. Circadian rhythm and the per ACNGGN repeat in the mole rat, Spalax ehrenbergi. *Behav Genet* **26**: 177–184.

Bergl RA, Vigilant L. 2007. Genetic analysis reveals population structure and recent migration within the highly fragmented range of the Cross River gorilla (Gorilla gorilla diehli). *Mol Ecol* **16**: 501–16.

Bjork A, Liu W, Wertheim JO, Hahn BH, Worobey M. 2011. Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. *Mol Biol Evol* **28**: 615–23.

Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. 2008. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet* **4**.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–8.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* **62**: 1408–15.

Carlson KD, Sudmant PH, Press MO, Eichler EE, Shendure J, Queitsch C. 2015. MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals. *Genome Res* **25**: 750–61.

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJH, Wilde A, Brudno M, et al. 2009. Conservation of core gene expression in vertebrate tissues. *J Biol* **8**: 33.

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**: 302–305.

Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR, Verstrepen KJ, Froyen G. 2014. Large-scale analysis of tandem repeat variability in the human genome.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**: 435–45.

Elmore MH, Gibbons JG, Rokas A. 2012. Assessing the genome-wide effect of promoter region tandem repeat natural variation on gene expression. *G3* **2**: 1643–9.

Ferris SD, Brown WM, Davidson WS, Wilson a C. 1981. Extensive polymorphism in the mitochondrial DNA of apes. *Proc Natl Acad Sci U S A* **78**: 6319–23.

Fischer A, Prüfer K, Good JM, Halbwax M, Wiebe V, André C, Atencia R, Mugisha L, Ptak SE, Pääbo S. 2011. Bonobos fall within the genomic variation of chimpanzees. *PLoS One* **6**: e21605.

Fondon JW, Hammock E a D, Hannan AJ, King DG. 2008. Simple sequence repeats: genetic modulators of brain function and behavior. *Trends Neurosci* **31**: 328–34.

Fondon JW, Martin A, Richards S, Gibbs R a, Mittelman D. 2012. Analysis of microsatellite variation in Drosophila melanogaster with population-scale genome sequencing. *PLoS One* **7**: e33036.

Fünfstück T, Arandjelovic M, Morgan DB, Sanz C, Breuer T, Stokes EJ, Reed P, Olson SH, Cameron K, Ondzie A, et al. 2014. The genetic population structure of wild western lowland gorillas (Gorilla gorilla gorilla) living in continuous rain forest. *Am J Primatol* **76**: 868–78.

Galvão R, Mendes-Soares L, Câmara J, Jaco I, Carmo-Fonseca M. 2001. Triplet repeats, RNA secondary structure and toxic gain-of-function models for pathogenesis. *Brain Res Bull* **56**: 191–201.

Garner KJ, Ryder O a. 1996. Mitochondrial DNA diversity in gorillas. *Mol Phylogenet Evol* **6**: 39–48.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* **44**: 445–77.

Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440**: 242–245.

Godde JS, Wolffe AP. 1996. Nucleosome assembly on CTG triplet repeats. *J Biol Chem* **271**: 15222–15229.

Goossens B, Chikhi L, Ancrenaz M, Lackman-Ancrenaz I, Andau P, Bruford MW. 2006. Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biol* **4**: e25.

Greminger MP, Stölting KN, Nater A, Goossens B, Arora N, Bruggmann R, Patrignani A, Nussberger B, Sharma R, Kraus RHS, et al. 2014. Generation of SNP datasets for orangutan population genomics using improved reduced-representation sequencing and direct comparisons of SNP calling algorithms. *BMC Genomics* **15**: 16.

Guilmatre A, Highnam G, Borel C, Mittelman D, Sharp AJ. 2013. Rapid multiplexed genotyping of simple tandem repeats using capture and high-throughput sequencing. *Hum Mutat* **34**: 1304–11.

Gymrek M, Golan D, Rosset S, Erlich Y. 2012. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* **22**: 1154–62.

Hammock E a D, Young LJ. 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* **308**: 1630–4.

Hefferon TW, Groman JD, Yurk CE, Cutting GR. 2004. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proc Natl Acad Sci U S A* **101**: 3504–3509.

Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* **41**: e32.

Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker JA, Nelson B, Alkan C, Sudmant PH, Huddleston J, et al. 2013. Rates and patterns of great ape retrotransposition. *Proc Natl Acad Sci U S A* **110**: 13457–62.

Hui J, Hung L-H, Heiner M, Schreiner S, Neumüller N, Reither G, Haas S a, Bindereif A. 2005. Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J* **24**: 1988–98.

Hvilsom C, Carlsen F, Heller R, Jaffré N, Siegismund HR. 2014. Contrasting demographic histories of the neighboring bonobo and chimpanzee. *Primates* **55**: 101–112.

Hvilsom C, Frandsen P, Børsting C, Carlsen F, Sallé B, Simonsen BT, Siegismund HR. 2013. Understanding geographic origins and history of admixture among chimpanzees in European zoos, with implications for future breeding programmes. *Heredity (Edinb)* **110**: 586–93.

IUCN. 2015. IUCN Red List of Threatened Species. *Version 20153* www.iucnredlist.org. www.iucnredlist.org.

Jordan IK, Mariño-Ramírez L, Koonin E V. 2005. Evolutionary significance of gene expression divergence. *Gene* **345**: 119–26.

Kaessmann H, Heissig F, von Haeseler A, Pääbo S. 1999a. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat Genet* **22**: 78–81.

Kaessmann H, Wiebe V, Pääbo S. 1999b. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286**: 1159–1162.

Kawakami K, Salonga D, Park JM, Danenberg KD, Uetake H, Brabender J, Omura K, Watanabe G, Danenberg P V. 2001. Different lengths of a polymorphic repeat sequence in the thymidylate synthase gene affect translational efficiency but not its gene expression. *Clin Cancer Res* **7**: 4096–4101.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol* **3**: 0282–0288.

Kelkar YD, Eckert KA, Chiaromonte F, Makova KD. 2011. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res* **21**: 2038–48.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res* **18**: 30–8.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Pääbo S. 2005a. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**: 1850–1854.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Pääbo S. 2005b. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**: 1850–1854.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–16.

Kita E, Katsui N, Emoto M, Sawaki M, Oku D, Nishikawa F, Hamuro A, Kashiba S. 1991. Virulence of transparent and opaque colony types of Neisseria gonorrhoeae for the genital tract of mice. *J Med Microbiol* **34**: 355–362.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Legendre M, Pochet N, Pak T, Verstrepen KJ. 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res* **17**: 1787–96.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100–1104.

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol* **5**: 182–7.

Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol* **11**: 2453–65.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth L V, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**: 529–33.

Loire E, Higuet D, Netter P, Achaz G. 2013. Evolution of coding microsatellites in primate genomes. *Genome Biol Evol* **5**: 283–95.

Mackay TFC, Richards S, Stone E a, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**: 173–8.

Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER. 2005. Microsatellite instability regulates transcription factor binding and gene expression. **102**: 3800–3804.

McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, Ryder OA, Ape Genome Project G, Kidd JM, Wall JD, et al. 2015. Inference of gorilla demographic and selective history from whole-genome sequence data. *Mol Biol Evol* **32**: 600–12.

Molla M, Delcher A, Sunyaev S, Cantor C, Kasif S. 2009. Triplet repeat length bias and variation in the human transcriptome. *Proc Natl Acad Sci U S A* **106**: 17095–100.

Nater A, Arora N, Greminger MP, Van Schaik CP, Singleton I, Wich SA, Fredriksson G, Perwitasari-Farajallah D, Pamungkas J, Krützen M. 2013. Marked population structure and recent migration in the critically endangered sumatran orangutan (Pongo abelii). *J Hered* **104**: 2–13.

Naylor LH, Clark EM. 1990. d(TG)n.d(CA)n sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. *Nucleic Acids Res* **18**: 1595–1601.

Necsulea A, Kaessmann H. 2014. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat Rev Genet* **15**: 734–748.

Payseur B a, Jing P, Haasl RJ. 2011. A genomic portrait of human microsatellite variation. *Mol Biol Evol* **28**: 303–12.

Pearson CE, Nichol Edamura K, Cleary JD. 2005. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* **6**: 729–42.

Pemberton TJ, DeGiorgio M, Rosenberg NA. 2013. Population structure in a comprehensive genomic data set on human microsatellite variation. *G3 (Bethesda)* **3**: 891–907.

Pemberton TJ, Sandefur CI, Jakobsson M, Rosenberg N a. 2009. Sequence determinants of human microsatellite variability. *BMC Genomics* **10**: 612.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* **499**: 471–5.

Reinartz GE, Karron JD, Phillips RB, Weber JL. 2000. Patterns of microsatellite polymorphism in the range-restricted bonobo (Pan paniscus): considerations for interspecific comparison with chimpanzees (P. troglodytes). *Mol Ecol* **9**: 315–328.

Riley DE, Krieger JN. 2009a. Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences. *Gene* **429**: 74–79.

Riley DE, Krieger JN. 2009b. UTR dinucleotide simple sequence repeat evolution exhibits recurring patterns including regulatory sequence motif replacements. *Gene* **429**: 80–86.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic Structure of Human Populations. **298**: 2381–2385.

Rothenburg S, Koch-Nolte F, Rich A, Haag F. 2001. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc Natl Acad Sci U S A* **98**: 8985–90.

Roy J, Gray M, Stoinski T, Robbins MM, Vigilant L. 2014. Fine-scale genetic structure analyses suggest further male than female dispersal in mountain gorillas. *BMC Ecol* **14**: 21.

Ruiz-Herrera A, Castresana J, Robinson TJ. 2006. Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol* **7**: R115.

Sandman K, Reeve JN. 1999. Archaeal nucleosome positioning by CTG repeats. *J Bacteriol* **181**: 1035–8.

Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. 2013. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One* **8**: e54710.

Sawyer LA, Hennessy JM, Peixoto AA, Rosato E, Parkinson H, Costa R, Kyriacou CP. 1997. Natural variation in a Drosophila clock gene and temperature compensation. *Science* **278**: 2117–2120.

Sawyer LA, Sandrelli F, Pasetto C, Peixoto AA, Rosato E, Costa R, Kyriacou CP. 2006. The period gene Thr-Gly polymorphism in Australian and African Drosophila melanogaster populations: Implications for selection. *Genetics* **174**: 465–480.

Scally A, Yngvadottir B, Xue Y, Ayub Q, Durbin R, Tyler-Smith C. 2013. A genome-wide survey of genetic variation in gorillas using reduced representation sequencing. *PLoS One* **8**: e65066.

Schlötterer C. 2004. The evolution of molecular markers--just a matter of fashion? *Nat Rev Genet* **5**: 63–69.

Stone AC, Battistuzzi FU, Kubatko LS, Perry GH, Trudeau E, Lin H, Kumar S. 2010. More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. *Philos Trans R Soc Lond B Biol Sci* **365**: 3277–3288.

Sturtevant AH. 1913. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *J Exp Zool* **14**: 43–59.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* **23**: 1373–82.

Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, et al. 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet* **44**: 1161–5.

Thomson N, Sebaihia M, Cerdeño-Tárraga A, Bentley S, Crossman L, Parkhill J. 2003. The value of comparison. *Nat Rev Microbiol* **1**: 11–12.

Tian B, White RJ, Xia T, Welle S, Turner DH, Mathews MB, Thornton C a. 2000. Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA* **6**: 79–87.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, et al. 2009. The genetic structure and history of Africans and African Americans. *Science* **324**: 1035–44.

Tomita N, Fujita R, Kurihara D, Shindo H, Wells RD, Shimizu M. 2002. Effects of triplet repeat sequences on nucleosome positioning and gene expression in yeast minichromosomes. *Nucleic Acids Res Suppl* 231–2.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*.

Usdin K. 2008. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res* **18**: 1011–9.

Vallender EJ. 2011. Expanding whole exome resequencing into non-human primates. *Genome Biol* **12**: R87.

Van Ham SM, van Alphen L, Mooi FR, van Putten JPM. 1993. Phase variation of H. influenzae fimbriae: Transcriptional control of two divergent genes through a variable combined promoter region. *Cell* **73**: 1187–1196.

Vigilant L, Bradley BJ. 2004. Genetic variation in gorillas. *Am J Primatol* **64**: 161–172.

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**: 1213–6.

Warnefors M, Kaessmann H. 2013. Evolution of the correlation between expression divergence and protein divergence in mammals. *Genome Biol Evol* **5**: 1324–35.

Warnefors M, Pereira V, Eyre-Walker A. 2010. Transposable elements: insertion pattern and impact on gene expression evolution in hominids. *Mol Biol Evol* **27**: 1955–62.

Warren KS, Nijmian IJ, Lenstra JA, Swan RA, Heriyanto, den Boer M. 2000. Microsatellite DNA variation in Bornean orangutans (Pongo pygmaeus). *J Med Primatol* **29**: 57–62.

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Hum Mol Genet* **2**: 1123–8.

Webster MT, Smith NGC, Ellegren H. 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc Natl Acad Sci U S A* **99**: 8748–53.

Wegmann D, Excoffier L. 2010. Bayesian inference of the demographic history of chimpanzees. *Mol Biol Evol* **27**: 1425–35.

Willems R, Paul a, van der Heide HG, ter Avest a R, Mooi FR. 1990. Fimbrial phase variation in Bordetella pertussis: a novel mechanism for transcriptional regulation. *EMBO J* **9**: 2803–2809.

Yáñez-Cuna JO, Arnold CD, Stampfel G, Boryń ŁM, Gerlach D, Rath M, Stark A. 2014. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res* **24**: 1147–1156.

Yu N, Fu Y-X, Li W-H. 2002. DNA polymorphism in a worldwide sample of human X chromosomes. *Mol Biol Evol* **19**: 2131–2141.

Zhang Y, Morin PA, Ryder OA, Zhang Y. 2001. A set of human tri- and tetra-nucleotide microsatellite loci useful for population analyses in gorillas (Gorilla gorilla gorilla) and orangutans (Pongo pygmaeus). *Conserv Genet* **9700**: 391–395.

Zsurka G, Kudina T, Peeva V, Hallmann K, Elger CE, Khrapko K, Kunz WS. 2010. Distinct patterns of mitochondrial genome diversity in bonobos (Pan paniscus) and humans. *BMC Evol Biol* **10**: 270.

Appendix

The following paper analyzes the genome diversity of great apes through the analysis of millions of SNPs identified from sequencing 79 individuals from six great ape species.

I collaborated in this project by validating the SNPs identified with the sequencing data, by checking if these were concordant with those genotyped with microarrays.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. Great Ape genetic diversity and population history. Nature. 2013 Jul 25;499(7459) :471-5. doi: 10.1038/nature12228.