University *of* Leicester

Gene conversion on the human Y chromosome

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

By

Georgina R Bowden

Department of Genetics

University of Leicester

October 2010

Acknowledgements

I would like to thank all the people whose help and advice over the past 4 years has helped contribute to this thesis. I would particularly like to thank my supervisor Professor Mark Jobling for all his valuable help and advice and giving me the opportunity to carry out this research. I would also like to thank the members of the Y chromosome group for all their support and creating an enjoyable environment to work in.

I would also like to thank the members of my thesis committee, Dr Richard Badge and Dr Celia May for their support and advice while carrying out this research.

Finally I would like to thank my family for all their support over that past 4 four years

Statement

Work on this thesis was carried out part-time between October 2006 and October 2010 while I was working full-time as a Research Technician at the University of Leicester. Writing up continued while I was working as a Research Assistant at the Institute of Cancer Research (ICR) in Surrey.

My work at the University of Leicester between January 2005 and August 2009 has led to one first-author publication and 4 publications on which I am co-author.

A predominantly Neolithic origin for European paternal lineages.
Balaresque P, Bowden GR, Adams SM, Leung HY, King TE, Rosser ZH, Goodwin J, Moisan JP, Richard C, Millward A, Demaine AG, Barbujani G, Previderè C, Wilson IJ, Tyler-Smith C, Jobling MA. *PLoS Biol*. (2010) 8, e1000285. doi:10.1371/journal.pbio.1000285.

Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis.
Balaresque P, Bowden GR, Parkin EJ, Omran GA, Heyer E, Quintana-Murci L, Roewer L, Stoneking M, Nasidze I, Carvalho-Silva DR, Tyler-Smith C, de Knijff P, Jobling MA. *Hum. Mutat*. (2008) 29: 1171-80

Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England.
Bowden GR, Balaresque P, King TE, Hansen Z, Lee AC, Pergl-Wilson G, Hurley E, Roberts SJ, Waite P, Jesch J, Jones AL, Thomas MG, Harding SE, Jobling MA. *Mol. Biol. Evol*. (2008) 25: 301-9.

Thomas Jefferson's Y chromosome belongs to a rare European lineage.
King TE, Bowden GR, Balaresque PL, Adams SM, Shanks ME, Jobling MA.
Am. J. Phys. Anthropol. (2007) 132: 584-9.

Structural variation on the short arm of the human Y chromosome: recurrent multigene deletions encompassing Amelogenin Y.
Jobling MA, Lo IC, Turner DJ, Bowden GR, Lee AC, Xue Y, Carvalho-Silva D, Hurles ME, Adams SM, Chang YM, Kraaijenbrink T, Henke J, Guanti G, McKeown B, van Oorschot RA, Mitchell RJ, de Knijff P, Tyler-Smith C, Parkin EJ. *Hum. Mol. Genet*. (2007) 16: 307-16.

Table of contents

Definitions

Paralog- Highly similar non-allelic sequences resulting from a duplication event

Paralogous sequence variant (PSV)- Sequence difference between two paralogous sequences

Gametolog – Highly similar non-allelic sequences located on the X and Y chromosomes due to their shared origin, or more recent transposition

Gametologous sequence variant (GSV) – sequence differences identifed between gametologous sequences

Palindrome – Class of Y chromosome paralog in which copies are inverted, and display >99.9% sequence similarity between the duplicated sequences, referred to as ‘arms’

Proximal palindrome arm – The arm of a palindrome situated closest to the centromere

Distal palindrome arm – The arm of a palindrome situated furthest away from the centromere

Inverted repeat – Class of Y chromosome paralog in which copies are inverted, and display 98-99% sequence similarity between the duplicated sequences; copies are more distant from each other, both physically and in divergence, than are the arms of a palindrome

Y phylogeny – phylogenetic tree which enables Y chromosomes to be classified based on the allelic states of binary markers

Phylogenetic analysis – analysis carried out using DNAs from chromosomes representing diverse haplogroups

Haplogroup - Class of Y chromosome determined from typing binary markers

Microsatellite – A tandem repeat of 2-6 bases which may occur in tandem up to 30 times

Abbreviations

µg	microgram
µl	microlitre
A	adenine
ABI	Applied Biosystems
AZF	*azoospermia factor*
BGCgc	biased gene conversion to C or G allele
bp	base pair
BSA	bovine serum albumin
C	cytosine
C	degree Celsius
CV	Craig Venter sequence
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
EDTA	ethylenediamine tetra acetic acid
G	guanine
g	gram
GSV	Gametelogous sequence variant
Hg	haplogroup
HJ	Holliday Junction
Indel	DNA insertion or deletion
IR	Inverted repeat
IR1Yp	paralog of IR1 which is located on Yp
IR1Yq	paralog of IR1 which is located on Yq
JW	James Watson sequence
Kb	kilobase

l	litre
M	molar
Mb	megabase
MDS	multidimensional scaling
ml	millilitre
MYA	million years ago
NAHR	Non-allelic homologous recombination
ng	nanogram
NRY	non-recombining region of the Y chromosome
P	Palindrome
PAR1	pseudoautosomal region 1
PAR2	pseudoautosomal region 2
PCR	polymerase chain reaction
pH	potential of hydrogen
PSV	Paralogous sequence variant
SAP	shrimp alkaline phosphatase
SNP	single nucleotide polymorphism
T	thymine
TBE	tris-borate EDTA
TE	tris-EDT
TMRCA	Time to the most recent common ancestor
Tris	2-amino-2 (hydroxymethyl)-1.3-propandiol
U	unit
V	volt
w/v	weight per volume
WGA	Whole genome amplification
Yh	YanHuang sequence

Abstract

For many years it was believed that recombination on the human Y chromosome was restricted to the XY-homologous pseudoautosomal regions, with over 95% of the Y chromosome believed to be non-recombining. Over the past 7 years gene conversion has been shown to occur between several classes of paralog situated outside of the pseudoautosomal regions. Gene conversion has been shown to occur both intrachromosomally on the Y chromosome, and between the X and the Y chromosomes (Cruciani et al. 2010; Rosser et al. 2009; Rozen et al. 2003; Trombetta et al. 2009) and several biases in the direction of gene conversion have been suggested (Bosch et al. 2004; Rozen et al. 2003; Trombetta et al. 2009). This study has used interspecies sequence comparisons to identify regions of the Y chromosome which are likely to be undergoing gene conversion. Phylogenetic analysis of paralogous sequence variants (PSVs) or gametologous sequence variants (GSVs) identified between these regions has been carried out. Significantly lower interspecies divergence was observed between orthologous palindrome arms in comparison to the non-duplicated spacers (P=0.0001, 2-tailed Fisher exact test) suggesting that conservative gene conversion occurs between the arms of palindromes. Significant evidence (P=0.0001, chi square test) of conservative gene conversion was observed between the arms of P6 with genotyping of 10 PSVs identifying 62 conversion events, of which 52 convert to the ancestral allele and 10 to the derived allele. Evidence of gene conversion was observed between the *VCX/VCY* and *TGIF2LX/Y* genes and between IR1 and P1. This study suggests that gene conversion between Y chromosome paralogs is conservative of the ancestral sequence via an unknown mechanism and that conservative gene conversion is not limited to genic regions of the Y chromosome. It also demonstrates that gene conversion can occur between multiple Yq paralogs and multiple XY-homologous genes.

Summary

Human DNA is divided into 46 ‘packages’, called chromosomes. Most chromosomes are the same between men and women, but one pair differs. Women have two X chromosomes, while men have one X and one Y chromosome, the Y determining male sex early in development. Apart from the Y, all chromosomes undergo a process of DNA exchange when eggs are sperm are made, called recombination. The isolation of the Y from recombination has led to the idea that it is gradually degenerating, and may, in the distant future, disappear. However, over the past 7 years evidence of recombination on the Y has emerged. In particular, a process called gene conversion - the one-way transfer of segments of DNA, has been observed between repeated DNA sequences (paralogs) within the Y chromosome, and between parts of the Y and parts of the X, in regions where they were thought to be isolated. This study aimed to examine this process in detail. First, comparisons between the human and chimpanzee Y-chromosome DNA sequences were used to identify regions of the Y chromosome which are likely to be undergoing gene conversion, and subsequently analysis of patterns of DNA sequence variation between these regions in different men was carried out. Sequence analysis indicated (as has been suggested before) that gene conversion among Y-chromosomal segments has been acting to conserve the DNA sequence against evolutionary change, and furthermore that this process has been operating during the evolution of humans over the last few tens of thousands of years. This process may in effect protect the Y chromosome from degeneration. Limited evidence of gene conversion was also observed between genes which have one copy on the X chromosome and one on the Y, and this supports the idea that change on the Y could be limited or modulated by the transfer of X-chromosome DNA segments.

Chapter 1: Introduction

1.1, General introduction

Gene conversion - the non-reciprocal transfer of sequence information is a fundamental evolutionary process (Hurles 2004) which occurs in diverse organisms. Gene conversion has been most intensely studied in fungi as all four products of a single meiosis may be observed, allowing easy identification of conversion events. In humans the study of allelic gene conversion is more complicated as only one product of a single meiosis can be observed, so that double-crossovers cannot be distinguished from gene conversions. Non-allelic gene conversion occurring between Y chromosome paralogs is uncomplicated by recombination and easier to identify as it is not necessary to observe all products of a single meiosis for conversion events to be identified. Despite extensive research being carried out into gene conversion the mechanism remains poorly understood, particularly in humans. Current research carried out in *Drosophila* and yeast has begun to reveal some of the mechanisms and proteins involved in the gene conversion process.

This thesis describes a population-based investigation of non-allelic gene conversion occurring between paralogous DNA sequences located on the human Y chromosome.

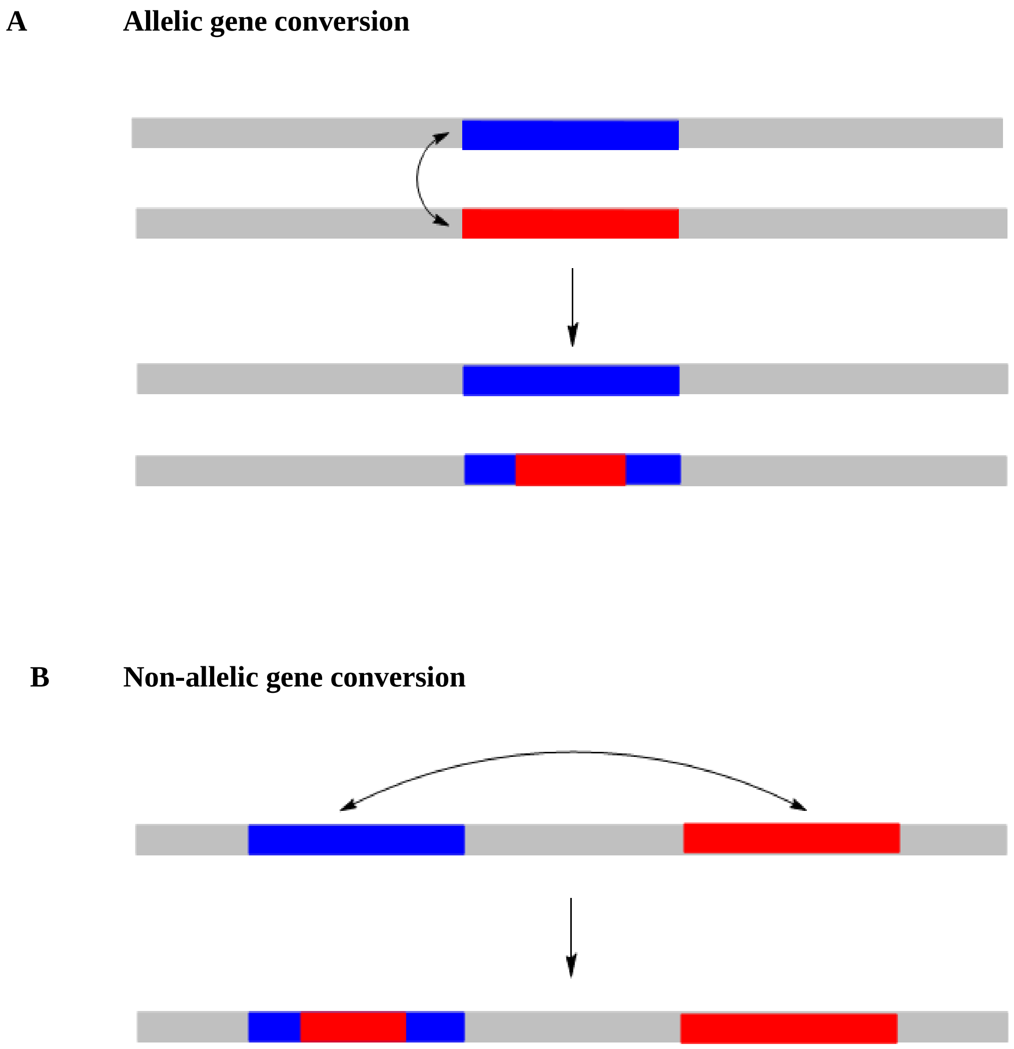
1.2, Aims of this chapter

This chapter will discuss what is currently known about gene conversion, and the properties of the Y chromosome which make it a useful tool to study this process. It will also review the available literature published about gene conversion occurring between Y chromosome paralogs, and how the Y chromosome phylogeny can be used to identify historical conversion events, and to estimate the rate of gene conversion

1.3, Gene conversion

Gene conversion is defined as the non-reciprocal transfer of genetic material between two homologous sequences with one sequence acting as a “sequence donor” and remaining unchanged, and a second sequence acting as a “sequence acceptor” and undergoing gene conversion. Gene conversion most commonly occurs between sequences which display >95% similarity although it has been observed when similarity is as low as 92% (Chen et al. 2007). Despite its name gene conversion is not limited to genes and has the potential to occur between any duplicated sequences in the genome, these duplicated sequences may be situated on homologous chromosomes, on sister chromatids, or paralogs located on the same chromosome (Chen et al. 2007).

Two types of gene conversion are known to occur in humans; these are allelic and non-allelic gene conversion. Allelic gene conversion (Figure 1.1a) is thought to be the most common (Chen and Li 2001), occurring between homologs located on separate chromosomes during recombination. Non-allelic gene conversion is less common, occurring between paralogs which may be situated within the same chromosome or on different chromosomes (Figure 1.1b). Allelic gene conversion events are more difficult to identify than non-allelic conversion events. During meiosis of diploid chromosomes recombination occurs along the entire chromosome length, and this along with normal mutational processes introduces the complication of inter-allelic diversity into the study of gene conversion. When carrying out population studies of allelic gene conversion there is the additional complication of the inability to observe more than one product of meiosis which makes it difficult to distinguish non-reciprocal exchange such as gene conversion from reciprocal exchange such as double crossover. In humans non-allelic gene conversion events are more easily



A Allelic gene conversion

B Non-allelic gene conversion

Figure 1.1: Allelic and non-allelic gene conversion

a) Allelic gene conversion occurs between duplicated sequences which may be located on homologous chromosomes or on sister chromatids.

b) Non-allelic gene conversion occurs between paralogs which in this example are located on the same chromosome, but can also lie on different chromosomes identified than allelic conversion events especially on the Y chromosome. As the Y chromosome is haploid it does not undergo recombination other than at the PARs where obligatory cross over with the X chromosome occurs during meiosis. The properties of the Y chromosome which make it a useful tool to study gene conversion will be discussed in more detail in section 1.6 of this chapter.

Non-allelic gene conversion has been observed to have four main effects on the human genome

1) Increasing sequence identity

Gene conversion is known to increase sequence identity between paralogs in a process known as concerted evolution. This is a phenomenon whereby paralogs within a single species appear to be more closely related than either is to their orthologous sequences in a closely related species (Hurles and Jobling 2001). Concerted evolution has been noted in multi-gene families that have arisen through gene duplication, leading to the homogenization of variants. This homogenization increases lengths of identity between duplicated sequences and renders paralogs better substrates for non-allelic homologous recombination (NAHR) (Hurles 2004). These observations have led to the suggestion that over an evolutionary time scale, gene conversion events may result in the formation of species-specific rearrangement hotspots (Hurles 2004) which may result in some individuals displaying an increased rate of chromosomal rearrangements in the germline.

2) Effects on the study of evolutionary history

Since the divergence between duplicated genes can be correlated to the time since duplication took place (Murphy et al. 2006) divergence calculations are often used to estimate when a duplication event first occurred. As gene conversion acts to increase sequence identity between parlogs the homogenizing effect can result in under-estimation of divergence time effectively eliminating evidence of evolutionary history. For example, high sequence identity between members of the *DAZ* multi-gene family led to the assumption that gene amplification had occurred within the last 200,000 years (Causio et al. 2000); however, multiple Y-linked *DAZ* copies also exist in apes and old world monkeys, indicating that amplification must have occurred before the human and chimpanzee lineages diverged 5-6 million years ago. The reason for their high intra-species similarity is due to gene conversion rather than recent duplication.

3) Increasing sequence diversity

As well as increasing sequence identity between paralogs gene conversion can also increase diversity between the orthologs of two closely related species. Gene conversion may also have the effect of increasing sequence diversity within a species above the normal rate as the result of different conversion events occurring in different chromosomes, a process recognized in organisms as diverse as humans, flies, and protozoans (Nielsen et al. 1983).

4) Introduction of pathological mutations

As well as eliminating mutations gene conversion has been shown to be involved in the occurrence of some pathological mutations in humans. Evidence for gene conversion between paralogs has been detected at a number of loci known to be involved in sponsoring pathogenic chromosomal rearrangements (Hurles 2004). Gene conversion of a functional gene by an inactive pseudogene can result in the production of inactivating mutations and is known to result in some human diseases such as adrenal hyperplasia and polycystic kidney disease (Chen et al. 2007). In duplicated genes, conversion has been shown to be a source for spreading disease mutations (Bailey et al. 2001). For example, allelic gene conversion on the X chromosome has been shown to maintain the identity of two duplicated sequences, nt1h-1 and int22h-1 which predispose to inversions breaking the F8 gene and causing hemophilia A (Bagnall et al. 2005).

1.4, Mechanisms and properties of gene conversion

As previously discussed, gene conversion has a number of different effects on the human genome but due to the difficulty in observing gene conversion events in humans relatively little is known about the mechanisms involved. Current research in both humans and prokaryotes is beginning to shed some light on the mechanisms involved in gene conversion and much more is being learnt about the effects gene conversion has on the genomes of humans and other diverse species.

1.4.1, Gene conversion in humans

In humans allelic gene conversion is known to occur within recombination hotspots, analysis of which have shown the ratio of conversion and crossover events to vary between hotspots (Jeffreys and May 2004). While allelic gene conversion is not known to be pathogenic, non-allelic gene conversion has been shown to be involved in multiple pathologies (Chen et al. 2007). On the Y chromosome differences in the direction of gene conversion have been observed between different classes of Y chromosome paralog. These observed directions of gene conversion will be discussed in more detail in section 1.9 of this chapter.

1.4.2, Gene conversion in bacteria and yeast

Yeast is a much simpler model for the study of gene conversion than human, and many of the proposed gene conversion mechanisms have come from the study of this organism. Many attempts have been made to determine the mechanism of gene conversion in both humans and protozoa and differences have been observed in the effects of gene conversion occurring in different organisms. While it has been suggested that conservative gene conversion in humans acts to maintain the functional state of human genes (Rozen et al. 2003) it appears that in viruses and pathogens gene conversion regularly alters protein coat expression which helps evade the immune response (Palmer and Brayton 2007).

1.5, Proposed gene conversion mechanisms

In 1964, the geneticist Robin Holliday first proposed a mechanism of DNA strand exchange that attempted to explain gene conversion events that occur during meiosis in fungi (Liu and West 2004). Under Holliday's mechanism (Figure 1.2), a break occurs in two homologous DNA single strands (A). The broken DNA strands cross and anneal to a region of homologous sequence (B) forming a four-stranded intermediate structure (C) which has become known as a Holliday junction (HJ). Migration of the HJ (D) and resolution of the intermediate structure leads to the formation of heteroduplex DNA, a DNA molecule which is a combination of the "donor" and the "acceptor" sequences.

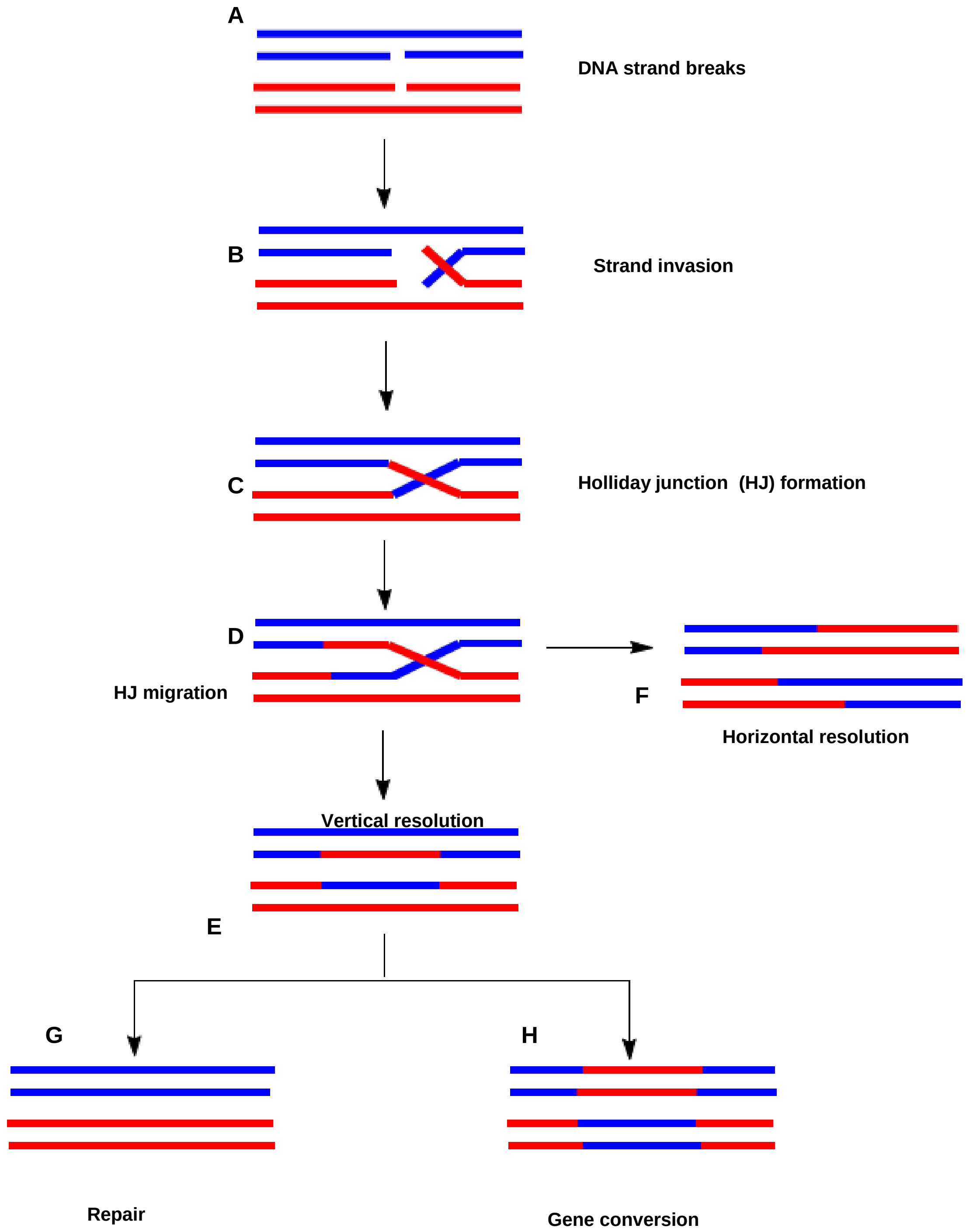


Figure 1.2: Holliday Junction formation and resolution

a) A break occurs in two homologous DNA strands

b) The broken strands cross and bind to a region of homologous sequence,

c) A four-stranded intermediate structure known as a Holliday junction (HJ) is formed.

d) Migration of the HJ occurs

e) and f) Horizontal or vertical resolution of the intermediate structure leads to the formation of heteroduplex DNA.

g) Correction of heteroduplex DNA via the mismatch repair mechanism

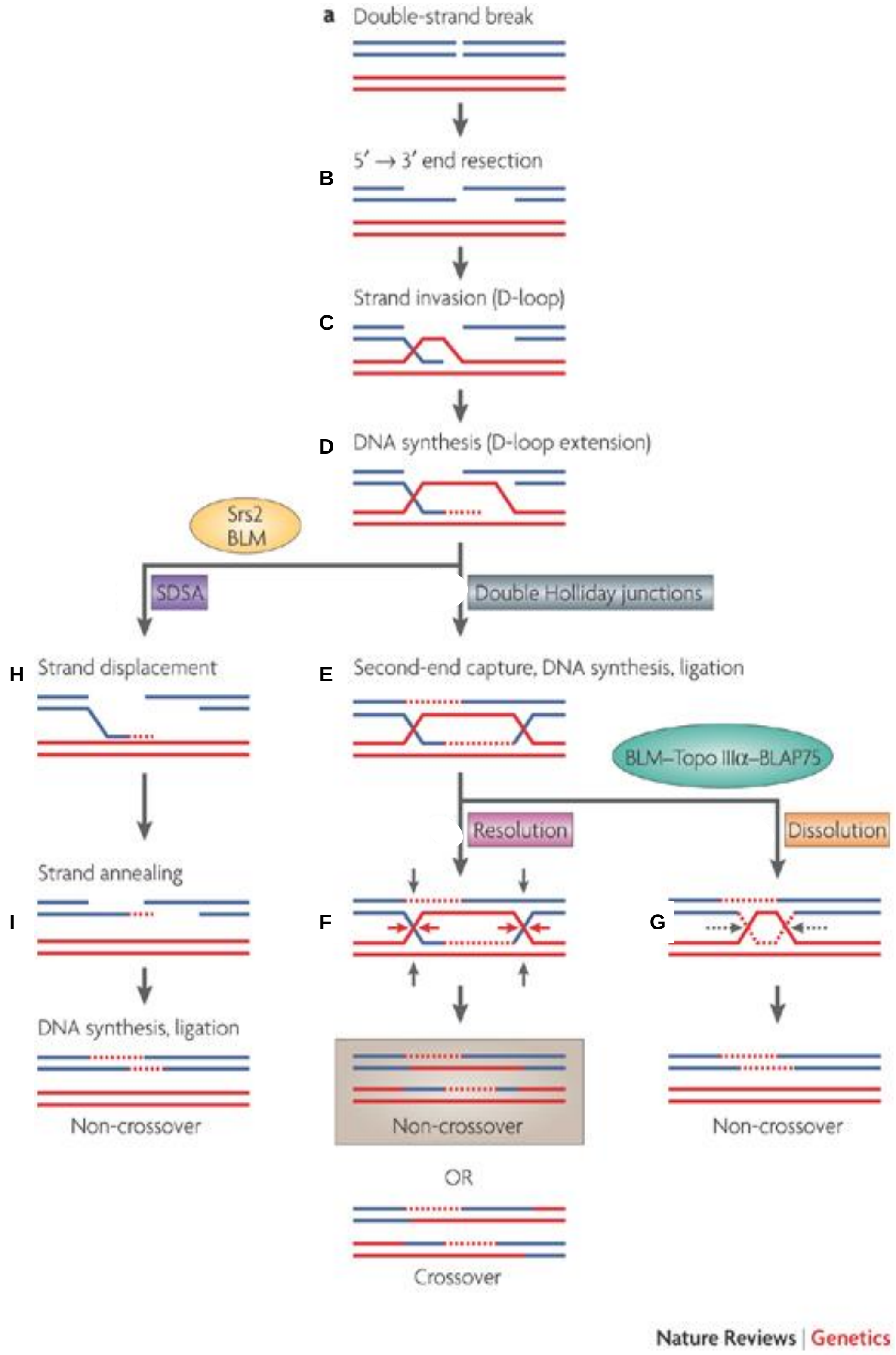
h) Correction of heteroduplx DNA via gene conversion

Resolution may occur vertically (E) or horizontally (F) producing two different types of heteroduplex DNA. This heteroduplex DNA contains mismatched bases which are either corrected via the mismatch repair mechanism (G) or via gene conversion (H). A heteroduplex DNA molecule formed by HJ resolution or via other mechanisms, may use the invaded segment as a template to correct the mismatch resulting in gene conversion. Although the mechanism has since been revised the HJ model of gene conversion is still widely accepted and has formed a basis for over 40 years of research.

A number of different mechanism have since been proposed based on the initial HJ model. These mechanisms are summarized in Figure 1.3. Currently there are three proposed mechanisms that aim to explain the processes involved in gene conversion with the initiating steps of all three proposed mechanisms being the same. A double-strand break occurs in one DNA molecule, followed by resection of the broken ends by a 5´- 3´ exonuclease, to produce two 3´-single-stranded tails which "scan" the genome for homologous sequences. Once a homologous sequence has been found, one 3´ tail invades the homologous duplex (first capture) to form a displacement (D) loop which is extended by DNA synthesis. After this point the three mechanisms diverge with different proteins believed to be involved in each mechanism.

1.5.1, The double-strand break repair mechanism (DSBR)

Following extension of the D-loop the second 3´ tail also pairs with the extended D-loop (second capture) which is followed by extension of the newly captured strand and ligation of the nicks. This produces an intermediate structure containing two HJs which



(Figure taken from Chen et al. 2007)

Figure 1.3: The proposed mechanisms of gene conversion

The three proposed mechanisms follow the same three steps before diverging into separate pathways.

a) A double-strand break occurs in one DNA molecule,

b) Resection of the broken ends by a 5' - 3' exonuclease, to produce two 3'-single-stranded tails

c) 3' tail invades a homologous duplex to form a displacement (D) loop which is extended by DNA synthesis.

The double-strand break repair mechanism (DSBR) (A-F)

d) D-loop extension

e) The second 3' tail pairs with the extended D-loop followed by extension and ligation of the nicks.

f) Cleavage of the HJs by HJ resolvase produces either crossover or gene conversion products. Produces non-cross over products only

The double Holliday Junction dissolution model (DHJD) (A-E, G)

g) The two HJs migrate towards each other and converge causing collapse of the structure.

The synthesis-dependent strand annealing model (SDSA) (A-D, H, I)

h) Following strand invasion and D-loop extension the newly synthesized strand is displaced from the template

i) Strand annealing followed by DNA synthesis and ligation of the nicks. Produces non-crossover products only.

are formed between the broken strand and its homologous sequence. Cleavage of the HJs by HJ resolvase produces either crossover or non-crossover (gene conversion) products. In this process the HJs are thought to be resolved randomly and this explains some of the features associated with meiosis; however, in this mechanism crossover and non-crossover products are expected to be produced in equal numbers but, in practice significantly more non-crossover products are observed which suggests that repair occurs more frequently by gene conversion than crossover.

1.5.2, The double Holliday Junction dissolution model (DHJD)

In the DHJD mechanism (Ira et al. 2003; Wu and Hickson 2003) the HJs are formed via the same pathway as the DSBR mechanism up to the point of double HJ formation. In contrast to the DSBR mechanism where cleavage of the HJs occurs in this process the two HJs migrate towards each other and converge causing collapse of the structure. Unlike the DSBR mechanism this gives rise to only non-crossover products. HJ dissolution has been shown to be promoted by BLM (the protein mutated in Bloom’s syndrome), topoisomerase IIIα, and BLM-associated protein (BALP75) which are believed to function together to cause convergent migration of the two HJs.

1.5.3, The synthesis-dependent strand annealing model (SDSA)

The SDSA mechanism involves the same initial steps as the DSBR and DHJD mechanisms, but does not involve the formation of HJs. In this case after strand invasion and D-loop extension have occurred, the newly synthesized strand is displaced from the template and anneals to the second 3´ tail. DNA synthesis and ligation of the nicks occurs. As with the DHJD mechanism this too only results in the production of non-crossover products. It is believed that in this pathway the BLM protein causes the newly synthesized strand to become displaced from the template (Bachrati, Borts, and Hickson 2006).

1.6, The Y chromosome

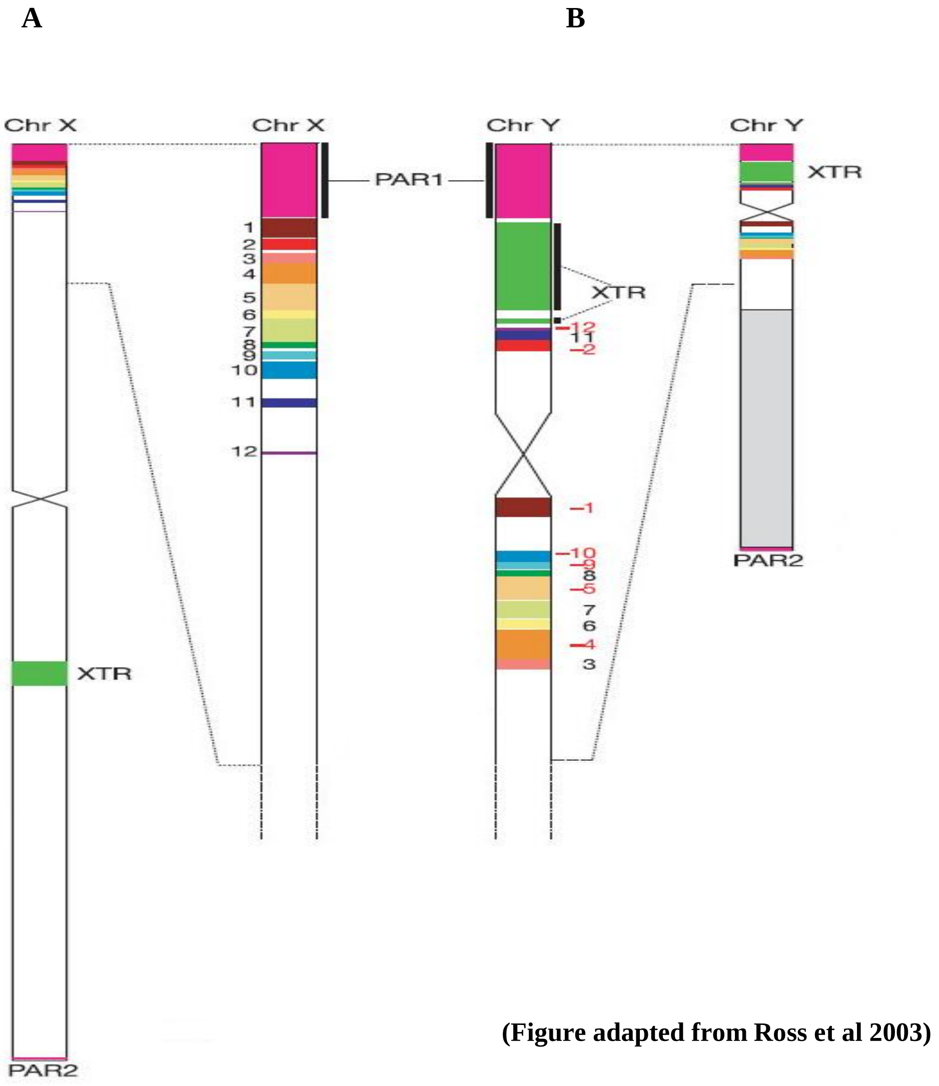
For many years the Y chromosome was considered to be a genetic wasteland, with no specific function beyond the determination of maleness. Despite this essential biological function, lack of recombination and apparently continuous gene decay led to the assumption that the Y chromosome would eventually become genetically inert (Quintana-Murci and Fellous 2001). During the past 15 years a clearer understanding of the Y chromosome has begun to emerge. What was once thought to be a gene-poor chromosome full of 'junk' repeats is now known to harbour genes vital for spermatogenesis, and has been implicated in pathologies such as gonadal sex reversal, Turner syndrome, graft rejection and male infertility (Willard 2003). The Y chromosome has since become a valuable tool which can be used in applications of genetics, such as forensic science (Jobling, Pandya, and Tyler-Smith 1997), paternity testing (Jobling, Pandya, and Tyler-Smith 1997) population studies and the study of human evolution (Tyler-Smith 2008).

The human Y chromosome is different from all other chromosomes in the human genome for a number of reasons. Firstly it is the only constitutively haploid chromosome with only 5% of its total length able to engage in pairing and crossing over with the X chromosome during meiosis. The remaining 95% does not cross over with the X and is considered to be non-recombining. It is also the only chromosome in the human genome to be strictly paternally inherited. This, combined with lack of recombination along the majority of its length, means that it is passed from father to son relatively unchanged. These properties which set aside the Y from all diploid chromosomes also make it increasingly valuable in the study of non-allelic gene conversion. Due to its haploid nature and absence of crossing over during meiosis the problems associated with inter-allelic diversity which complicates the study of gene conversion in diploid chromosomes are effectively eliminated. As the Y chromosome can only be inherited paternally and is passed from father to son, paralogs on the chromosome cannot have independent evolutionary histories. The Y chromosome also has a well-established phylogeny based on binary markers which allows individual Y chromosomes to be classified into haplogroups defining their evolutionary relationships. This cannot be achieved for any of the diploid chromosomes due to the complexity introduced by recombination.

1.6.1, Y chromosome evolution

The X and Y chromosomes are believed to have evolved from an ancestral pair of autosomes which existed approximately 300 million years ago (Bagnall et al. 2005). Regions of homology between the two chromosomes suggest that they originally existed as a homomorphic pair of autosomes which through the acquisition of sex-determining genes and evolution of mechanisms that prevent inter-chromosomal recombination have undergone substantial divergence. Comparative studies between human and primate sequences have shown that some of the inter-chromosomal homology has also arisen from recent duplication and transposition events (Stone et al. 2002).

As over 95% of the Y chromosome is non-recombining, correction via recombination with a homologous chromosome cannot occur, which has led to numerous structural changes and loss of gene content. The human Y chromosome is well known for its high level of structural variability (Jobling 2008). Cytogenetic and molecular studies have shown that many structural rearrangements exist within human populations, including several deletions (Jobling, 1996; Jobling et al., 2007; Jobling et al., 1996; Repping et al., 2003; Repping et al., 2006), duplications (Bosch and Jobling, 2003; Jobling et al., 1996; Repping et al., 2006), and inversions (Verma, Rodriguez, and Dosik 1982; Affara et al. 1986; Bernstein et al. 1986; Page 1986; Repping et al. 2006), not all of which lead to infertility. The Y chromosome is now one of the smallest chromosomes in the human genome, and while the X is estimated to contain approximately ~1000 genes (Ross, Bentley, and Tyler-Smith 2006) the non-recombining region of the Y contains approximately 80 genes which encode 27 distinct proteins (Skaletsky et al. 2003). This startling loss of genes in comparison to the X chromosome led to the prediction that eventually the Y chromosome would become devoid of all genes (Graves, Koina, and Sankovic 2006); however, despite this prediction no gene decay or loss has occurred during the last 5–7 MY of human evolution (Hughes et al. 2005). Many of the genes which remain on the Y chromosome are specialized in spermatogenesis (Lahn and Page 1999; Skaletsky et al. 2003) and in the differentiation of male structures during embryonic development. Of the 25 genes which remain homologous between the human X and Y chromosomes (Ross et al. 2005) all X chromosome copies are concentrated towards the telomeric region of Xp, whereas on the Y chromosome they are scattered across the total chromosome length. The order of the genes is not consistent between the two chromosomes (Figure 1.4), which suggests that a series of



(Figure adapted from Ross et al 2003)

Figure 1.4: X – Y homologous genes

a) On Xp the genes which remain homologous between the human X and Y chromosomes are concentrated towards the telomeric region of Xp.

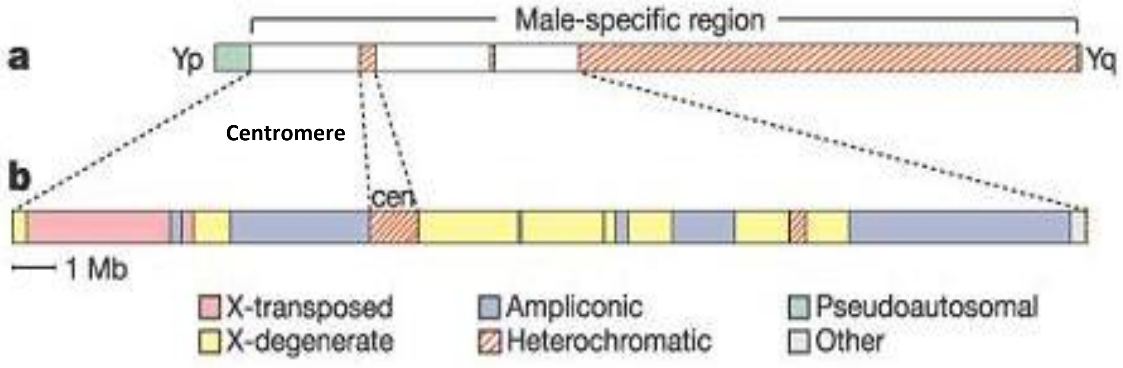
b) On the Y chromosome a series of inversions have occurred changing gene position and order and the genes are distributed across the total chromosome length.

inversions have occurred on the Y chromosome which has changed gene position and order. This has led to suggestions that the Y chromosome may be uniquely tolerant of inversions and other rearrangements during evolution (Schwartz et al. 1998). While it is thought that inversions and structural rearrangements occurring on diploid chromosomes would disrupt meiosis, sex-linked inheritance and absence of recombination on the Y chromosome has allowed many structural rearrangements to persist throughout evolution.

1.6.2, Structure of the Y chromosome

Structurally the Y chromosome can be divided into two main regions. These are the pseudoautosomal regions (PAR1 and PAR2) where obligatory crossover with the X occurs during meiosis, and the male-specific region (MSY – sometimes called the non-recombining region, NRY) (Figure 1.5).

The Pseudoautosomal regions are located at the tips of the chromosome arms, with PAR1 being located at the tip of Yp and PAR2 at the tip of Yq. PAR1 is the major pseudoautosomal region and spans approximately 2.6 Mb (Rappold 1993), while PAR2, the minor pseudoautosomal region, spans only 320 kb (Morris and Mangs 2007). While crossover at PAR2 is infrequent and not essential for normal male meiosis (Hamer and Li 1995) crossover at PAR1 is essential for normal male meiosis and chromosomal segregation to occur (Morris and Mangs 2007). Genes located within the pseudoautosomal regions can be inherited from either parent, via the same mechanisms as autosomal genes with one copy being located in the pseudoautosomal region of the Y and the other located in the gametologous portion of the X chromosome.



(Skaletsky et al. 2003)

Figure 1.5: Structure of the Y chromosome
a) 95% of the Y chromosome is male specific and does not recombine with the X chromosome during meiosis.This region has become known as the MSY
b) The euchromatic regions of the MSY consists of three classes of DNA, the X-degenarate, X-Transposed and Ampliconic.

The MSY is formed from a combination of euchromatin and heterochromatin. The heterochromatic regions are highly condensed DNA segments which contain no genes and are transcriptionally inert (Skaletsky et al. 2003) and of little apparent biological significance. The euchromatic regions on the other hand contain the majority of genes which are responsible for important biological functions including those which are vital for normal male development. There are three different classes of euchromatin located within the MSY, all of which have arisen via different mechanisms during Y chromosome evolution.

1) X-degenerate euchromatic sequences are relics of the ancestral autosomes from which the X and Y chromosomes originally evolved. These regions contain 16 functional genes and 13 pseudogenes which show between 60% and 96% sequence similarity to their X-linked gametologs. Located within the X-degenerate segment of the short arm is the important sex-determining gene *SRY*. When SRY is expressed in the developing embryo, the gonads specialize as testes, which in turn secrete two types of hormones that trigger the differentiation of Sertoli cells (Sekido and Lovell-Badge 2008) during male development.

2) X-transposed euchromatin originated as the result of an X to Y transposition between Xq21 and Yp 3-4 million years ago (Schwartz et al. 1998) after the human and chimpanzee lineages had diverged. On the Y chromosome the X-transposed region spans approximately 3.4 Mb and exhibits 99% sequence similarity to DNA sequences located within Xq21. This region contains the two genes *TGIF2L* and *PCHDH11* which have homologous copies on Xq21 and Yp.

3) Ampliconic euchromatic sequences result from intra-chromosomal duplications and comprise over 45% of the MSY (Skaletsky et al. 2003). The ampliconic regions contain the highest density of genes with over 60 genes having been identified (Skaletsky et al. 2003). Many of the genes located within the ampliconic regions are duplicated and comprise nine gene families all of which are involved in spermatogenesis (Bhowmick, Satta, and Takahata 2007).

1.6.3, The Y chromosome paralogs

Multiple paralogs exist within the ampliconic regions of the Y chromosome (Figure 1.6) The two main classes of paralog are those that are formed into palindromes and inverted repeats. These paralogs exhibit >98% sequence similarity and have the potential to undergo gene conversion. Other types of repeated sequence also exist on the Y chromosome including structures such as Alu and L1 elements and HERVs which have previously been shown to undergo gene conversion (Batzer et al. 2003; Bosch et al. 2004) . As well as paralogs which are located solely on the Y chromosome, parology also exists with the X chromosome and some autosomes.

1.6.3.1, The Palindromes

DNA palindromes are the most pronounced structural feature of the ampliconic regions of Yq (Skaletsky et al. 2003) ranging in size from 30kb to 2.9Mb and comprising over 25% of the MSY euchromatin. Eight palindromes are located on the long arm of the Y chromosome; these palindromes each consist of two virtually identical paralogs or “arms” which are located on the same DNA strand and are in most cases separated by a non-duplicated spacer sequence. One arm is situated in the forward orientation while the second arm is in the reverse orientation. When distinguishing between palindrome arms in this thesis the arm located closest to the centromere will be referred to as the

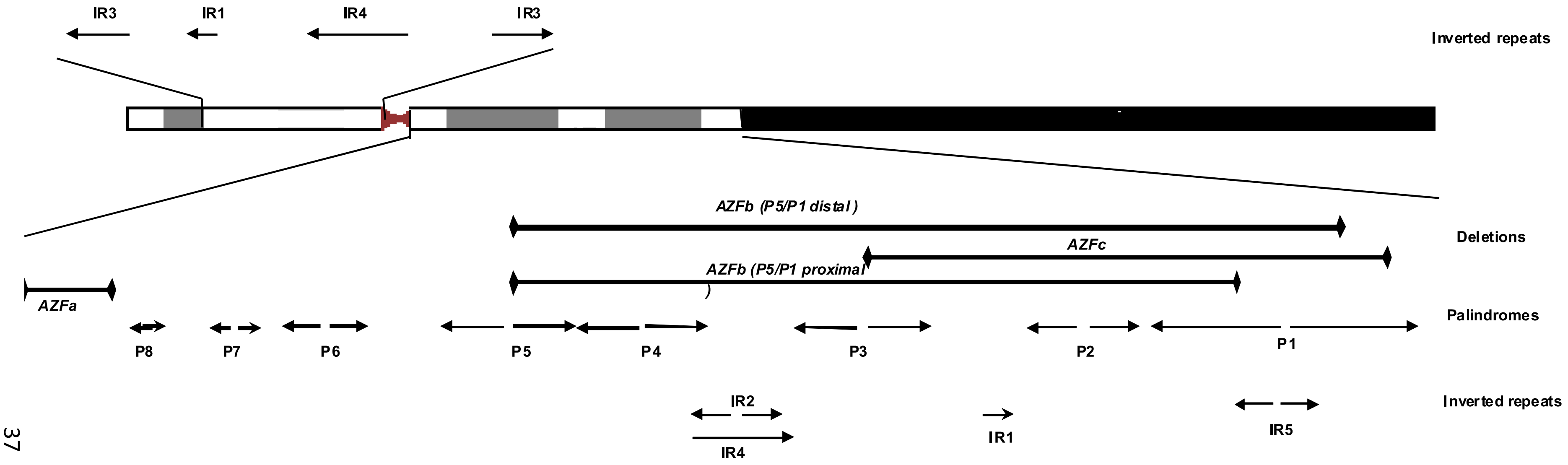


Figure not drawn to scale

Figure 1.6: Location of Y the chromosome paralogs
Multiple paralogs exist within the ampliconic regions of the Y chromosome, the two main classes of paralog are palindromes (P1-P8) and inverted repeats (IR1-IR4). Both classes of paralog are very similar in structure and while the palindromes are located solely on the long arm of the chromosome (Yq) the IRs are distributed on both the long and the short arm (Yp) some IRs are situated in regions of the Y chromosome which are known to be involved in multiple deletions.

proximal palindrome arm while the arm situated further away from the centromere will be referred to as the distal palindrome arm. Comparative sequencing data suggest that abundant gene conversion has driven the concerted evolution of these palindrome arms (Rozen et al. 2003), leading to >99.5% sequence similarity between palindrome arms compared to only 90% similarity observed between duplicated sequences in the rest of the genome (Samonte and Eichler 2002). Of the eight palindromes identified on the human Y chromosome, six bear protein-coding genes, all of which are primarily expressed in the testis. Each of the genes located within a palindrome has at least one identical copy located on the opposite arm of the palindrome. Some genes such as *DAZ* exist in multiple copies as a result of multiple duplication events. By comparison, most genes located outside of the palindromes such as *AMELY* and *PRKY* have only one copy located on the Y chromosome, with the only exception being the *TSPY* genes of which there are multiple copies situated on Yp.

1.6.3.2, The Inverted repeats

In addition to the eight palindromes, four sets of inverted repeats are also located within the ampliconic regions of the Y chromosome. Inverted repeats are similar in structure to palindromes but contain much larger spacer regions and exhibit lower sequence similarity between paralogs (Table 1.1). Inverted repeats exhibit sequence similarity of 99.6% - 99.9%, (Skaletsky et al. 2003) suggesting that gene conversion has the potential to occur. Similarly to the palindromes, it is hypothesized that these structures are capable of folding around the spacer region forming hairpin structures that undergo gene conversion. In contrast to the palindromes which are only located on the long arm of the Y chromosome, inverted repeats are spread across both chromosome arms. While

	Location	**Arm length (kb)**	**Spacer length (kb)**	**% similarity**	**Genes**
Palindromes					
P1	Yq	1 450	2.1	99.95	*DAZ, CDY*
P2	Yq	122	2.1	99.99	*DAZ, BPY2*
P3	Yq	283	170	99.99	*RBMY, PRY*
P4	Yq	190	40	99.99	*HSFY*
P5	Yq	496	3.5	99.99	*CDY, XRKY*
P6	Yq	110	46	99.96	None
P7	Yq	8.7	12.6	99.99	None
P8	Yq	36	3.4	99.99	*VCY*
Inverted repeats					
IR1	Yp & Yq	65	15,102	99.66	None
IR2	Yq	62	249	99.95	*RBMY*
IR3	Yp	298	3,601	99.75	*TSPY*
IR4	Yp &Yq	275	12,353	93.76	*RBMY*

Table 1.1: Different classes of paralog located on the Y chromosome
Two main classes of paralog, palindromes (P) and Inverted Repeats (IRs) are situated on the Y chromosome. Both are very similar in structure: however, the paralogs of palindromes are typically separated by smaller spacers and display higher inter-paralog sequence similarity than IRs. While palindromes are situated solely on Yq paralogs of IRs are situated on both Yp and Yq (Skaletsky et al. 2003).

both copies of the IR2 repeat are located on Yq and both copies of IR3 are located on Yp, IRs 1 and 4 both have one copy located on Yp and the second located on Yq. Inverted repeats are located in more complex regions than palindromes, exhibiting multiple regions of paralogy within the Y chromosome. This makes the study of gene conversion between inverted repeats more complicated as there is the potential for gene conversion to occur between multiple paralogs. IR2 and the Yq paralog of IR1 are also located within regions of the Y chromosome which are known to undergo non-pathogenic deletion (Repping et al. 2002). This can complicate the study of gene conversion and could lead to the false identification of gene conversion events as what appears to be two homogeneous copies might in fact be a single copy. While gene conversion has been shown to occur between some Y chromosome palindromes (Rozen et al. 2003), conversion between inverted repeats has not been studied. Evidence of gene conversion between paralogs of IR1 or IR4 would be interesting as this would suggest that the Y chromosome can fold around the centromere to allow recombination between Yp and Yq. While gene conversion has not been studied between paralogs of IRs, NAHR is known to occur which results in Y chromosome inversions. It has previously been reported that a paracentric inversion occurs as the result of NAHR between the paralogs of IR3 (Repping et al. 2006). It has also been hypothesised that IR1 and IR4 may sponsor a pericentric inversion that occurs between Yp and Yq which cause INV(Y)(p11q11) (Causio et al. 2000). As NAHR can occur between paralogs of IRs and cause inversions, there is also the potential for gene conversion to occur between these regions.

1.6.3.3, Other Y chromosome repeat elements

Various other classes of repeat elements exist on the Y chromosome, which are not classed as paralogs have been shown to mediate gene conversion, of these, LINE (Tremblay, Jasin, and Chartrand 2000) and Alu (Sen 2006; Zhi 2007) elements have both been shown to mediate gene conversion. Alus comprise 10% of the human genome, generally spanning 300bp and sharing 70-100% sequence similarity (Batzer et al. 2003). Gene conversion has been shown to occur frequently between neighbouring Alus with 15 000 – 85 000 point mutations thought to be caused by gene conversion events (Chen et al. 2007). Gene conversion has also been shown to occur between LINEs although at a lower frequency than between Alus (Chen et al. 2007). Although many LINE-mediated conversion events have been observed *in vitro* and in transgenic mice, relatively few LINE-mediated conversion events have been observed in humans (Vincent 2003; Myers et al. 2005). LINE-mediated gene conversion is thought to be less frequent than that involving Alus, due to lower sequence similarity, larger “spacer” distance between paralogous copies, and their being located in AT-rich, gene-poor regions (Chen et al. 2007).

1.6.3.4, Gametology with the X chromosome

The Y chromosome also contains multiple regions of gametology with the X chromosome. This gametology can be between X-degenerate regions on the X and Y chromosomes or between the Xq21 and the 3.2Mb of the X transposed region of the Y chromosome. Gene conversion has been shown to occur at a translocation hotspot adjacent the X-degenerate *PRKX* and *PRKY* genes (Rosser et al. 2009) while no evidence of gene conversion between gametologous genes on Xq21 and the XTR has been observed.

1.6.3.5, Sequence variation between Y chromosome paralogs

When paralogous sequences are aligned, differences may be identified between the two sequences which are known as paralogous sequence variants (PSVs). Similarly, differences between paralogous regions of the X and Y chromosomes represent gametologous sequence variants (GSVs). These variants may be single nucleotide differences, insertions and deletions or microsatellite length variations.

As gene conversion acts to homogenize two sequences, conversion events can only be identified in regions where PSVs have previously been identified. It has been hypothesized that during gene conversion between Y chromosome paralogs the DNA strand folds around the spacer sequence forming a hairpin structure which enables the two arms to align (Figure 1.7). This aligning of the paralogs allows the exchange of material between the two sequences via NAHR. Spacer size is thought to influence the rate of gene conversion with paralogs which are separated by smaller spacers being shown to exhibit a higher degree of sequences similarity in comparison to those which are separated by larger spacers (Chen et al. 2007).

1.6.4, Polymorphic markers on the Y chromosome

There has been keen interest in using polymorphisms on the Y chromosome to examine questions about paternal genetic relationships among human populations since the mid

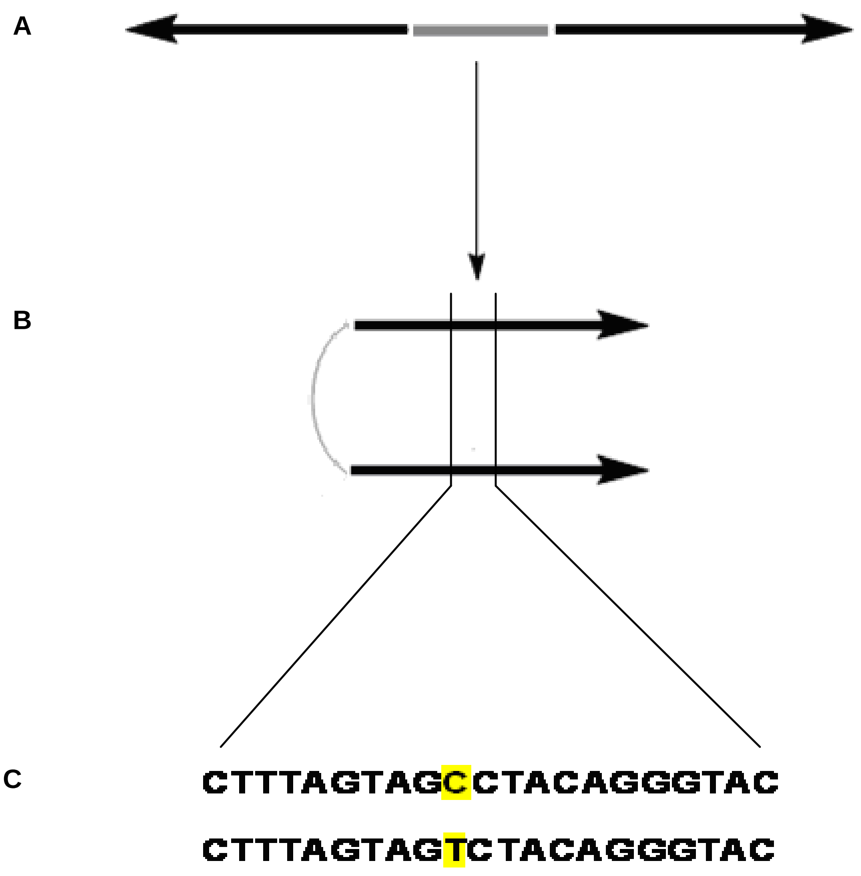


Figure 1.7: Inter-paralog gene conversion

a) Palindromes consist of two duplicated sequences (paralogs) which are separated by a non-duplicated spacer sequence.

b) It has been hypothesised that palindromes can fold around the spacer region forming a hairpin structure allowing the paralogs to align.

c) Paralogous sequence variants (PSVs), which have arisen through mutation can be corrected by gene conversion.

1980s (Casanova et al. 1985). Mutations occurring on the Y chromosome result from intra-allelic processes which also occur in diploid chromosomes but their interpretation is not complicated by allelic diversity or recombination on the haploid Y chromosome. There are two types of polymorphisms that are commonly typed on the Y chromosome - binary markers which include SNPs and small insertions and deletions (indels), and microsatellites. Currently, there are over 500 characterized binary markers as well as over 200 informative microsatellites on the Y chromosome (Karafet et al. 2008). Since they are plentiful and relatively easy to type and interpret, binary markers and microsatellites are the most informative markers in population studies and for following evolutionary history. Slowly evolving binary markers can be used to define haplogroups while the faster evolving microsatellites can be used to study more recent events within haplogroups and populations. By virtue of its many polymorphisms, the Y chromosome is now considered to be the most informative haplotyping system, with applications in evolutionary studies, forensics, medical genetics, and genealogical reconstruction.

1.6.4.1, Binary markers and the Y phylogeny

The first Y-chromosome DNA polymorphisms were published in 1985, but over the following decade very few additional polymorphisms were identified and by the end of 1996 less than 60 polymorphisms had been discovered (Hammer and Zegura 2002). In 1997 Underhill et al. (Underhill et al. 1997) published 19 additional polymorphisms which had been identified by denaturing high performance liquid chromatography (DHPLC). Since the introduction of this method and some systematic resequencing projects, many more polymorphisms have been discovered.

Binary markers (SNPs) are the main class of marker typed on the Y chromosome. These are particularly useful as, on the Y chromosome, they represent unique events in human evolution and have a low mutation rate of 10^{-8} per base per generation (Nachmana and Crowella 2000). As the Y chromosome is haploid all SNPs located outside of the PARs are observed in either the ancestral or derived state as determined from comparisons with the chimpanzee ortholog. In contrast, SNPs located within the PARs or on diploid chromosomes can be observed in one of three states, homozygous for either the ancestral or derived allele and heterozygous. As Y chromosome binary markers present only one allele they may be used in combination to define monopyletic haplogroups.

From 1997 the sudden increase in Y chromosome polymorphisms being identified led to a number of different nomenclature systems being introduced for the same haplogroups. To overcome this problem an attempt was made to unite the growing number of nomenclature systems into one that could easily be shared between publications. As a result, in 2002 the Y Chromosome Consortium first introduced the YCC tree which was subsequently updated in 2003 (Jobling and Tyler-Smith 2003) after the identification of additional polymorphisms (Figure 1.8). The 2003 version of the YCC tree included 243 binary markers which when typed in combination allowed individual Y chromosomes to be subdivided into haplogroups which are arranged alphabetically into clades termed A-R. Each clade may also be further subdivided and arranged into alphanumerically named subclades. This allows Y chromosomes to be assigned to a particular haplogroup based on a combination of allelic states for certain binary markers. In 2008 an updated version of the tree was introduced which comprises 586 binary markers defining 311 haplogroups (Karafet et al. 2008) arranged into clades A-T. This has led to a change in nomenclature of some branches of the tree

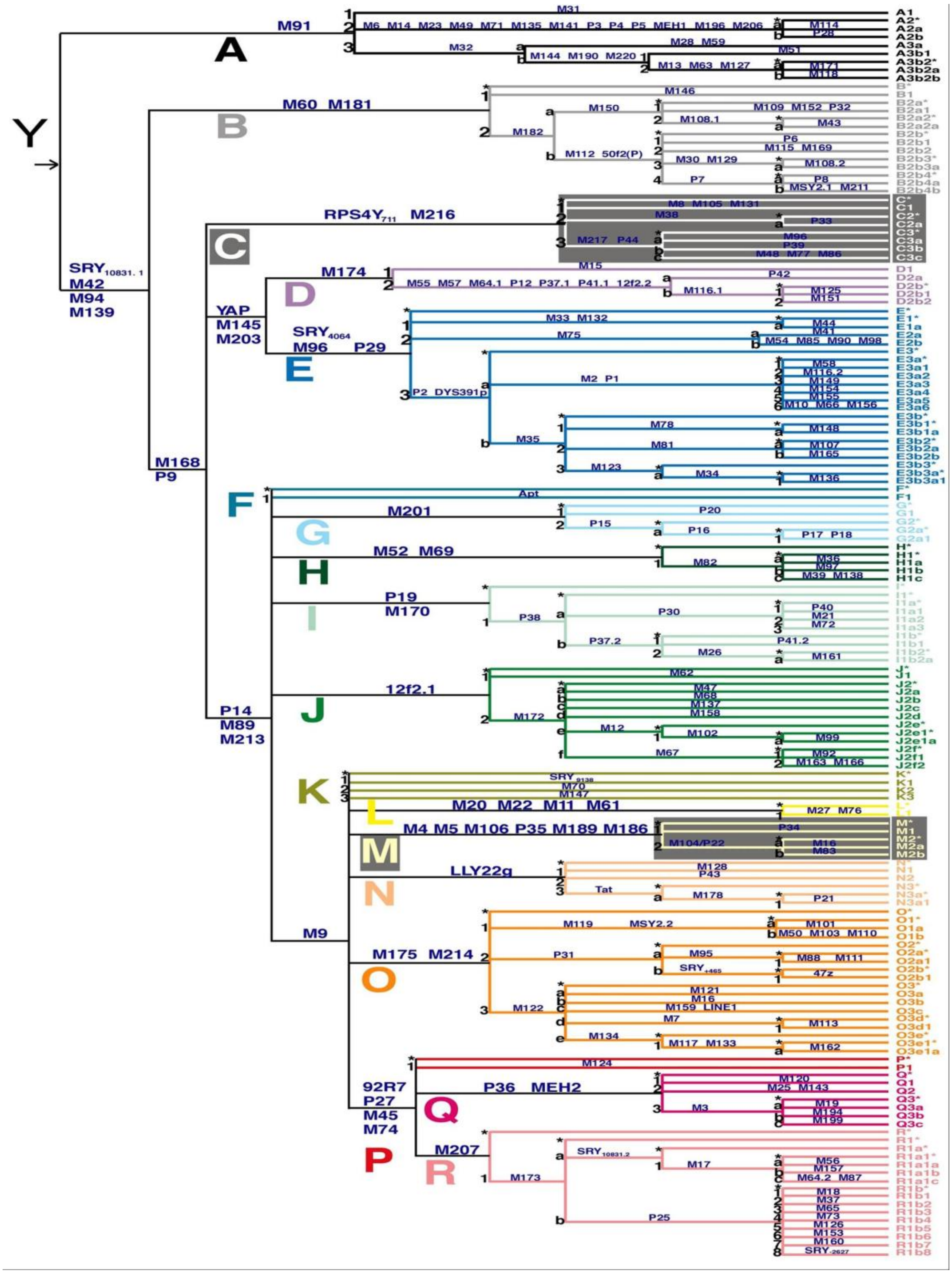


Figure 1.8: The 2003 YCC tree

The Y phylogeny allows Y chromosomes to be assigned to a particular haplogroup based on a combination of allelic states for certain binary markers. The 2003 YCC tree (Jobling and Tyler-Smith 2003) comprised 243 binary markers which allows Y chromosomes to be subdivided into haplogroups. Haplogroups are arranged alphabetically in clades termed A-R and each clade may be subdivided and arranged alphanumerically into subclades.

while additional haplogroups and subclades have been introduced (Figure 1.9). With improved sequencing technologies and decreased costs of genome sequencing, many more binary markers are likely to be identified which will lead to further changes in YCC tree topology and nomenclature (Karafet et al. 2008). Publication of the 1000 Genomes Project (http://www.1000genomes.org) is likely to identify additional binary markers which will lead to further changes to the Y phylogeny.

1.6.4.2, The depth of the Y phylogeny

In the Y phylogeny a new haplogroup is defined when a unique marker arises on a single Y chromosome and as the haplogroup grows in frequency diversity accumulates through mutation of the linked markers such as microsatellites. The amount of intra-haplogroup diversity among members of a population can be related to the age since they last shared a common ancestor. In contrast to the Y chromosome, the process of recombination complicates the dating of alleles at autosomal and X-linked loci. Intra-allelic diversity at markers linked to a specific allele is generated not only through introduction of a new allele by mutation but also by replacement of the ancestral allele at the linked locus through recombination.

A number of different methods have been used to estimate the time to the most recent common ancestor (TMRCA) of a set of Y chromosomes sharing a mutational change at a unique marker. Introduction of such methods has led to estimates of the time at which all Y chromosomes shared a common ancestor as well as the TMRCA of individual clades of the phylogeny. Estimates of the TMRCA for the Y phylogeny have varied and range from 50 000 years (Thomson et al. 2000) to 188 000 years, with a 95%

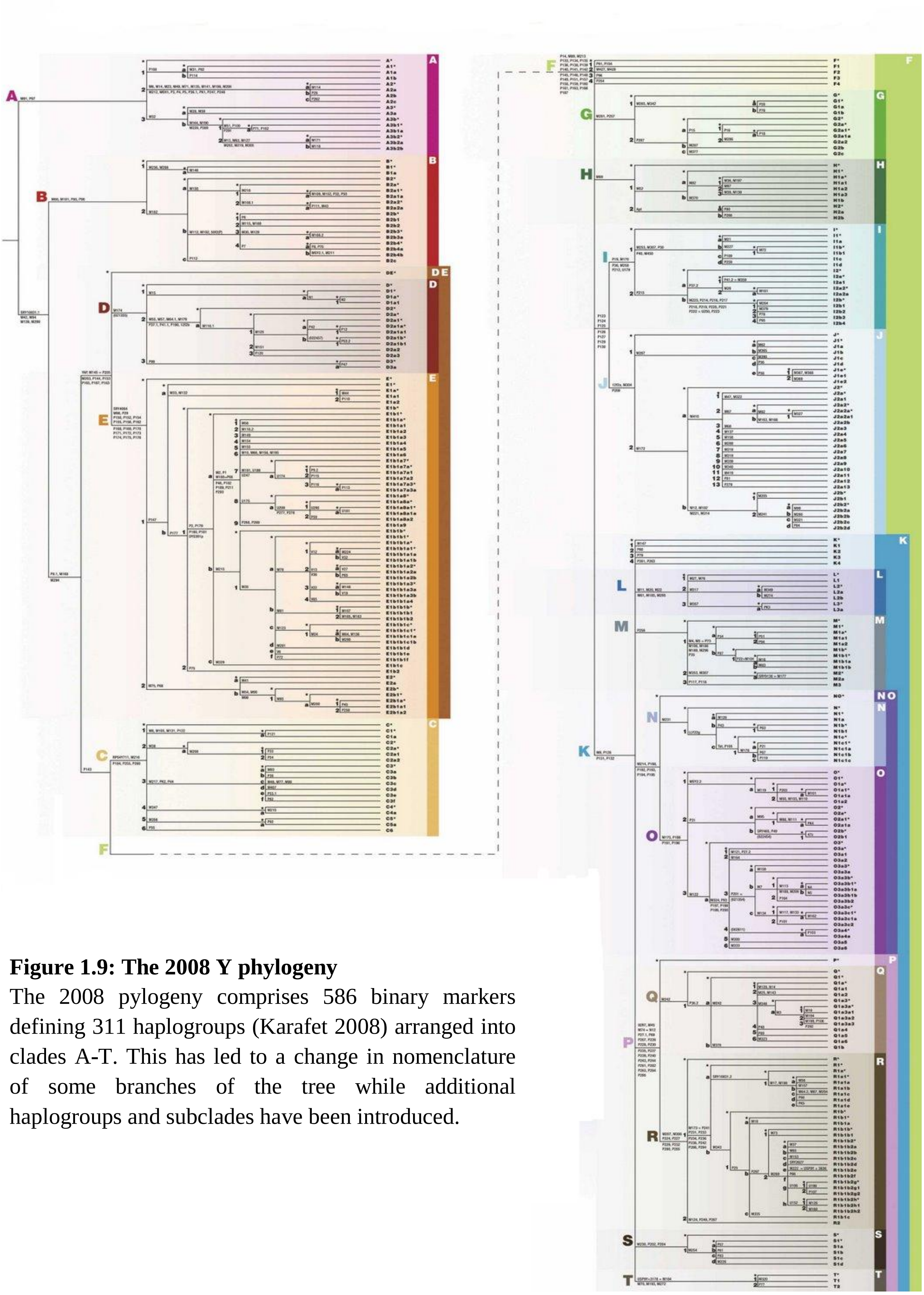


Figure 1.9: The 2008 Y phylogeny

The 2008 pylogeny comprises 586 binary markers defining 311 haplogroups (Karafet 2008) arranged into clades A-T. This has led to a change in nomenclature of some branches of the tree while additional haplogroups and subclades have been introduced.

Figure adapted from Karafet et a1. 2008

confidence interval from 51,000 to 411,000 years (Hammer 1995). Zegura and Hammer (2002) used GENETREE (Bahlo and Griffiths 2000) to determine different clades. The recent 2008 YCC tree (Karafet et al. 2008) has led to the TMRCA of some clades to change, with some binary markers having originated earlier than previously thought, although the overall TMRCA of the Y phylogeny remains the same. An additional advantage of knowing the TMRCA of a set of Y chromosomes is that it can be used to estimate the TMRCA of the entire Y phylogeny to be 90,040 years. As updating the YCC tree changes nomenclature and topology it can also change the TMRCA estimates of the rate of gene conversion occurring between a diverse set of Y chromosomes. The methods used will be discussed in section 1.7.2 of this chapter.

1.6.4.3, Microsatellites

Despite binary markers forming stable haplotypes their use in population studies is limited due to the slow rate of mutation. While combinations of allelic states can be used to divide Y chromosomes into haplogroups, microsatellites can be used to define more informative haplotypes within a haplogroup.

Microsatellites are comprised of repeats of 2-6 nucleotides which occur typically up to 30 times in tandem arrays. Microsatellites are faster mutating than binary markers with mutation rate generally increasing as the number of repeats increases, with a typical mean mutation rate estimated at $1x10^{-3}$ per microsatellite per generation. Microsatellites can be highly polymorphic, making them valuable genetic markers in many fields of genetics.

The majority of widely used microsatellites are tri- and tetra-nucleotide repeats of which there are >200 on the Y chromosome (Kayser et al. 2004). Due to the lack of recombination on the Y chromosome, variation occurring within microsatellites can only be mutational and microsatellite diversity is not influenced by other factors such as unequal crossover. Typing Y chromosome microsatellites is a highly efficient way both to distinguish between Y chromosomes and to indicate haplotype relationships, and they can even be used to predict Y chromosome haplogroups (Schlecht et al. 2008) using software such as haplogroup predictor available from http://www.hprg.com/hapest5.

Relationships between Y chromosome microsatellites are often displayed as median-joining networks (Bandelt, Forster, and Röhl 1999). Microsatellite networks can be constructed based on data from individual or multiple microsatellite markers (Figure 1.10). Within a network, chromosomes which share a particular haplotype are grouped together within a node and the size of the nodes are proportional to the number of chromosomes sharing that particular haplotype. The nodes are separated by lines which represent the number of mutational differences between the haplotype for each node and the length of each line is proportional to the number of mutational differences between haplotypes. Each individual chromosome can be colour-coded based on the information required from the network. For example chromosomes may be colored based on the population, haplotype or Y-chromosomal haplogroup.

1.6.5, The chimpanzee Y chromosome

There are four subspecies of *Pan troglodytes* (Pt), *P.t.verus, P.t. vellerosus, P.t. troglodytes and P.t. schweinfurthi,* each of which are resident in different regions of Africa. In this study, reference to the chimpanzee will refer to the *P.t. troglodytes*

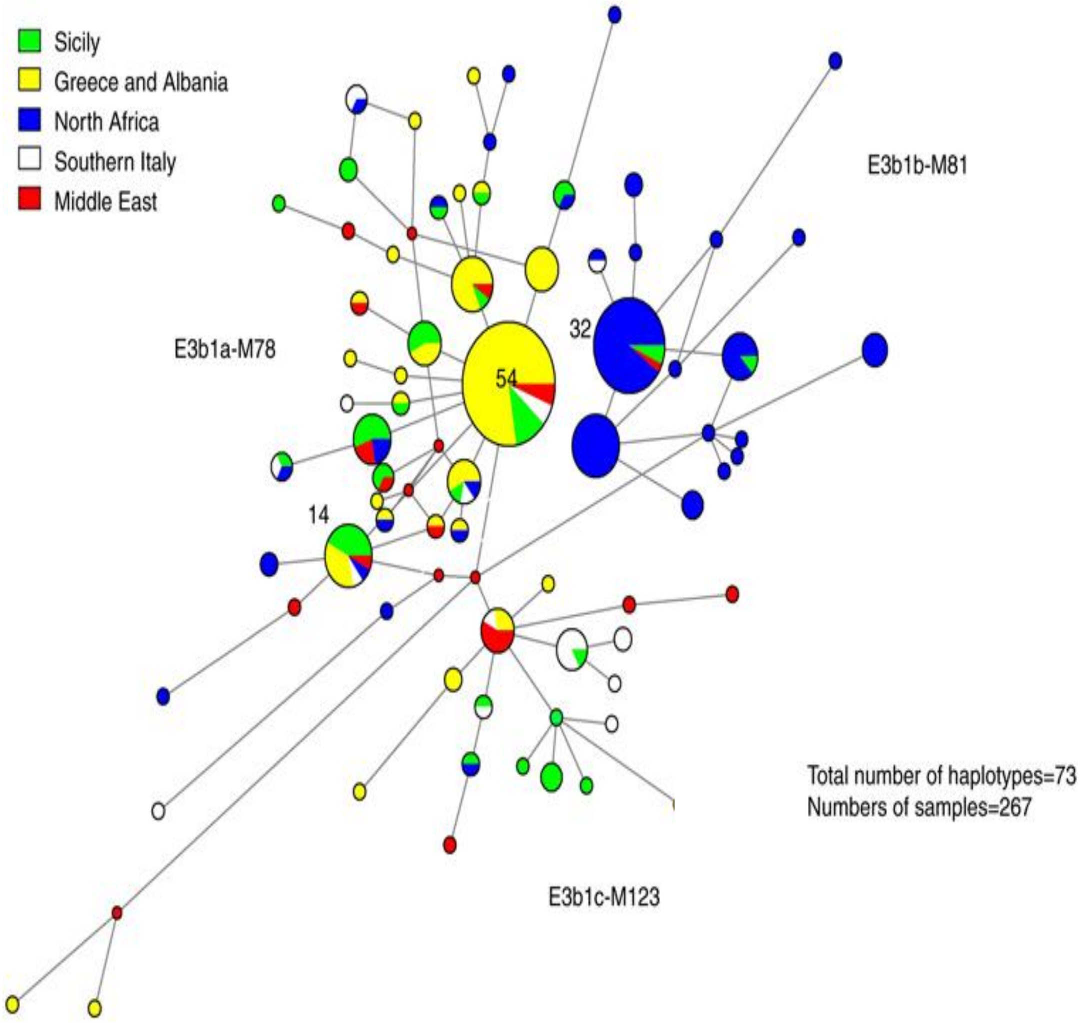


Figure 1.10: Microsatellite networks

Microsatellite networks display the relationships between Y chromosome haplotypes. Each node within a network represents a particular haplotype and the node size is proportional to the number of chromosomes which carry that haplotype. Nodes are joined by a series of lines which are proportional in length to the number of mutational steps which separate each haplogroup Each chromosome can be colour coded based on the desired information, such as haplogroup or population.

subspecies which are found in central Africa. DNA from chimpanzee mainly comes from a captive member of subspecies *Pan troglodytes troglodytes.*

Evolutionarily, chimpanzees are the closest living species to humans and sequence comparisons have begun to reveal a spectrum of genetic changes that have accompanied human evolution (Kehrer-Sawatzki and Cooper 2007). Following the divergence from their common ancestor humans and chimpanzees have clearly evolved in different ways with some studies suggesting that gene loss may contribute significantly to the divergence between the two species (Olson 1999). Despite the clear phenotypic differences between humans and chimpanzees, at a genomic level the two species are very similar with only 1.2% - 1.4% sequence divergence observed between species in alignable DNA (Stone et al. 2002). Divergence between the human and chimpanzee Y chromosome sequences is known to be higher at approximately 1.7%, (Stone et al. 2002) due to the higher mutation rate of the Y chromosome. The human Y chromosome is much larger than the chimpanzee Y chromosome covering apporoximatly 60Mb compared to 35Mb for the chimpanzee (Ross et al. 2005). However, this is largely due to human-specific heterochromatin and therefore likely to be of little biological significance. The human and chimpanzee Y chromosomes both show many structural differences which must have occurred after species divergence. Over the past 6 million years the human Y chromosome has retained all 27 protein-coding genes while some of the orthologous genes in chimpanzee have sustained inactivating mutations (Stone et al. 2002).

Analysis of the chimpanzee Y chromosome sequence is valuable in the study of gene conversion for several reasons. Not only can the chimpanzee sequence provide evidence of the ancestral state of human sequences, divergence calculations can also give an indication as to whether gene conversion may be occurring (Rozen et al. 2003) and of any possible biases in the direction of gene conversion. Comparisons of the human and chimpanzee Y chromosome have been complicated due to the gaps in the chimpanzee sequences and until recently the reliability of the available chimpanzee sequence has been questioned. In 2010 Hughes et al. published a finished chimpanzee reference sequence (Hughes et al. 2010); this has allowed more detailed human and chimpanzee comparisons to be carried out. All currently available chimpanzee sequences will be discussed in more detail in Chapter 3.

Comparisons with an independent primate species such as gorilla or macaque can also provide a deeper rooting evidence of the ancestral state of human PSVs, as gene conversion can occur in both humans and chimpanzee. However, currently sequence is only available for female gorilla and macaque and offers no additional information on the evolution of Y chromosome palindromes.

1.6.5.1, The chimpanzee Y phylogeny

Humans and chimpanzees are believed to have diverged from a common ancestor approximately 5-6 million years ago (MYA) while the gorilla is believed to have shared a common ancestor with humans approximately 5-7 MYA (Chen and Li 2001; Brunet et al. 2002) (Figure 1.11). Although a Y chromosome phylogeny does exist for the chimpanzee (Stone et al. 2002) it is not as well defined as it is for humans. A problem with studying the Y chromosome of the chimpanzee is that sequence data are only available for chimpanzees which are in captivity the majority of which have been shown

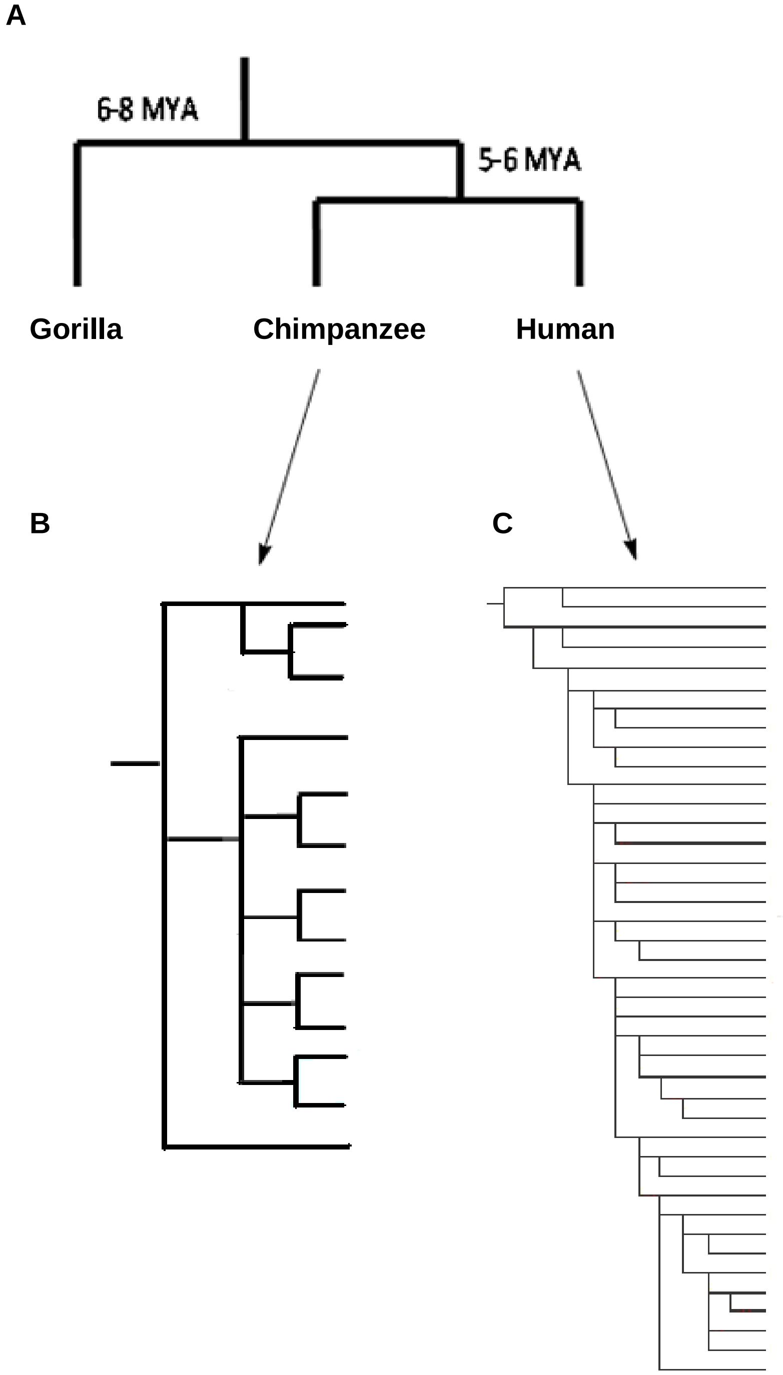


Figure 1.11: The human and chimpanzee Y phylogeny.

Humans and chimpanzee diverged from a common ancestor approximately 5-7 million years ago (MYA) while the gorilla is believed to have shared a common ancestor with humans approximately 6-8 MYA (A). While the human Y chromosome has a well defined evolutionary phylogeny (B) that of the chimpanzee is less well defined (C).

to belong to the *Pan troglodytes* troglodytes subspecies (Stone et al. 2002). This is analogous to a human sample containing many individuals from a single population with very limited haplogroup diversity and therefore multiple chimpanzee sequences offer only limited information

1.6.5.2, Chimpanzee palindromes

Rozen et al. (2003) have previously sought evidence that palindromes existed in the common ancestor of humans and chimpanzees. This was achieved by looking for orthologs of palindromes in common chimpanzee, bonobos and gorillas. PCR was used to amplify the inner and outer palindrome boundaries in all three species, with the presence of a PCR product for both boundaries being taken as evidence that the palindromes exist in a particular species. From this it was determined that P1, P2, P6, P7 and P8 are all present in the chimpanzee genome and P1, P2, P6 and P7 are present in bonobo, while only P4 and P6 are present in gorilla. This provides evidence that the five palindromes P1, P2, P6, P7 and P8 existed before the human and chimpanzee lineages diverged and therefore cannot be the result of a human-specific duplication event. As the other palindromes have been shown to predate speciation, the low divergence cannot be attributed to a more recent duplication event and is most likely to be due to ongoing gene conversion. Three palindromes P3, P4 and P5, do not appear to be present in chimpanzee and may have arisen as the result of a more recent duplication event.

At the time this study was commenced the majority of available sequence for the chimpanzee Y chromosome represented the X-degenerate regions of the Y chromosome and little sequence was available for the ampliconic regions due to their repetitive nature and consequent difficulties of sequencing. The limited sequence data for the ampliconic regions came from the work of Rozen et al. (2003) which established the existence of some palindrome boundaries in the chimpanzee genome prior to speciation: however, this offers little information on the internal palindromic structure, and also little is known about the presence of IRs in the chimpanzee genome. Recent publication of the complete chimpanzee sequence has identified 19 palindromes in the chimpanzee genome (Hughes et al. 2010) and should offer more information on the ancestral state of human palindromic structures.

1.6.6, Gene conversion on the Y chromosome

Over the past seven years substantial evidence has begun to emerge that gene conversion occurs between paralogs located on the Y chromosome as well as between X and Y copies of some X-Y homologous genes. This section discusses the findings of four of these studies and will go on to discuss the methods used to detect gene conversion events.

1.6.6.1, The study of Rozen et al. 2003

The first study published by Rozen et al. in 2003 presented evidence of gene conversion occurring between Y chromosome palindromes and focused mainly on the *CDY* genes located within the arms of P1. The first part of the study compares sequence divergence between human and chimpanzee palindrome sequences to determine whether gene conversion occurs between the arms of palindromes. It then goes on to seek more direct evidence of gene conversion occurring between the arms of P1 in humans. Sequencing of Bacterial Artificial Chromosomes (BACs) which correspond to P1, P2, P6 and P7 in chimpanzee was carried out and the intra- and inter-specific sequence divergence calculated. Divergence between palindrome arms in both humans and chimpanzees was very similar at 0.028% between chimpanzee arms and 0.021% between human arms while interspecies divergence between orthologous palindrome arms was calculated as 1.44%. This in itself provides evidence of gene conversion in that conversion of a mutation into a sequence creates diversity between species but at the same time homogenizes sequences within a species. Interspecies divergence between palindrome arms was also compared to that of the spacer region in order to assess how non-duplicated sequences have diverged compared to duplicated sequences. As spacers are non-duplicated they cannot undergo gene conversion and should more closely represent normal human and chimpanzee divergence following speciation. Interspecies sequence divergence between the spacers was calculated as 3.2% which is significantly higher (P =0.0001, 2-tailed Fisher exact test) than the 1.4% observed between palindrome arms. From these data Rozen et al. (2003) suggest that gene conversion effectively favours the ancestral sequence. The study then goes on to provide direct evidence of gene conversion occurring between the *CDY* genes located within P1 in humans. Three variant sites were observed to undergo gene conversion to both the derived and ancestral allele as determined from human and chimpanzee sequence comparisons. For example (Figure 1.12), a synonymous C/T variant was typed in 171 males, covering 42 branches of the YCC tree. Conversion events to and away from the ancestral state were observed within five separate branches of the phylogenetic tree. From this study it was concluded that - assuming an average generation time of 20 years - the rate of gene conversion which would be needed to explain the observed divergence between *CDY* genes would be 2.2 x 10^{-4} conversions per duplicated nucleotide per generation. From this it was estimated that in each new-born male up to 600bp of sequence are converted.

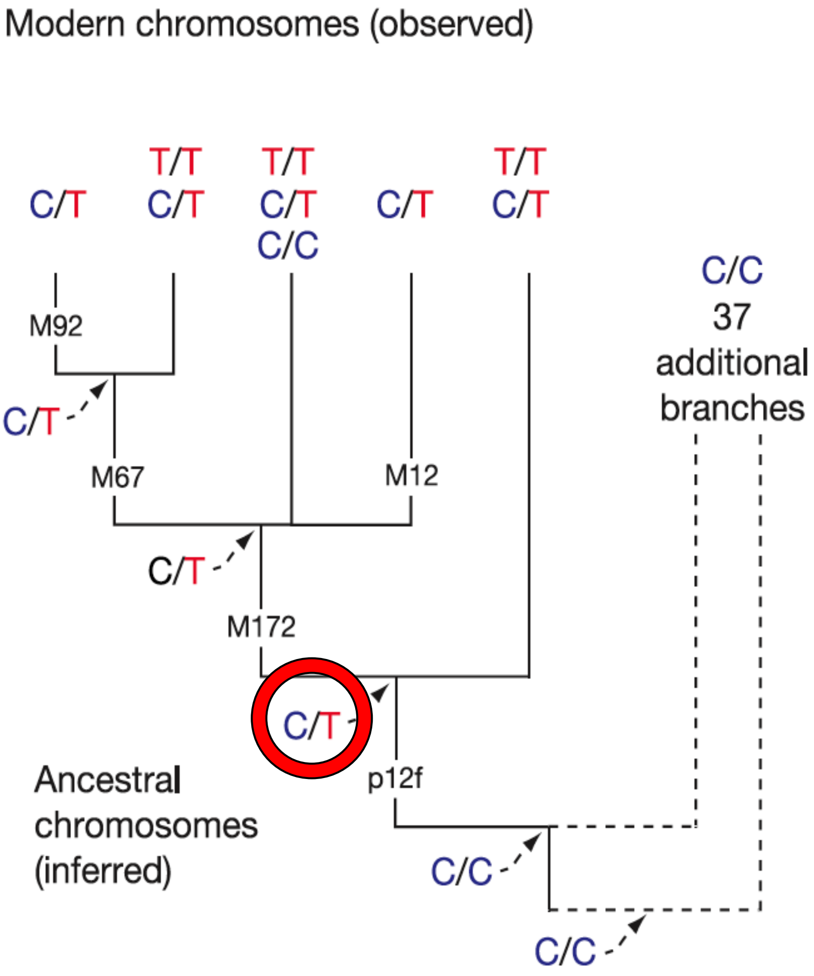


Figure adapted from Rozen et al. 2003

Figure 1.12: Evidence of inter-paralog gene conversion
A synonymous C/T variant was typed in 171 males, covering 42 branches of the YCC tree. Conversion events were observed within five branches of the Y phylogeny with three variant sites undergo gene conversion to both the derived and ancestral allele. the gene conversion rate was calculated as 2.2 x 10^{-4} conversions per duplicated nucleotide per generation - assuming an average generation time of 20 years.

1.6.6.2, The study of Bosch et al. 2004

The second study by Bosch et al. published in 2004 looked at conversion occurring between two directly repeated ~10-kb Human Endogenous Retroviral Sequences (HERVs) which flank the *AZFa* region located on Yq. *AZFa* is a region of ~780kb which contains genes that are essential for spermatogenesis. It has previously been shown that NAHR occurring between the two HERVs can result in deletion (Blanco et al. 2000; Kamp et al. 2000; Sun et al. 2000) and duplication (Bosch and Jobling 2003) of the *AZFa* region. Overall 94% sequence similarity is exhibited between HERV sequences; however, four blocks of complete sequence identity have been observed (Blocks A-D) and break points have been identified within blocks A B and C. Bosch and Jobling (2003) previously identified two gene conversion events occurring in a region lying between identity blocks A and D known as the inter-AD region.

The study by Bosch et al. (2004) examined conversion events occurring within this region in more detail. From alignment of the proximal and distal HERV sequences, 24 PSVs were identified and sequencing in 33 Y chromosomes from across the Y phylogeny did not identify any additional PSVs. Of the 24 PSVs observed, one distal-to-proximal conversion event was observed while 22 proximal-to-distal conversion events were seen showing a directional bias which favoured proximal-to-distal conversion (Figure 1.13). From these data the rate of gene conversion occurring between HERV sequences was estimated to between $2.5x10^{-4}$ and $1.3x10^{-3}$ per base per generation.

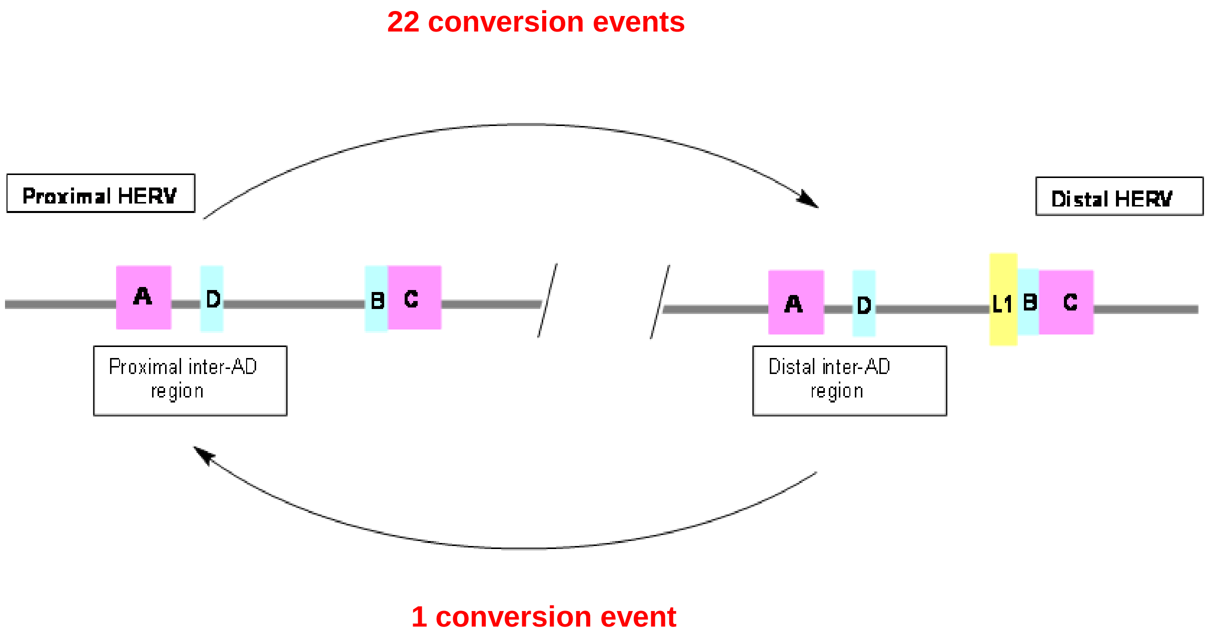


Figure not drawn to scale

Figure 1.13: Evidence of gene conversion occurring between two Human Endogenous Retroviral Sequences (HERVs)

Analysis of 24 PSVs identified between proximal and distal HERV sequences, identified one distal-to-proximal conversion and 22 proximal-to-distal conversion. This shows a directional bias which favours proximal-to-distal conversion. The overall rate of conversion was estimated to be $2.5\text{x}10^{-4}$ and $1.3\text{x}10^{-3}$ per base per generation

1.6.6.3, The study of Trombetta et al. 2009

In addition to gene conversion occurring between paralogs located on the Y chromosome, gene conversion has also recently been shown to occur between several X-Y homologous genes. Trombetta et al. (2009) analyzed sequence variation between X-Y homologous genes located within three different regions of the MSY, the X-degenrate, X-transposed and ampliconic regions. This study carried out analysis of the *PCDH11Y* gene located in the X-transposed region (X-Y identity 99%), the *TBL1Y* gene located in the X-degenerate region (X-Y identity 86%-88%), and the *VCY* genes situated in P8 of the ampliconic regions (X-Y identity 95%). In this study the Y chromosome gene was sequenced in individuals from diverse Y chromosome haplogroups and the sequence aligned with the March 2006 X chromosome reference sequence. No evidence of gene conversion was observed between the *PCDH11Y* or *TBL1Y* genes; however, evidence was provided which suggests that the *VCY* genes act as a sequence acceptor from *VCX* during gene conversion. The resulting *VCY* sequences were shown to share homology with 1-4 regions of the X chromosome – which is most likely to be due to copy number variation of the *VCX* genes. From sequencing of 122 males from diverse Y chromosome haplogroups, it was determined that gene conversion occurs at a rate of 1.8 x 10^{-7} X-to-Y conversion events per nucleotide per year. As the study of Trombetta et al. (2009) only carried out comparisons of the *VCY* genes with the X chromosome reference sequence it is possible that gene conversion may occur between *VCY* and different *VCX* genes therefore *VCY-to-VCX* gene conversion events may have been unidentified.

1.6.6.4, The study of Rosser et al. 2009

In a study by Rosser et al. (2009) resequencing of X and Y copies of a known translocation hotspot adjacent to the *PRKX* and *PRKY* genes provided evidence of historical bidirectional gene conversion between the human MSY and the X chromosome. Sequencing of 1.9-kb X- and Y-specific segments in twelve males representing diverse Y chromosome haplogroups and populations showed eleven of the twelve Y-chromosome sequences to be identical to the reference sequence. However, in one chromosome carried by a Namibian male from haplogroup A2c, two GSVs separated by 4 bp were shown to carry the allelic state of the X-chromosome suggesting that both GSVs lie within the same conversion tract. This sub-region was sequenced in an additional 23 diverse Y chromosomes and the same tract was observed in two additional haplogroup A2c Y chromosomes suggesting a common ancestry for the conversion event. Within the same sub-region a longer tract of conversion containing four GSVs was observed in a chromosome belonging to haplogroup Q. Analysis of a further 32 haplogroup Q chromosomes from diverse populations failed to identify any additional conversion events suggesting a single conversion event has occurred. In this study the conversion tract was shown to vary between chromosomes with the average tract length estimated to be approximately 100bp.

In this study the rate of gene conversion was estimated to be 3.8 x 10^{-8} and 1.7 x 10^{-6} per base per generation, with the lower value being similar to the average Y chromosome base mutation rate of 2.3 x 10^{-8} per base per generation (Repping et al. 2006), and the upper value being two orders of magnitude slower than the 2.2 x 10^{-4} per base per generation rate of gene conversion occurring between palindrome arms (Rozen et al. 2003).

1.7, Estimating the rate of gene conversion

The studies of gene conversion outlined in this chapter have each used different methods to estimate the rate of gene conversion occurring between Y chromosome paralogs. As the Y chromosome has a known time-depth over which gene conversion events have occurred, it is possible to estimate the rate of conversion between Y chromosome paralogs. This is not possible for the diploid chromosomes which are complicated by recombination, which introduces inter-allelic diversity and precludes the construction of a single coherent phylogeny. This section will discuss the methods which have previously been used to estimate the rate of gene conversion occurring between different classes of Y chromosome paralogs.

1.7.1, The method of Rozen et al. 2003

The first method that will be discussed is from Rozen et al. (2003) who used the known MSY mutation rate and observed divergence between palindrome arms to determine the rate of conversion that would be required to explain the observed sequence divergence. In this study the following equation was used.

$$c=2u/d$$

Where ***c*** represents the gene conversion rate, ***u*** is the known human MSY mutation rate (1.6×10^{-9}) per nucleotide per year which is multiplied by two as the sequence is duplicated, and ***d*** is the observed divergence between palindrome arms (3×10^{-4})

Therefore **c** = $(2 \times 1.6 \times 10^{-9}) / 3 \times 10^{-4}$

$$c = 1.1 \times 10^{-5}$$ gene conversions per duplicated nucleotide per year

To determine the conversion rate per generation an assumption is made that the average generation time is 20 years:

Therefore **c** = $2.2\text{x}10^{-4}$ gene conversions per duplicated nucleotide per generation

From this Rozen et al. (2003) were able to determine how many base pairs are converted per generation. The human palindromes are known to cover 5.4Mb, so as these are duplicated regions the palindromes contain $2.7 \text{ x } 10^{6}$ duplicated bases. From this it was calculated that up to 600bp of sequence undergo conversion in each new born male.

1.7.2, The method of Bosch et al. 2004

In the study by Bosch et al. (2004) the rate of gene conversion was calculated using the known TMRCA of the Y phylogeny. To estimate the rate of gene conversion it was first of all necessary to know the total amount of time over which the observed conversion events have occurred. This was done by estimating the maximum and minimum plausible elapsed time since all the chromosomes analyzed shared a common ancestor, by using published TMRCA estimates based on coalescent analysis from Hammer and Zegura (2002) for each haplogroup. This produced a range of 18,686 to 90,274 generations which was most likely to encompass the actual time over which the conversion events occurred. As this study showed gene conversion events to be directional the rate of conversion was calculated for each direction, proximal-to-distal and distal-to-proximal. For proximal to distal gene conversion 22 conversion events were shown to have occurred over 18686 to 90274 generations.

Therefore 22 / 18686 = 1.2x10^{-3} and 22 / 90274 = 2.4x10^{-4}

Giving an average rate of conversion for proximal to distal conversion of 2.4x10^{-4} - 1.2x10^{-3} conversion events per generation.

For distal to proximal conversion only 1 conversion event was observed and using the same principle as above an average rate of 1.1x10^{-5} – 5.3x10^{-5} conversion events per generation was estimated, which is approximately 20 fold lower than that of proximal-to-distal gene conversion.

From the sum of these individual rates the overall conversion rate between the two sequences is estimated to be between 2.5x10^{-4} and 1.3x10^{-3} conversion events per generation.

1.7.3, The method of Trombetta et al. 2009

In the study by Trombetta et al. (2009) a modified version of the equation of Repping et al. (2006) was used to estimate the rate of X-to-Y gene conversion in the *VCY* region.

The equation for ***C*** is:

$$\boldsymbol{C = \frac{N/ttot}{l\ x\ d}}$$

Where ***N*** represents the number of X-to-Y gene conversions (*N*=9), ***ttot*** represents the total time spanned by all branches in the tree for the 122 chromosomes analyzed (*ttot*=899,750) ***l*** represents the length of the region analyzed (*l* =1,616 bp) and ***d*** represents the average X-Y sequence diversity between each of the 122 Y chromosomes and the gametologous regions on the X *(d=0.035).* From this a rate of 1.8 x 10^{-7} X-to-Y conversion events per base per year was estimated between genes.

1.8, Problems associated with identifying gene conversion events

Despite the haploid nature of the Y chromosome which eliminates some complications in the study of non-allelic gene conversion, additional problems in the identification of conversion events still remain. As the Y chromosome is prone to various structural variations which have the potential to arise in any Y chromosome without affecting fertility (reviewed by Jobling, 2008), pseudohemizygosity due to deletion must be distinguished from pseudohomozygosity caused by gene conversion. Also due to the identical or near identical sequence which surround PSVs located in palindromes and IRs, sections of both paralogs will be co-amplified during PCR, producing pseudohomozygous or pseudoheterozygous results when variants are analyzed (Figure 1.14). This means that for pseudoheterozygous PSVs it is not known which allele of the variant lies within each paralog. This is analogous to the problem of "phase", as it is not known which allelic state is associated with which arm of the palindrome. While it is likely that the PSVs will remain in the same phase within each chromosome, NAHR may also cause the alleles to "switch" between paralogs in some chromosomes through rearrangements such as inversions. The issue of phase also creates problems when typing duplicated microsatellites as it is not known which arm of the palindrome is associated with each microsatellite haplotype. The issue of phase is more of an issue when typing duplicated microsatellites as each microsatellite will mutate independently. When typing duplicated microsatellites, population geneticists have traditionally used an arbitrary method to assign the alleles with the shorter allele being assigned to locus 1 and the larger allele to locus 2 (Balaresque et al. 2007).

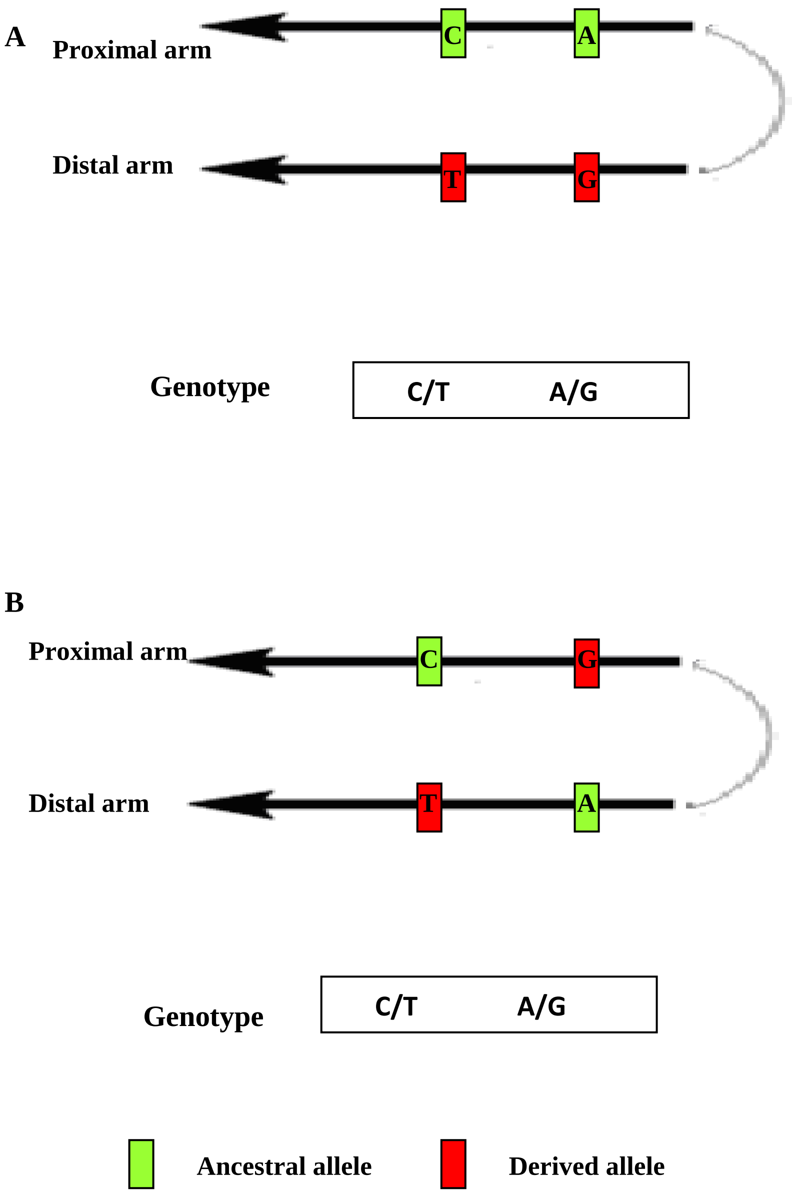


Figure 1.14: The problem of phase

a) Co-amplification of both paralogs during PCR creates the issue of “phase” as a geneotype will be produced but it is not known which allele is associated with each arm of the palindrome

b) Rearrangements such as inversions can switch alleles between palindrome arms, however, due to co-amplification during PCR the resulting genotypes for each PSV will remain the same

When typing GSVs identified between X and Y gametologous regions, the issue of phase is less of a problem as X-Y differences can be taken advantage of to design chromosome-specific primers. Gene conversion from the X to the Y chromosome is relatively easy to identify: however, the problem of meiotic segregation and interallelic diversity poses a problem when looking for evidence of Y-to-X gene conversion. As the X chromosome is three times more prevalent in the population than the Y chromosome, a gene conversion event may be passed on multiple times leading to over representation of a single conversion event: additionally, a conversion event may be lost during meiosis. The issues of identifying gene conversion events and how they may be overcome will be discussed in more detail in Chapter 5.

1.9, Observed biases in the direction of gene conversion

Growing literature is beginning to suggest that gene conversion between different regions may be directional. On the Y chromosome Rozen et al. (2003) compared interspecies sequence divergence between the palindrome arms and the spacer region and suggested that gene conversion between palindrome arms is conservative of the ancestral state, while Bosch et al. (2004) showed that gene conversion between HERVs flanking the *AZFa* region favours proximal-to-distal conversion.

Gene conversion has also been shown to occur between duplicated mammalian genes (Galtier et al. 2001; Galtier and Duret 2007). It has been suggested that AT-rich regions of DNA are more prone to mutation than GC-rich regions, due to the reduced number of hydrogen bonds. A bias towards increased GC content of a region is thought to stabilise AT-rich regions making mutations less likely. Gaiter et al. (2007) hypothesise that a G or C allele will convert an A or T allele with a higher probability than the reverse, resulting in an increase of GC content of genomic regions undergoing frequent gene conversion (Galtier et al. 2001). When considering BGCgc between palindrome arms, frequent gene conversion would be expected to increase the CG content of palindrome arms relative to the spacers which do not undergo gene conversion. There are several factors which could also influence the GC content of palindrome arms. Conversion of Alus which are GC-rich and LINEs which are GC-poor into the palindrome arms would alter the GC content of palindrome arms relative to the spacer. Another problem when comparing palindrome arms to the spacers is that GC content varies between genic and non-genic sequences, and since the arms of palindromes but not the spacers tend to contain genes this could lead to a bias in GC content due to the presence of genes, rather than BGCgc.

1.9, Aims of this thesis

The aim of this study is to determine whether gene conversion occurs between various different classes of paralog located on the Y chromosome as well as between regions of paralogy between the X and Y chromosomes outside of the pseudo-autosomal regions. As the Y chromosome is rich in paralogs which display as little as 80% and up to 99.9% sequence similarity, a thorough bioinformatic exploration will be carried out in order to identify regions which have the potential to reveal historical gene conversion events by sequence comparisons among human Y chromosomes. Regions where conversion has already been shown to occur will not be included in this analysis. From this bioinformatic exploration, regions which appear to be good candidates for gene conversion will be examined experimentally making use of DNAs representing diverse populations carrying haplogroups from across the Y phylogeny. This will involve typing snPSVs and microsatellites in a panel of males in order to identify conversion events. Regions where gene conversion is shown to occur will be further examined in an attempt to estimate the rates of conversion, tract length and whether there is a bias in the direction of gene conversion.

Chapter 2: Materials and Methods

2.1, Materials

2.1.1, Suppliers

Applied Biosystems (Warrington, Cheshire), National Diagnostics (Hessel, UK), Qiagen (Crawley, UK), New England Biolabs (Hertfordshire, UK), Sigma (Dorset, UK), Amersham (Buckinghamshire, UK), Promega (Southampton UK). Kappa Biosystems (Essex, UK), Abgene (Epsom, UK), Invitrogen (Paisley, UK) Edge Biosystems (Gaithersburg, USA).

2.1.2, Commonly used reagents

AmpliTaq gold (Applied Biosystems), Big dye Terminator V1.1 (Applied Biosystems), GeneScan™ 120 LIZ™ Size Standard (Applied Biosystems), GeneScan™ 500 LIZ™ Size Standard (Applied Biosystems), SNaPshot master mix (Applied Biosystems), Shrimp Alkaline phosphatase (Amersham), Exonuclease I (New England BIolabs), Formamide (National Diagnostics), Kappa Taq (Kappa Biosystems), BSA (New England Biolabs), Oligonucleotides (Sigma), RepliG mini kit (Qiagen), dNTPs (Promega), *Phi*X *Hae*III DNA ladder (Abgene), λ/*Hin*d III (Invitrogen).

Reagents prepared at The University of Leicester

11.1 x PCR Buffer (Jeffreys et al. 1990)	45 mM Tris-HCI (pH 8.8) 11 mM $(NH_4)_2SO_4$, 4.5mM $MgCl_2$ 8.7 mM B-mercaptoethanol, 4.5 uM EDTA 1mM each of dATP, dCTP, dGTP, dTTP, 110 ng/ml bovine serum albumin
10xTBE	0.89M Tris borate, 2mM EDTA (pH8.3)

2.1.3, DNA samples

2.1.3.1, Genomic DNAs

Genomic DNAs from the CEPH-HGDP panel (Cann et al. 2002) were diluted to a concentration of 5ng/µl. Genomic DNAs from a Himalayan sample set (de Knijff et al. 2009) were diluted to a concentration of 5ng/µl.

2.1.3.1, Whole genome amplified (WGA) DNAs

64 male DNA samples representing 31 different Y chromosome haplogroups and 17 populations were selected from the CEPH–HGDP diversity panel (Cann et al. 2002) (Supplementary table S2.1). All samples were subject to whole-genome amplification (WGA) by the multiple-displacement amplification method (Dean et al. 2002) using the RepliG midi Kit (Qiagen).

A subset of eight chromosomes which represent the major haplogroups of the Y phylogeny were selected for sequencing. These chromosomes represented individuals from haplogroups A(xA3b2a), B2b4, E1b1b1c, G, J2, P*, O3a3c and R1b1b2.

2.1.4, Oligonucleotides

All primers were designed based on the March 2006 version of the human reference sequence. The Tm for each PCR primer was estimated to be 60°C – 62°C based on the A/T = 2°C and G/C = 4°C rule. For microsatellite typing, the 5´ ends of forward primers were labelled with a fluorescent dye (FAM or HEX) which allows fragment detection during capillary electrophoresis.

SNaPshot extension primers were designed based on the human reference sequence and are comprised of the 20 nucleotides immediately adjacent to the PSV of interest. For multiplex SNaPshot reactions a poly-A tail may be added to the 5´ end of the extension primer to alter mobility during capillary electrophoresis and allow clear definition of products, typically allowing a minimum of 4bp of separation.

2.1.5, The ABI 3130xl

Fragment and sequence analysis was carried out on an ABI 3130xl Prism Genetic Analyzer (Applied Biosystems) using Pop 4 polymer (Applied Biosystems) and a 36cm 16 capillary array. For SNaPshot and microsatellite typing, fragment size can be determined through running the product along with a fluorescently labeled size standard which allows the size of DNA fragments to be compared to a set of fragments of known size. This is a more sensitive method for determining fragment size than gel electrophoresis, and is particularly useful for microsatellite typing as variations of as little as 1bp can easily be visualized.

2.2, Methods

2.2.1, Whole genome amplification (WGA)

Genomic DNA from the CEPH-HGDP panel (Cann et al. 2002) were amplified using the Qiagen RepliG mini kit. This method allows amplification of limited DNA samples utilizing a WGA technique called Multiple Displacement Amplification and provides unbiased and accurate amplification of whole genomes to produce an effectively unlimited supply of DNA. 10ng of genomic DNA was amplified to a final concentration of approximately 10µg in a 50µl reaction (according to the manufacturers' protocol).

2.2.2, Rehydrating oligonucleotides

Oligonucleotide primers were supplied lyophilised, and were rehydrated to a final concentration of 100µM using dH_20 according to the manufacturer's recommendations. Primers for PCR were generally diluted to a 10µM working stock using dH_20, with the final primer concentration varying from 0.5µM to 10µM in PCR reactions.

2.2.3, Determining optimal PCR conditions

The optimum conditions for each primer pair were determined by carrying out titrations of temperature, (at 55°C, 58°C, 60°C, 62°C and 65°C), cycle number (20, 25,30,35) and annealing and extension times (10s - 60s).

2.2.4, PCR amplification

PCR was carried out in a Tetrad Thermocycler (MJR) using 1-2µl of WGA DNA or 5-10ng of genomic DNA using either Buffer II (Applied Biosystems) and 0.1U Taq Gold

(Applied Biosystems) or the buffer of Jeffreys et al. (Jeffreys, Neumann, and Wilson 1990) with 10µl BSA and 1U Kappa Taq in a final volume of 10µl.

Cycling conditions for PCR were generally: 95°C 5 minutes followed by 94°C 30s, 60°C 30s, 70°C 60s, for 35 cycles, while for microsatellite typing conditions were generally 94°C 30s, 60°C 20s, 62°C 30s, for 20 cycles, although annealing temperature and cycle number were varied where stated.

2.2.5, Gel electrophoresis

To verify the success of PCR, products were run on an agarose gel ranging from 1-3% (w/v) depending on fragment size. Agarose gels were comprised of 100ml 1XTBE, 1-3g agarose powder and 0.02 µg/ml ethidium bromide. 2µl of PCR product were run against 3µl of *Phi*X *Hae*III or λ/*Hin*d III size markers depending on expected fragment size. Gels were run in 1xTBE at 8V/cm over 1-2 hours. Fragments were visualised and photographed via a Syngene Geneflash transilluminator.

2.2.6, PCR product purification

Unincorporated primers and dNTPs were removed from the remaining 8µl of PCR product by addition of 1µl (1U) Exonuclease I (New England BIolabs) and 4µl (4U) Shrimp Alkaline Phosphatase (Amersham) to produce a final volume of 12µl. Products were incubated at 37°C for 2 hours followed by enzyme inactivation at 80°C for 5 minutes.

2.2.7, Sequencing reaction

Samples were either sequenced using Big Dye V1.1 and run on an ABI 3130xl, or using Big Dye V3.1 and run on an ABI 3070xl via the Protein and Nucleic Acid Chemistry Laboratory (PNACL) at the University of Leicester.

2.2.7.1, Big Dye v3.1 protocol

Where stated, samples were sequenced by the Protein and Nucleic Acid Chemistry Laboratory (PNACL), at the University of Leicester. 8µl of purified PCR template was supplied to PNACL with 1 pmol of forward or reverse primers. Reaction conditions were 94°C 30s, 96°C 10s, 40°C - 60°C 5s, 60°C 4 minutes, for 25-30 cycles, and annealing temperature and the number of cycles were varied according to primer Tm and fragment size. Products were purified by adding 2µl of 2.2% (w/v) SDS followed by heating to 96°C for 5 minutes and were then passed through an EDGE (EdgeBio) spin column. 10µl of formamide were added to each sample and products were run on an ABI 3730xl using the parameters detailed in supplementary table S2.2a.

2.2.7.2, Big Dye v 1.1 protocol

2-3µl of purified PCR template were sequenced using 3.2µM of forward or reverse primers and 3µl of Big Dye terminator v 1.1 (Applied Biosystems) to give a final volume of 8µl. Reaction conditions were generally 94°C 30s, 60°C 60s, 70°C 2 minutes, for 29 cycles. Annealing temperature and cycle numbers were varied where stated.

Unincorporated dye terminators were removed by using the DyeEx Spin Kit v2.0 (Qiagen) and dried down by heating at 80°C for 60 minutes. Products were resuspended in 10µl formamide (National Diagnostics) and the purified products were run on an ABI 3130xl using the parameters detailed in supplementary table S2.2b.

2.2.7.3, Sequence analysis

Sequence data were obtained on an ABI 3130xl (big dye v1.1) or 3730xl (big dye v3.1) Prism Genetic Analyzer and analyzed using Sequence Analysis v3.7 (Applied Biosystems) or BioEdit v7.0.5 (Hall 2005) and sequences aligned using ClustalW (Higgins 2003). Pseudoheterozygous sites were indicated by using the standard IUB codes (Figure 2.1).

2.2.8, Overview of the SNaPshot minisequencing assay

The SNaPshot (Applied Biosystems) minisequencing assay involves two main steps. First, primary PCR amplification of the region encompassing a PSV of interest is carried out followed by purification of the PCR product using SAP and ExoI. The purified PCR product forms a template on which the SNaPshot reaction is performed. During the SNaPshot reaction the 3´ end of the extension primer anneals to the nucleotide adjacent to the PSV resulting in amplification of the PSV site. The SNaPshot master mix contains fluorescently labelled ddNTPs producing a different colour for each allele, T red, C, yellow, G blue and A green. During the reaction only the nucleotide of interest is amplified producing a single peak for pseudohomozygous samples and two peaks of different colours for pseudoheterozygous samples. Following

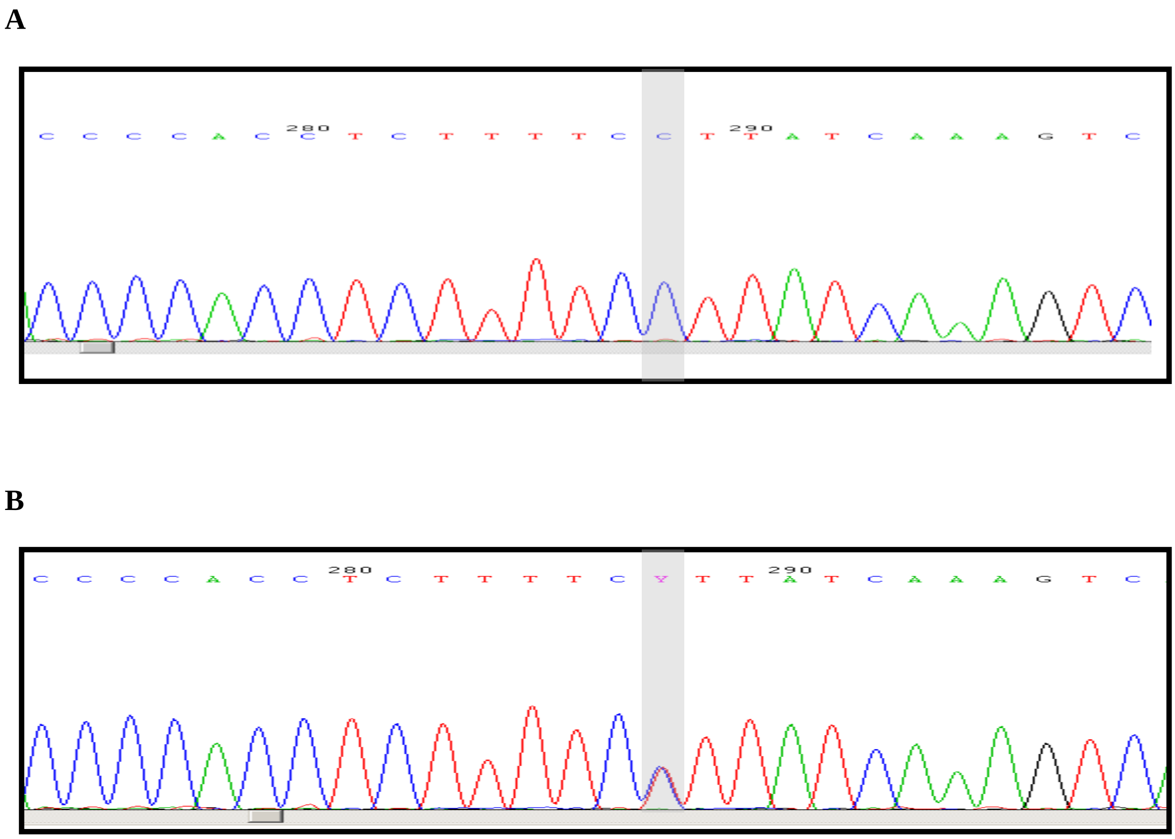


Figure 2.1 Sequence elctropherograms

a) Electropherogram demonstration a pesudohomozygous PSV.

b) Electropherogram representing a pseudoheterozygous PSV. In this study standard IUB codes will be used to identify PSV sites.

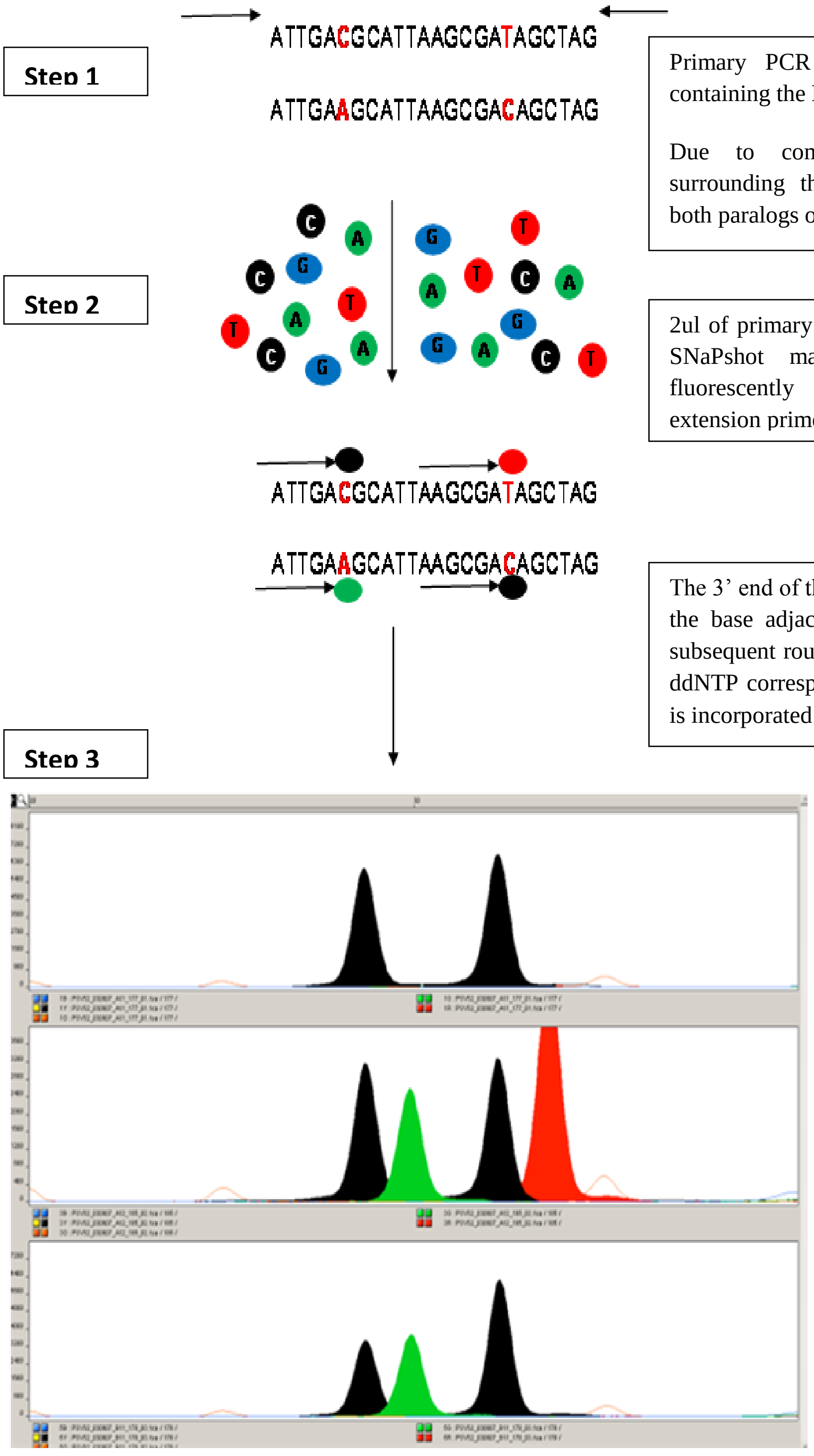
dephosphorylation of unincorporated ddNTPs size and peak colour can be detected by running products on an ABI 3130xl against the GeneScan™ 120 LIZ™ Size Standard which contains nine single-stranded fragments of: 15, 20, 25, 35, 50, 62, 80, 110 and 120 nucleotides labelled with the orange dye LIZ (Applied Biosystems).

2.2.8.1, Primary PCR

PCR primers were designed to amplify regions encompassing 1-2 PSVs based on the reference sequence alignment. PCRs were multiplexed allowing the amplification of multiple PSV sites in one reaction; for samples where multiplex PCR was unsuccessful separate PCRs were carried out and the products pooled. Products were purified using SAP and ExoI as described in section 2.2.6.

2.2.8.2, SNaPshot reaction

2µl of purified PCR template was added to 5µl SNaPshot master mix (Applied Biosystems), 1µl dH_20 and 2µl of SNaPshot primer mix with each primer at a final concentration of 0.5-3 µM forming a final volume of 10µl. During this reaction the extension primer binds to the complementary PCR template in the presence of fluorescently labelled ddNTPs and Amplitaq gold DNA polymerase which are contained within the SNaPshot master mix. The polymerase extends the primer by one nucleotide adding a single ddNTP to its 3´end. Each ddNTP is labeled with a fluorescent dye which is incorporated into the product according the base present in the template (Figure 2.2).



Primary PCR amplification of a region containing the PSVs of interest.

Due to complete sequence homology surrounding the PSVs co-amplification of both paralogs occurs during PCR.

2ul of primary PCR product is added to 5ul SNaPshot master mix which contains fluorescently labelled ddNTPs and an extension primer for each PSV.

The 3’ end of the extension primer anneals to the base adjacent to the PSV site. During subsequent rounds of thermocycling a single ddNTP corresponding to the base of interest is incorporated into the prod duct

2ul of product is run on an ABI 3130xl genetic analyser against GeneScan™ 120 LIZ™ Size Standard

For pseudohomozygous samples a single coloured peak is produced on the elctropherogram.

For pseudoheterozygous samples two peaks of different colours are produced on the electropherogram.

Figure 2.2: Workflow for the SNaPshot minisequencing assay

2.2.8.3, SNaPshot reaction conditions

Reaction conditions were generally 96°C 10s, 50°C 30s, 60°C 60s for 39 cycles although cycle number varies where stated.

2.2.8.4, ddNTP dephosphorylation

Unincorporated ddNTPs were dephosphorylated using 1U SAP (Amersham) incubated at 37°C for 1 hour, followed by inactivation at 80°C for 5 minutes.

2.2.8.5, Sample analysis

2µl of product were added to 10µl formamide containing 0.02µl of fluorescently labelled GeneScan™ 120 LIZ™ Size Standard (Applied Biosystems). Products were denatured by heating to 95°C for 5 mins and run on an ABI 3130xl capillary electrophoresis apparatus (Applied Biosystems) using the parameters detailed in supplementary table S2.2c.

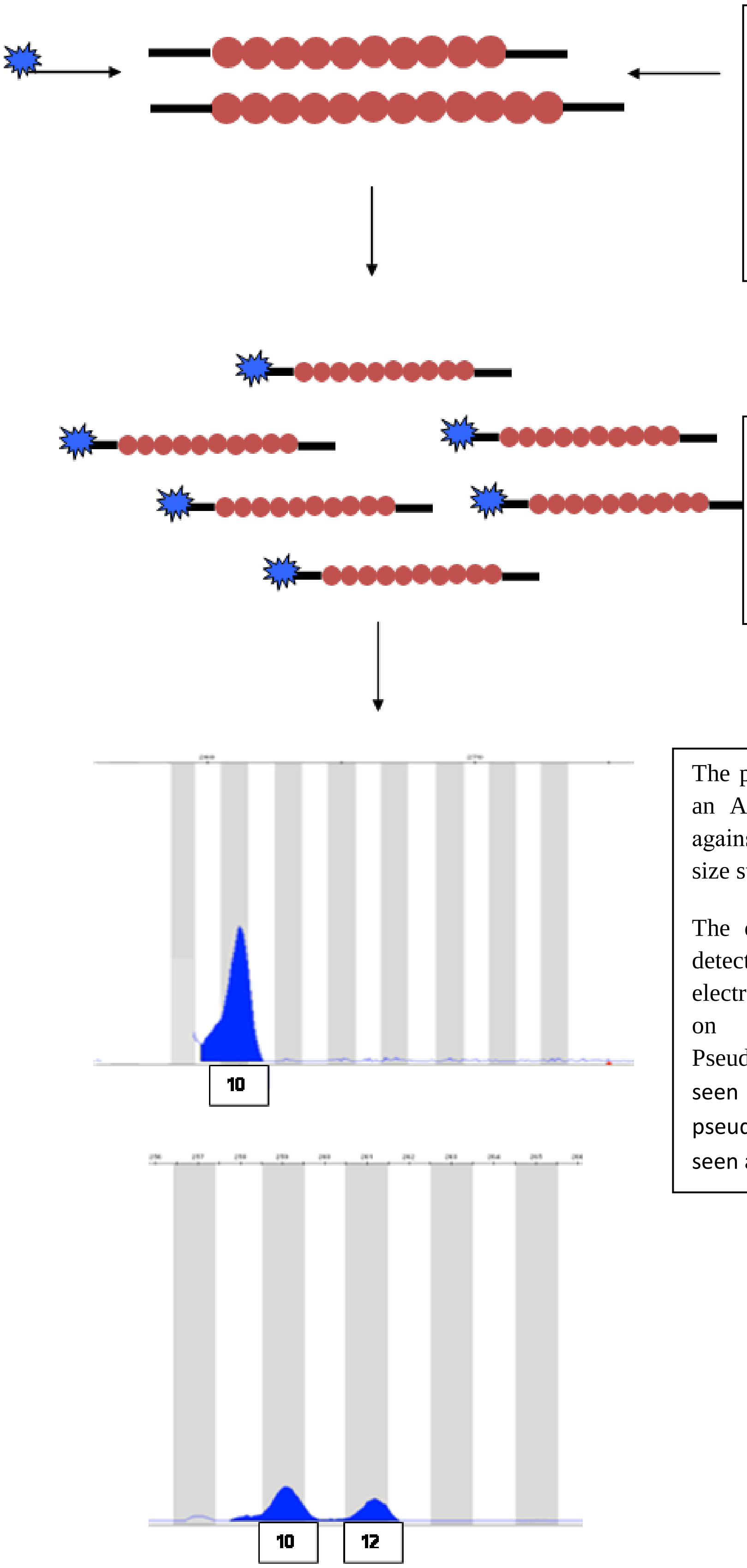
Analysis of fluorescently labelled products was carried out using Gene Scan v3.1 software (Applied Biosystems). Pseudohomozygous samples are identified by a single peak while pseudoheterozygous samples are identified by two peaks of different colours.

2.2.9, Overview of microsatellite typing

This is a one-step assay in which primers are designed to amplify a 100 – 400bp region flanking each microsatellite. The 5´ end of each forward primer is labeled with a fluorescent dye which allows fragment detection during capillary electrophoresis. During PCR the fluorescent dye is incorporated into the PCR product, which is detected after laser excitation as a peak on the electropherogram. Following laser excitation primers labeled with FAM, emit a wavelength of 518 (nm) producing blue peaks while primers labeled with HEX emit a wavelength of 556 (nm) producing green peaks. As microsatellites within palindromic repeats are duplicated, both copies are amplified during PCR producing either pseudohomozygous results which are seen as a single peak, or pseudoheterozygous results seen as two peaks. The sizes of the amplification products and dyes used can be chosen to allow multiplex microsatellite typing (Figure 2.3). Following PCR the products are diluted 1/50 with dH_2O and 2µl of product is added to 10µl of formamide containing 0.02µl of fluorescently labelled GeneScan™ 500 LIZ™ Size Standard (Applied Biosystems) which contains 16 single-stranded fragment of: 35, 50, 75, 100, 139, 150, 160, 200, 250, 300, 340, 350, 400, 450, 490 and 500 nucleotides labelled with the orange dye LIZ. Products are denatured by heating at 95°C for 5 minutes and run on an ABI 3130xl using parameters detailed in supplementary table S2.2d.

2.2.9.1, Microsatellite analysis

Products were analyzed using GeneMapper v 4 software (Applied Biosystems). Repeat number corresponding to a given fragment size for each microsatellite was determined by sequencing a pseudohomozygous sample using Big Dye v1.1 as described in Section 2.2.7.2.



Primary PCR amplification of a 100-400bp region flanking the microsatellite. Due to complete sequence homology surrounding the microsatellites amplification of both paralogs occurs during PCR.

The 5’ end of the forward primer is labelled with a fluorescent dye which is incorporated into the PCR product during subsequent rounds of PCR.

The product is diluted and run on an ABI 3130xl genetic analyzer against GeneScan™ 500 LIZ™ size standard.

The dye in the PCR product is detected during capillary electrophoresis producing a peak on the electropherogram. Pseudohomozygous samples are seen as a single peak while pseudoheterozygous samples are seen as two peaks

Sequencing of a pseudohomozygous sample is performed to determine the microsatellite repeat number for each peak.

Figure 2.3; Workflow for typing duplicated microsatellites

2.2.10, The Y phylogeny

In this study the Y phylogeny has been modified to be consistent with Karafet et al. (Figure, 2.4) in particular for the branch patterns leading to haplogroups B and C, and the coancestry of haplogroups I and J. However, the Karafet tree has not been reproduced in its entirety, as many of the markers and haplogroups shown there were not typed in this study.

2.3, Bioinformatic Analysis

The bioinformatic tools detailed in table 2.1 were used where stated in this thesis. A detailed description of these tools and their uses in the study of gene conversion is included in the introduction to Chapter 4.

2.3.1, The reference sequences

Human and chimpanzee Y- and X-chromosomal reference sequences were obtained from the UCSC genome browser while gorilla sequence (X chromosome only) was obtained by performing BLAST searches of the chimpanzee sequence against the gorilla trace archive. To establish prior existence in the human-chimpanzee common ancestor, complete orthologs were sought in the chimpanzee genome. Human and chimpanzee sequences were obtained from the March 2006 version of the reference sequence available from UCSC.

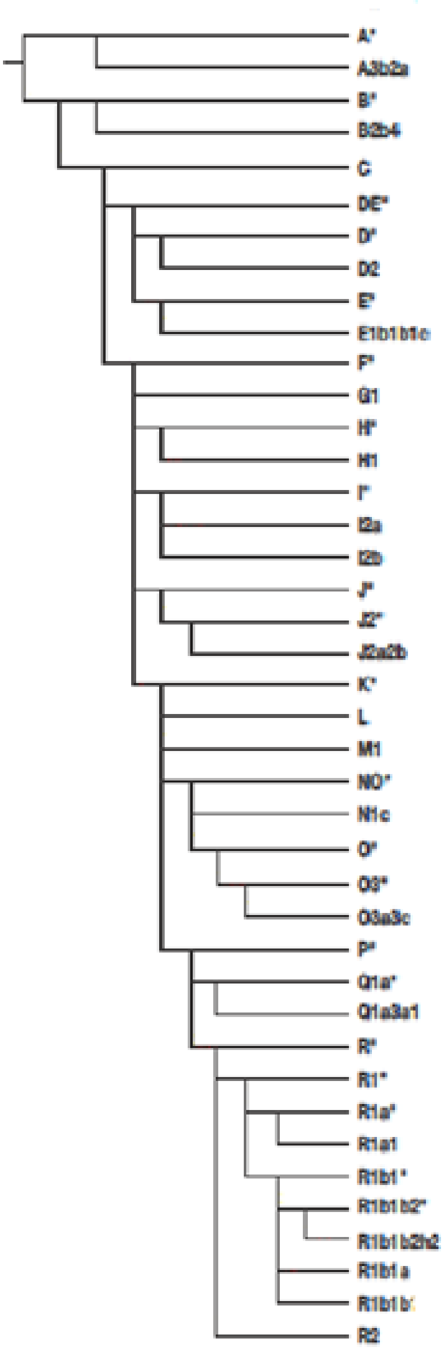


Figure 2.4. The Y phylogeny
The Y phylogeny used in this study is based on that of Karafet (2009). Branch patterns leading to haplogroups B and C, and the coancestry of haplogroups I and J are consistant with that of Karafet. However, many of the markers and haplogroups shown in the Karafet phylogeny were not typed in this study therefore the tree has been adapted to show the haplogroups typed in this study

2.3.2, Identification of PSVs

Sequences were aligned using clustalW (http://www.ebi.ac.uk/Tools/clustalw) or VISTA lagan http://genome.lbl.gov/vista/lagan/submit.shtm and PSVs identified from sequence alignments. Sequence aligments using clustalW were performed using default settings while the output for VISTA lagan alignments was set to identify regions which exhibit 95-100% similarity between sequences.

2.3.3, Identification of microsatellites

To detect microsatellites which are not variable within the reference sequence alignment but may be polymorphic in other chromosomes, sequences were analyzed using Tandem Repeats Finder software (Benson 1999) (http://tandem.bu.edu/trf/trf.html) using default settings.

2.3.4, Repeat Masker

As the GC content of sequences may be influenced by insertions such as Alus and LINEs all sequences were repeat-masked using the repeatmasker software (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) using default settings. Regions containing repeat elements were removed befor analysis of GC content was carried out.

2.3.5, Network construction

A network can be constructed based on microsatellite data for a single or multiple mirosatellite markers. Networks were constructed using Network 4 (Bandelt et al. 1995), and the Network Publisher software using the median joining option. This

Table 2.1: Bioinformatic tools

Program	**URL**	**Reference**
UCSC database	http://genome.ucsc.edu/	(Haussler et al. 2002)
The Watson and Venter sequences	http://jimwatsonsequence.cshl.edu/cgperl/gbrowse/cvsequence	(Levy et al. 2007; Wheeler et al. 2008)
Yanhuang sequence	http://yh.genomics.org.cn/search.jsp	(Zhang et al. 2008)
clustalW	http://www.ebi.ac.uk/Tools/clustalw/	(Higgins 2003)
VISTA lagan	http://genome.lbl.gov/vista/lagan/submit.shtml	(Brudno et al. 2003)
NCBI Blast	http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome	(Schwartz et al. 2003)
tandem repeats finder	http://tandem.bu.edu/trf/trf.submit.options.html	(Benson 1999)
splitsTree4	http://www.splitstree.org/	(Huson 1998)
Repeat masker	http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker	Smit &. Green unpublished data
Network	http://www.fluxus-engineering.com/nwpub.htm	(Bandelt, Forster, and Röhl 1999)
DnaSP	http://www.ub.edu/dnasp/	(Rozas and Rozas 1999)
Chi square and Fisher exact test	http://www.graphpad.com/quickcalcs	Smit & Green unpublished data

programme constructs a network by calculating the number of mutational differences between different Y chromosome haplotypes. Chromosomes sharing a particular haplotype are placed into a node the size of which is proportional to the number of chromosomes which carry the haplotype. Nodes are joined by a line which represents the number of mutational differences between each node, the length of the line being proportional to the number of mutations separating each haplotype.

A weighting scheme was employed for each microsatellite as described by (Qamar et al. 2002) with specific weights assigned to each microsatellite based on the variance observed between chromosomes within the sample set. The following weights were employed: a weight of 5 was used for variance of 0 - 0.09, weight 4 for variance of 0 - 1-0.19; weight 3 for variance of 0.2 - 0.49, weight 2 for variance of 0.5 - 0.99, and weight 1 for variance 1.00.

2.3.6, Use of SplitsTree to create phylogenetic split networks

Phylogenetic split networks (Huson 1998) were constructed using SplitsTree4 which can be downloaded from http://www-ab.informatik.uni-tuebingen.de/software/splitstree4/welcome.html. SplitsTree is a program used for analyzing and visualizing evolutionary data. Using split decomposition, evolutionary data are conically decomposed into a sum of weakly compatible splits which are represented as a single graph (Huson 1998) or phylogenetic networks. Within these phylogenetic networks the evolutionary history of a set of taxa (such as DNA sequences) is represented by a non-reticulated phylogenetic tree. This works well when non-recombining sequences are diverging in a simple way without any exchange through processes such as recombination or gene conversion. However, in more complex evolutionary scenarios, phylogenetic networks are useful, because they allow reticulate events such as recurrent mutation or gene conversion to be visualized.

In figure 2.5, SplitsTree 4 was used to create phylogenetic split networks, via the ‘‘NeighborNet’’ method and the ‘‘uncorrectedP’’ distance. Within the networks the lengths of edges represent the proportion of sites at which sequences differ, as indicated by a scale-bar. A simple mode of evolution without conversion will produce a simple non-reticulated network (Figure 2.5a); when there has been a history of gene conversion, this is reflected in a reticulation(s) (Figure 2.5b).

2.4, Statistical analysis

2.4.1, Divergence calculations

Divergence calculations were performed by loading the MFA files available from the VISTA output into the DnaSP software (http://www.ub.es/dnasp/). Polymorphism calculation was carried out to determine the number of identical nucleotides and the number of variable sites. From this the percentage sequence divergence was calculated.

2.4.2, Chi square

To determine whether gene conversion events are significantly conservative of the ancestral state, the P-value was determined using the chi square test available from http://www.graphpad.com/quickcalcs/chisquared1.cfm.

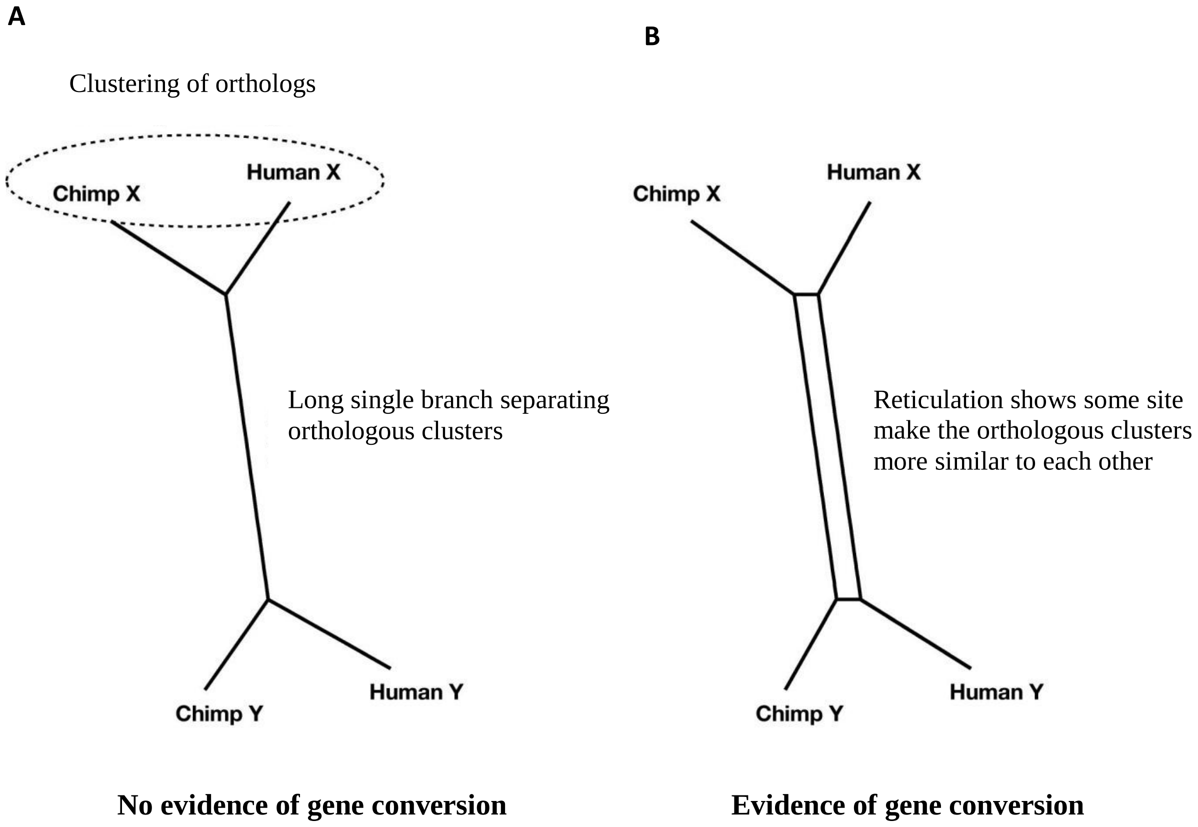

A
B
Clustering of orthologs
Chimp X
Human X
Long single branch separating orthologous clusters
Chimp Y
Human Y
No evidence of gene conversion
Chimp X
Human X
Reticulation shows some site make the orthologous clusters more similar to each other
Chimp Y
Human Y
Evidence of gene conversion

Figure 2.5. Phylogenetic split networks and evidence of gene conversion.

a) Network constructed from hypothetical X- and Y-linked paralogs in human and chimpanzee. The non-reticulated structure shows no evidence for gene conversion.
b) The reticulated structure represents a more complex history which could include gene conversion, or recurrent mutation.

2.4.3, Fisher exact test

To determine the significance of sequence divergence and whether GC-content is significantly different between duplicated and non-duplicated regions of the Y chromosome the P-value was determined using a two-tailed Fisher exact test available from http://www.graphpad.com/quickcalcs/contingency1.cfm.

2.5, Work carried out by others

Sequencing reactions using Big Dye v3.1 were carried out by Dr Sharad Mistry and Joan Sutherland at the Protein and Nucleic Acid Chemistry Laboratory (PNACL) at the University of Leicester.

Chapter 3: The reference sequences

This study relied on the available human and chimpanzee reference sequences to detect PSVs and to gain an understanding of how paralogs have evolved since speciation. Reference sequences are valuable for detecting PSVs without the need for extensive resequencing of a region; however, identifying PSVs from an individual sequence creates the problem of ascertainment bias. Analysis of a single reference sequence will only identify sites which are variable within the sequence and other PSVs which have been homogenised through gene conversion will not be identified. Some PSVs may also be 'private' to the individual chromosome and will add no additional information to the study of gene conversion. Analysis of a single chromosome sequence also poses other problems in addition to that of ascertainment bias. Chromosomal rearrangements such as inversions and deletions are common on the Y chromosome many of which do not affect male fertility (Kehrer-Sawatzki and Cooper 2007; Jobling 2008). Such rearrangements could have occurred in the reference sequence chromosome and this adds an additional problem as it could potentially make the reference sequence atypical of the majority of Y chromosomes, and therefore a poor general model.

Chapter 4 of this thesis relies on the analysis of the Y chromosome reference sequence to identify regions where gene conversion may be occurring. When this study was commenced, only the NCBI build 36.1 Y-chromosomal reference sequence (Lander et al. 2001) and the Celera sequence (Venter et al. 2001) were publicly available. With advances in sequencing technologies the number of available sequences is increasing, with four additional sequences being published within the past four years (Levy et al. 2007; Wheeler et al. 2008; Zhang et al. 2008; Blanchard 2009). These sequences potentially allow the identification of additional PSVs and may help overcome the ascertainment bias associated with identifying PSVs from an individual reference sequence. However, before these sequences can be used they must be assessed to see how reliable they are and whether they will be of further use in the study of gene conversion.

3.1, Chapter aims

Due to the potential variability of the human Y chromosome, it is useful to have a number of sequences for comparison. This Chapter will evaluate all known human and non-human primate Y chromosome sequences which were publicly available in April 2010 in order to determine the reliability and identify sequences which may be from different Y chromosome haplogroups.

As data for this study were obtained between October 2006 and May 2009, not all of the sequences discussed in this Chapter have been used in this study. However: the additional Y chromosome sequences which were published after completion of this study will be discussed to determine whether they may of further use in future studies of gene conversion.

3.2, The human Y chromosome reference sequences

In October 2006 when this study was first commenced, only two Y chromosome sequences were available (Skaletsky et al. 2003) and over the past four years four additional sequences have been made publicly available (Levy et al. 2007; Wheeler et al. 2008; Zhang et al. 2008; Blanchard 2009). This section will discuss all seven available human sequences and determine their reliability and whether they will be of additional value in the study of gene conversion.

3.2.1, The human reference assembly

The main reference sequence for the human genome is available from the UCSC genome browser (http://genome.ucsc.edu) and was produced by the International Human Genome Sequencing Consortium (Lander et al. 2001). The Y chromosome sequence available from the UCSC genome browser was produced by Skaletsky et al. (2003) which is mainly based on a single male donor.

The Y chromosome sequence was obtained by sequencing a tiling path of 220 bacterial artificial chromosome (BAC) clones, each containing a portion of the MSY from a single male. In total the sequence covered approximately 23Mb including 8Mb of sequence from Yp and 14.5Mb of sequence from Yq. Three gaps remain in the final sequence, two of which are approximately 50-kb in size and complete coverage of a 0.7Mb tandem array on Yp was not achieved. The third gap corresponds to the centromere of the chromosome. It is estimated that approximately 97% of the MSY euchromatin has been covered with approximately 60% of the euchromatin being sequenced in two independent BAC clones. The error rate was determined to be 1 nucleotide per 10^5 bases of sequence. To confirm the organization of MSY sequences, PCR amplification of the inner and outer boundaries of all palindromes in ten men from diverse Y chromosome haplogroups was carried out. This confirmed that each palindrome boundary is present in the majority of human Y chromosomes.

Analysis of Y chromosome binary markers for this sequence shows it to carry the derived allele for M65 and therefore represents a chromosome from haplogroup R1b1b2b; However, the *AZFa* region is known to be from a haplogroup G chromosome (Jobling 2008). In the remainder of this thesis “the reference sequence” will refer the sequence produced by Skaletsky et al. (2003) which is available from the UCSC genome browser.

In February 2009 the human reference sequence was updated with version GRCh37 (NCBI Build 37.1) which was produced by the Genome Reference Consortium. Analysis carried out in Chapter 4 uses the March 2006 version (NCBI build 36.1) of the reference sequence as the 2009 version was not available when the study was commenced. The 2009 version NCBI Build 37.1 has replaced build 36.1 in NCBI BLAST and ENSEMBL; however, build 36.1 can still be obtained from the UCSC genome browser.

3.2.2, The Celera assembly

The Celera sequence was produced in 2001 (Venter et al. 2001) and was the second sequence publicly available when this study was commenced. Celera produced 14.8 billion nucleotides of sequence generated from over 27 million shotgun sequence reads, producing 5.11 times coverage of the genome. Sequence was generated through sequencing of plasmid clones made from five individuals, (two males and three females) using the whole genome shotgun sequencing method. Reads had an average length of 543bp which have been organized into pairs by virtue of end sequencing 2-kb, 10-kb and 50-kb inserts from shot gun clone libraries (Istrail et al. 2004).

Two assembly strategies were employed - a whole genome assembly and a regional chromosome assembly - which combine sequence data from Celera and the publicly funded genome sequence. The resulting sequence effectively covered the euchromatic regions of the chromosome and filled gaps in the NCBI build 34 reference sequence.

Comparison of the Celera sequence (Venter et al. 2001) with the Y chromosome reference sequence (Skaletsky et al. 2003) reveals that only data for one arm of each palindrome is available which appears to be due to a mis-assembly of material from both the proximal and distal palindrome arms. In many cases the Y chromosome sequence appears to be erroneously labeled as the X chromosome - this was determined by performing BLAST searches of sections of the Y chromosome which are known not to have X homology. Due to the presence of both male and female donors in Celera's shotgun sequence, coverage of the X and Y chromosomes is reported to be lower than that of the other chromosomes resulting in a lower-quality assembly for the sex chromosomes (Istrail et al. 2004) and this may explain the observed problems with obtaining data for the Y chromosome sequence.

Analysis of binary markers shows the Celera sequence to carry the derived state for marker M65 and this chromosome appears to belong to the haplogroup R1b1b2b: however, as two males were sequenced it is possible that that this sequence represents a combination of two Y chromosome haplogroups. As this sequence appears to belong to the same haplogroup as the reference sequence it offers only limited additional information in this study and given the misassembly of palindrome arms, no additional PSVs could be identified. For these reasons this sequence will not be included in further analysis.

3.2.3, Craig Venter assembly

Craig Venter's Y chromosome sequence was sequenced jointly by The Craig Venter Institute, The Toronto Hospital for Sick Children, and The University of California and was made public in 2007 (Levy et al. 2007). Sequence was produced from ~32 million random DNA fragments, sequenced by Sanger dideoxy technology, and comprises 2,810 Mb of contiguous sequence with approximately 7.5-fold coverage for any given diploid region (Levy et al. 2007). Sequence data can be obtained from the http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/cvsequence/ website, which contains a browser where sequence co-ordinates can be entered to obtain data from specific chromosomal regions. The browser also offers additional information such as SNPs and GC content for any given region.

Analysis of binary markers shows this sequence to carry the derived state for the SNP M65 and therefore represents an individual from haplogroup R1b1b2b, and will only offer limited additional information in this study. From preliminary analysis this sequence appears to be of further use in this study and will be referred to as the "CV" sequence in the remainder of this thesis.

3.2.4, James Watson assembly

The James Watson genome was sequenced jointly by The Baylor College of Medicine Genome Sequencing Center, 454 Life Sciences Technology, and The Rothberg Institute. This sequence was made public in 2008 (Wheeler et al. 2008) and is also available from the http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/cvsequence/.website website.

Sequence was produced from 234 runs on a Genome Sequencer FLX instrument (454 technology), producing over 105 million bases per run. In total 93 million reads each of approximately 250 nucleotides were obtained, producing 7.4-fold coverage of the diploid genome. Sequences were mapped on to NCBI build 36 of the human reference sequence, by alignment using Basic Local Alignment Search Tool (BLAST)-like alignment tool (BLAT). All were realigned to local genome segments using Cross match software.

Analysis of binary markers shows the sequence to carry the derived state for the SNP M65 and therefore also represents an individual from haplogroup R1b1b2b and will offer only limited additional information in this study. However, this sequence may also be of further use in the study of gene conversion and will be referred to as the "JW" sequence in the remainder of this thesis.

3.2.5, Yoruban human genome NA18507

A sequence obtained from a Yoruban male from the HapMap panel was published in 2009 (Blanchard 2009). Sequence data for NA18507 are available from the NCBI short read archive, accession SRA000271, and were obtained using a combination of mate-paired libraries and fragment libraries with the Applied Biosystems SOLiD System analyzer. From this analysis 76.53Gb of paired reads were generated and were aligned to the NCBI build 36 reference sequence. In total 17.9 fold coverage of the diploid genome was produced with accuracy reported to be >99.99%.

At the time of writing data were not available in a format where specific sequence information was easily obtained; however, it is reported that a Resembl viewer is currently under construction (Blanchard 2009). As this sequence was only made public towards the end of this study, the data were not in a format that could readily be used. Data available from the International Hapmap project website (http://hapmap.ncbi.nlm.nih.gov) reveals this individual to represent a haplogroup E3a7 chromosome, which is likely to be of further use in the identification of additional PSVs.

3.2.6, The YangHuang genome

Sequence data from a Han Chinese male were published in November 2008 as part of the YangHuang Project, which aims to sequence 100 Chinese individuals in 3 years (Zhang et al. 2008). The YangHuang genome sequence was assembled from 3.3 billion reads generated using an Illumina Genome Analyzer. In total 102.9Gbp were mapped against the NCBI Build 36 version of the human reference sequence using the self-developed software SOAP (Short Oligonucleotide Alignment Program). A 36-fold average diploid sequence coverage was achieved which is estimated to cover 99.97% of the genome. Sequence data have been deposited in the EBI/NCBI Short Read Archive (accession number ERA000005).

Sequence data are available from http://yh.genomics.org.cn/download.jsp, where the sequence has been put into FASTA format and can be downloaded for each chromosome. Using the mapview function in the browser, regions of different chromosomes can be downloaded by entering sequence co-ordinates which allows large segments of sequence to be obtained. An ideogram shows the location of sequences on the chromosome and also shows additional information such as genes and SNPs as well as HapMap data. Mapview is similar to the UCSC browser but currently does not provide as much information as the UCSC browser. BLAST searches of a sequence can also be carried out against the YangHuang genome using the BLAST function in the browser. This is useful when looking at short read lengths but can only align approximately 1kb at a time, making it difficult to obtain sequence for large regions such as Y-chromosomal palindromes. This function is quite slow and mapview is much more efficient for obtaining large segments of sequence.

Analysis of binary markers reveals this sequence to carry the derived allele for M122 representing a haplogroup O3 chromosome which may potentially identify additional PSVs which will be of further use in the study of gene conversion. The YangHuang sequence will be referred to as the "Yh" sequence in the remainder of this study.

3.2.7, The DFNY sequence

The DFNY sequence was produced in 2009 (Xue et al. 2009) by flowsorting of two closely related Y chromosomes DFNY1-66 and DFNY1-101. Paired-end libraries of 200 bp fragments were constructed, and 35 bp from each end were sequenced with Illumina (Solexa) technology. After quality control and removal of duplicate reads, 11x and 20x mapped coverage of the Y reference sequence was obtained from DFNY1-66 and DFNY1-101, respectively. Following application of the filter parameters an 11-fold and 20-fold coverage for each individual Y chromosome was achieved. The haplogroup for the two DFNY1 individuals was determined to be O3 by typing a standard set of Y-SNPs including M122 (Xue et al. 2009). This sequence was made available in 2009 after this study was complete; however, advance sequence for palindromes 6, 7 and 8 was kindly provided by Chris Tyler-Smith which enabled additional PSVs to be identified in these regions.

3.2.8, Determining the reliability of available human database sequences

From the initial analysis of the recently published sequences the CV, JW, Yh and DFNY sequences appear to be of further use in this study for identification of additional PSVs.

To assess the reliability of the CV, JW and Yh sequences, ten STS markers from across the Y chromosome were compared to the reference sequence obtained from the UCSC database. As this study carries out analysis of the ampliconic regions and x transposed region (XTR), five STS markers from the ampliconic regions and XTR were selected for analysis while an additional 5 from the X-degenerate regions were included to give an idea of the reliability of the sequence as a whole. All ten STS markers were shown to be present in all three sequences and complete sequence was obtained in all cases. Secondly sequences from ten 1-kb segments and three genes (*AMELY*, *PCDH11*, *VCY*) were obtained using chromosome co-ordinates from the UCSC database (Supplementary information S3.1. These regions were also present and complete in all three sequences. As DFNY sequence was only available for palindromes 6, 7 and 8, sequences were compared to the reference sequence and while no additional PSVs were identified in P7 four additional PSVs were identified in P8 and 13 in P6.

From this analysis it was determined that all four sequences were reliable enough to be included in sequence comparisons. Using the co-ordinates obtained from the UCSC database (Supplementary information S3.2), sequences were obtained for palindromes 6-8, IR1 and the XTR.

3.3, Non-human primate reference sequences

When studying gene conversion occurring between human Y chromosome paralogs, having sequence from a non-human primate available for comparisons is useful. Not only do comparisons offer information about how the human Y chromosome has evolved since speciation, they can also provide evidence of gene conversion by providing evidence of the ancestral state of human PSVs and to some extent allow the direction of conversion to be determined. Comparisons of the human and chimpanzee sequences to an independent primate species such as the gorilla or macaque can also identify interspecies sequence variants which are due to gene conversion occurring in either species.

Currently the gorilla and macaque sequences can be obtained from the NCBI trace archive; however, data are only available for females in each case, which only offers additional information relevant for the X/Y homologous regions analyzed in this study, and not the Y-specific regions. As the XTR has arisen through transposition of X chromosome sequence on to the Y chromosome (Page et al. 2009) following human and chimpanzee speciation, the gorilla sequence will provide additional information on the ancestral state of X chromosome variants and X/Y chromosome GSVs.

3.3.1, The chimpanzee reference sequence

As the gorilla and macaque genome projects offer no additional information on the study of gene conversion between Y chromosome paralogs, sequence comparisons can only be made with the chimpanzee reference sequence. The chimpanzee sequence is valuable in the study of gene conversion for several reasons, Firstly it can provide evidence of whether Y chromosome paralogs were present prior to speciation and also provides evidence of the ancestral state of human PSVs. Determining the presence of paralogs prior to speciation suggests that the low intraspecific sequence divergence observed between paralogs is not due to a recent duplication event in humans and suggests gene conversion has been occurring.

3.3.2, The March 2006 reference assembly

The March 2006 (panTro2) assembly available from the UCSC genome browser is based on the sequence of Hughes et al. (2005) and is the main chimpanzee reference sequence used in this study. This version of the chimpanzee reference sequence is not complete and this must be taken into account when it is being used for human and chimpanzee sequence comparisons. The March 2006 (panTro2) assembly comes from whole genome shotgun data derived primarily from a male *Pan troglodytes*, named 'Clint'. This assembly is estimated to cover approximately 97% of the chimpanzee genome. The chimpanzee Y chromosome sequence was assembled via alignment against the human genome at UCSC utilizing BLASTZ (Schwartz et al. 2003) and structures such as the centromeres were introduced based on their positions in the human sequence.

3.3.3, Sequence of Kuroki et al. 2006

The second chimpanzee sequence was published by Kuroki et al. in 2006 and is available through NCBI BLAST. Only part of the Y chromosome was sequenced in this study comprising approximately 12.7 Mb of the chimpanzee Y chromosome. This sequence was produced from isolated clones from the whole-genome BAC library and

Y-specific BAC and fosmid libraries originating from a single male chimpanzee named "Gon"**.** Four clone contigs were produced which cover 905,338 nucleotides of chimpanzee Yp and 10,074,253 nucleotides of chimpanzee Yq. The sequence was estimated to be 99.9998% accurate based on a 5.6 Mb segment of sequence. The regions sequenced correspond mainly to the X-degenerate region of the human Y chromosome and partly to the ampliconic regions of the chimpanzee Y chromosome and therefore not all of the regions included in this study are covered by this sequence.

3.3.4, The 2009 chimpanzee sequence

Since this project was carried out, a new and essentially complete chimpanzee sequence has been published. (Hughes et al. 2010) extended sequencing of the chimpanzee MSY using large insert BAC clones and the same BAC tiling path/shotgun approach employed by Skaletsky et al. (2003) for the human MSY. From this a tiling path of 219 BAC and 12 fosmid clones from across the chimpanzee MSY were assembled. All but 17 of the 230 BAC or fosmid clones were from one male. The resulting 25.8Mb of euchromatic sequence is formed from eight contigs, the largest of which spans 10.1Mb. The finished sequence is estimated to have an error rate of one nucleotide per Mb of sequence.

3.3.5, The reliability of the chimpanzee reference sequences

When this study was commenced only two chimpanzee Y chromosome sequences were available for comparison. The March 2006 (panTro2) chimpanzee Y chromosome sequence was claimed to be completely sequenced however there has been much debate as to the reliability of the available sequence. (Hughes et al. 2006; Tyler-Smith, Howe, and Santos 2006) have questioned its reliability and whether the sequence is truly complete. Analysis carried out in this study has revealed many gaps in the sequence while some regions of the human Y chromosome could not be found in chimpanzee. As the 2009 version of the chimpanzee sequence was not available when this study was carried out, this has limited the analysis of some regions of this study and for some regions human- chimpanzee sequence comparisons could not be performed.

The March 2006 reference sequence available from the UCSC database will be the main source of the chimpanzee sequence used in this study and will be referred to as “the chimpanzee reference sequence” in the remainder of this thesis. Where complete sequence could not be obtained from the March 2006 sequence the new 2009 version will be used. As this sequence is not currently in an easy obtainable format only regions where complete sequence from the March 2006 reference sequence could not be obtained will be re-analyzed using the new 2009 sequence.

Chapter 4: A bioinformatic exploration of Y chromosome paralogs

4.1, Introduction

The use of bioinformatics to explore regions which display high sequence similarity can give some insight as to which regions of the Y chromosome may be undergoing gene conversion. It can also distinguish paralogs which have high sequence similarity through gene conversion from those which have high similarity due to a recent duplication. A number of bioinformatic tools are available which are be able to indicate whether gene conversion is occurring within a region; however, only experimental analysis can definitively determine whether gene conversion is ongoing. This study of gene conversion will follow the example set by Rozen et al. (2003) to identify regions where sequence divergence between Y chromosome paralogs is significantly lower than non-duplicated regions. In this Chapter, the bioinformatic software used in this study will be described; however, other programmes available to study gene conversion which have not been used in this study will be discussed in Chapter 11 of this thesis.

4.1.1, The UCSC genome browser

The UCSC http://genome.ucsc.edu genome browser provides access to sequence data for over 40 different genomes. Sequence information can be obtained by carrying out searches based on information such as a chromosomal region, gene name or by using the BLAST-like alignment tool (BLAT). The information which is displayed in the browser window can be selected from a list of extensive settings and can include information such as SNPs, genes, conservation between species and segmental duplications.

In this study the UCSC browser is useful for identifying segmental duplications and regions of paralogy between the Y and other chromosomes. Coloured blocks represent duplications and the degree of sequence identity, with orange blocks representing 99% sequence similarity and grey representing less than 99% similarity. These regions can also be compared to the genomes of other species such as the chimpanzee by using the convert or BLAT functions.

Details of snPSVs and microsatellites from additional sources such as dbSNP are also available from the browser which provides an additional way to identify variant sites which are not variable within the reference sequence. The browser has recently been updated with a 2009 version of the human reference sequence; however, as this study was commenced before the 2009 sequence was made public, coordinates used throughout this thesis will refer to the March 2006 NCBI Build 36.1 version. The UCSC browser still provides a link to Build 36.1, which was produced by the International Human Genome Sequencing Consortium (Lander et al. 2001)

4.1.2, ClustalW

ClustalW is available from http://www.ebi.ac.uk/Tools/clustalw2/index.html and can align multiple sequences into one alignment which is useful for analysing sequences from multiple regions, individuals or species. Sequences must be in the FASTA format and also in the same orientation as the software does not reverse complement sequences. Output is given as a single sequence alignment showing regions of complete identity and also produces a phylogenetic network showing the relationship between sequences. In this study this allows PSVs between sequences to easily be identified and gives an indication of the degree of sequence similarity exhibited between sequences.

4.1.3, VISTA lagan

VISTA lagan is available from http://lagan.stanford.edu/lagan_web/index.shtml and can be used to carry out multiple sequence alignments. Unlike clustalW, sequences cannot be viewed as one single alignment and multiple output files are produced. A graphical output of the sequence alignment can be viewed, from which the degree of sequence identity and location of PSVs can easily be observed. In the output the pink shading represents the degree of similarity between the two sequences with the white regions representing sites where there is variation between sequences. Parameters can be adjusted to display different percentages of similarity between two sequences, and in this study the parameters have been set to display regions which exhibit between 95% and 100% sequence similarity. Unlike clustalW, VISTA lagan produces a number of different alignment files which can be imported into different software such as DnaSP (Rozas and Rozas 1999) and SplitsTree (Huson and Bryant 2006).

4.1.4, SplitsTree4

The SplitsTree4 software produces a graphical view of the relationship between sequences and is useful in this study as it can indicate graphically whether conversion is occurring. A number of different parameters can be adjusted which change the information displayed in the network. As divergence calculations are typically based on single nucleotide variations, in this study the parameters have been set to ignore gap sites. Additionally, parameters have been set to calculate Jukes Cantor distances which assumes that the rate of substitution is the same for each of the four nucleotides. Reticulations between sequences provide evidence of an ancestral relationship between sequences while the lengths of edges represent the proportion of sites at which sequences differ, as indicated by the scale bar.

4.2.5, DnaSP

In the study of gene conversion it is important to establish the existence of paralogs in the human genome prior to speciation, as high sequence similarity can be due to a recent duplication event as well as to long-term gene conversion. Low sequence divergence between paralogs which are known to predate speciation indicates that gene conversion may be occurring. Divergence calculations are most informative when comparing sequences between different species such as the human and chimpanzee which are known to have diverged from a common ancestor over 5MYA. While low divergence between paralogs which have arisen following speciation would be expected, low divergence between paralogs which are known to predate speciation is suggestive of gene conversion.

DnaSP has many functions that can be used for different types of analysis such as identifying polymorphic sites, calculating linkage disequilibrium and divergence between populations; however, the sequences must be the same length and gap sites aligned prior to use. In this study DnaSP has been used to identify the number of variations between sequences and to calculate sequence divergence between two paralogs which are known to predate human-chimpanzee speciation. Divergence calculations are useful in this study as low sequence divergence between paralogs which are known to predate speciation is suggestive of gene conversion. MFA files produced by VISTA lagan in which the sequences have been aligned to include gap sites can be imported into DnaSP and divergence between sequences can be determined.

4.1.6, Chapter aims

This Chapter will carry out bioinformatic analysis of Y chromosome paralogs which exhibit >98% sequence similarity but are not known to have previously been studied for evidence of gene conversion. Paralogs will be analyzed to determine whether high sequence similarity is more likely attributable to gene conversion or duplication. Paralogs which are shown to predate speciation and which are most likely to be undergoing gene conversion will be studied in more detail in the remainder of this thesis.

4.2 Results

4.2.1, Exploration of Yq Palindromes

Eight palindromes termed P1-P8 are situated on the long arm of the Y chromosome. These palindromes each consist of two duplicated “arms” which typically exhibit >99% sequence similarity and are separated by a non-duplicated spacer sequence. The presence of orthologous palindromes in the chimpanzee genome shows that at least five of these structures existed prior to speciation (Rozen et al. 2003), which occurred approximately 5MYA, and therefore the high sequence similarity cannot be attributed to more recent duplication event. Evidence of gene conversion has previously been reported between the *CDY* gene copies located within the arms of P1 (Rozen et al. 2003). This study also suggested that gene conversion is conservative of the ancestral state which has been suggested is to maintain the function of spermatogenic genes.

While gene conversion has been shown to occur in the larger palindromes, relatively little is known about the smaller palindromes. This section concentrates on the three smallest palindromes P6-P8. Gene conversion within these palindromes has not yet been studied, and their relatively small size makes them more tractable than the larger palindromes. Rozen et al. (2003) have previously shown that P6 – P8 exist in the chimpanzee genome indicating that these structures were present before the human and chimpanzee lineages diverged. Sequence divergence between human palindromes is reported to be <0.1% despite the structures having existed for over 5MY, which suggests that historical gene conversion has been occurring within these palindromes.

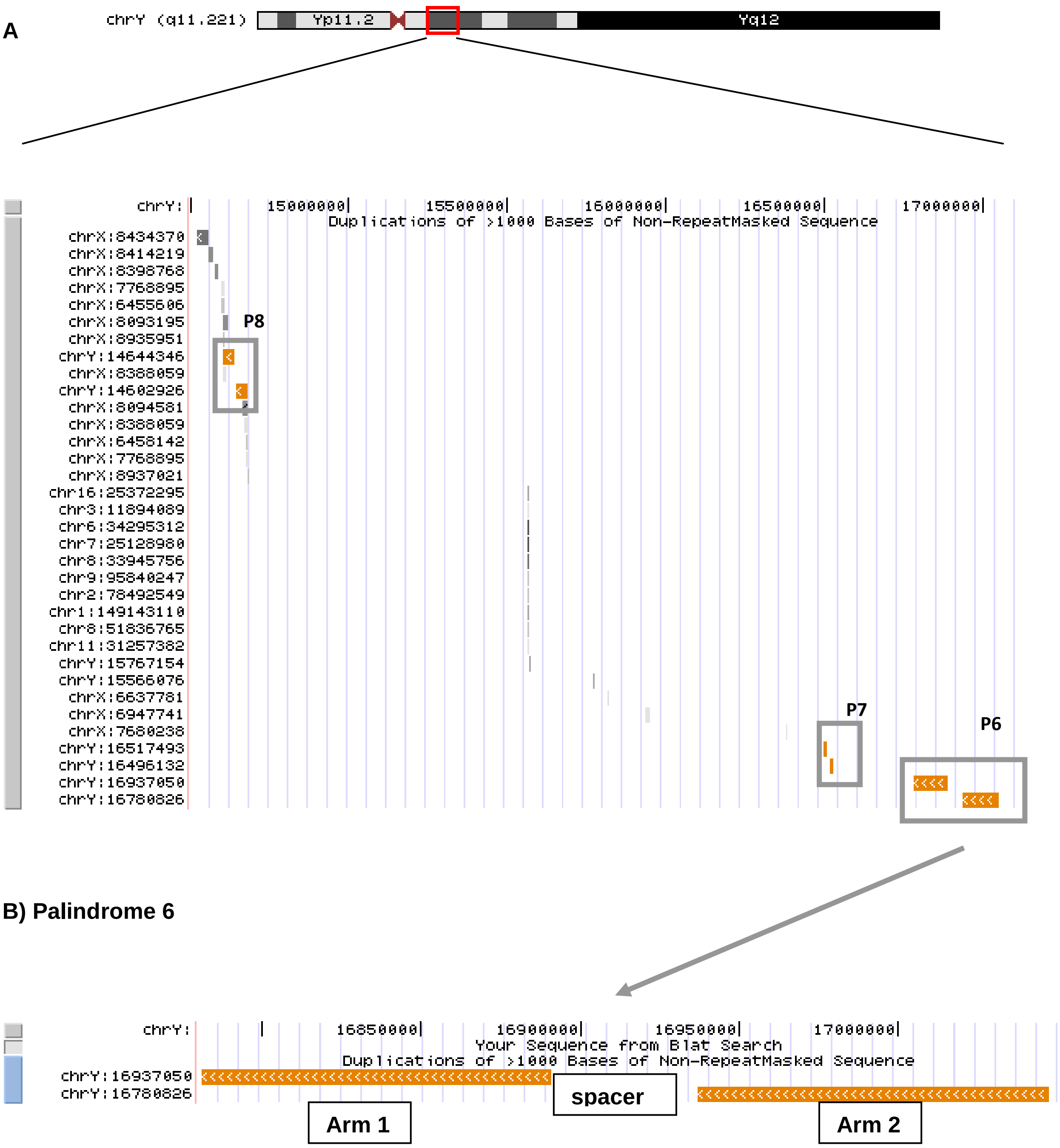


Figure 4.1: Parology features of the region containing palindromes 6-8

In the UCSC browser (A) the palindromes can be easily identified with the orange rectangles representing the duplicated palindrome arms which exhibit >99% similarity between sequences and the gap between the rectangles represents the non-duplicated spacer region (B).

4.2.1.1, Sequence comparison

Palindrome sequences were identified in the UCSC database using landmark STSs described by Skaletsky et al. (2003) (Figure 4.1). Sequence alignments were carried out using VISTA lagan and PSVs were identified (Figure 4.2). From sequence alignments three classes of variants were observed between palindrome arms; these were single nucleotide PSVs (snPSVs), insertions/deletions (indels), and microsatellite length variations (repeat unit size ≥2bp).

As previously reported by Rozen et al. (2003) a high degree of sequence similarity was observed between palindrome arms, with P7 and P8 displaying a higher degree of sequence similarity than P6. In total fifty PSVs were identified between the arms of P6: of these, twenty-eight consist of snPSVs, eleven of microsatellite length variations and the remainder of poly A-tail variations. In contrast, only five PSVs were identified between the arms of P7, two of which were microsatellite length variations, while the remaining three were snPSVs. P8 showed the lowest number of variations between arms with only two PSVs being identified. (supplementary table S4.1)

4.2.1.2, Sequence divergence

To gain further insight into the degree of similarity exhibited between palindrome arms, divergence calculations were performed for each palindrome using DnaSP. As is conventional, indels or microsatellite length variations were excluded in these calculations. P7 and P8 were both shown to display lower inter-arm sequence divergence (0.01%) than P6 (0.03%) and since all three palindromes are known to predate speciation this low divergence is suggestive of gene conversion in the human

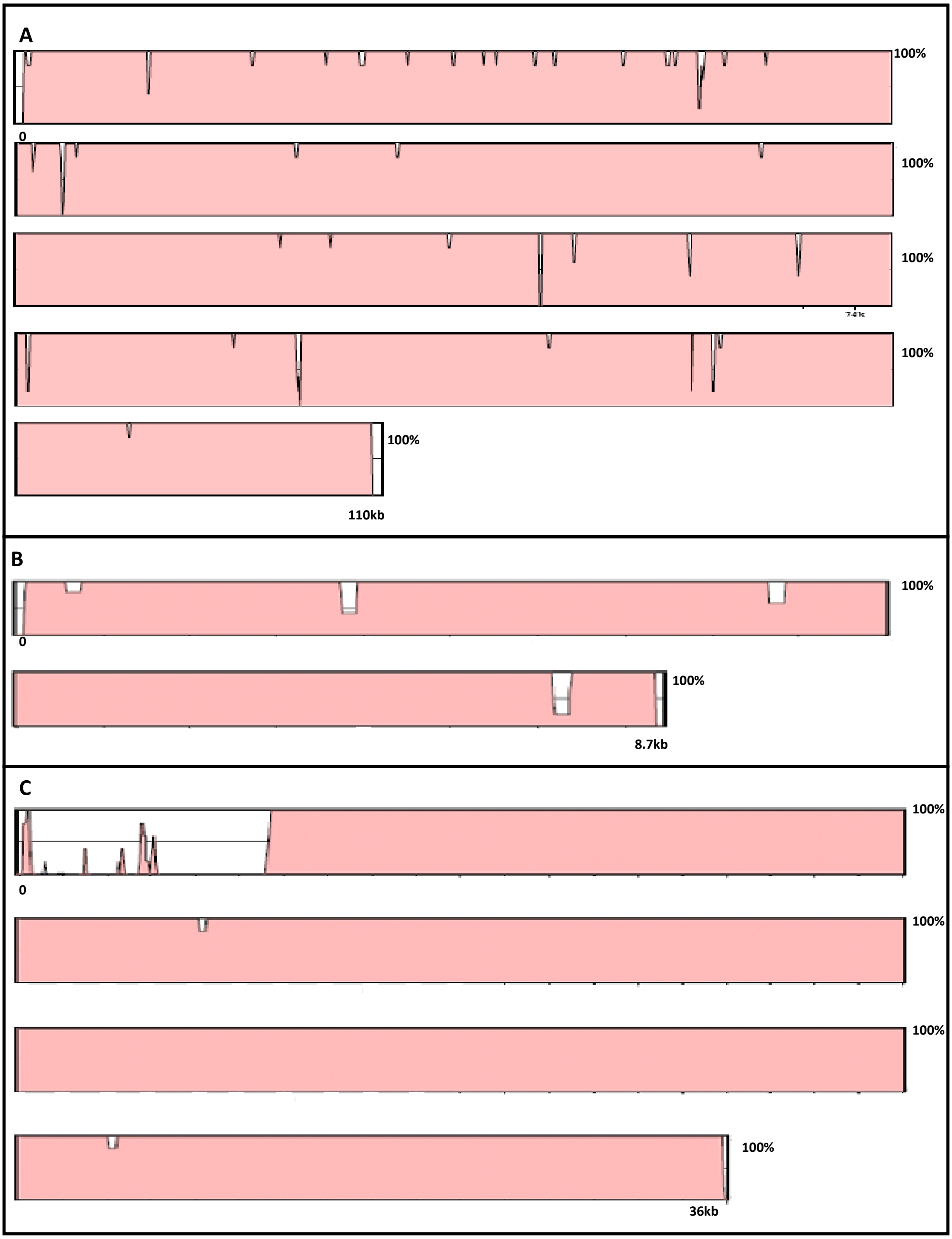


Figure 4.2: Inter-arm sequence similarity for palindromes 6- 8
VISTA lagan outputs representing regions of the palindromes which display >95% similarity between sequences (pink shading) and the location of paralogous sequence variants (PSVS (white gaps) From alignment of the reference sequence 50 PSVs were identified between the arms of P6 (A) 5 between arms of P7 (B) and 2 between the arms of P8 (C).

lineage. The higher sequence divergence and number of PSVs observed between the arms of P6 suggests that if gene conversion is occurring it is likely to be at a slower rate than for P7 and P8.

4.2.1.3, Human and chimpanzee sequence comparison

Each palindrome was located in the chimpanzee reference sequences (using the UCSC genome browser) and four-way sequence alignments were carried out using clustalW. Interspecies sequence comparisons were performed for each palindrome and interspecies divergence between palindrome arms and between the spacer sequences was calculated (Figure 4.3). At the time this study was commenced the chimpanzee reference sequence was not complete and only 10kb of sequence was available for P8. Where stated the 2009 version of the chimpanzee reference sequence was used.

From interspecies sequence comparisons three types of variant were observed between species; these were microsatellite length variations, indels and single nucleotide variations. As previously reported by Rozen et al. (2003), P7 was shown to exhibit the highest degree of sequence similarity between human and chimpanzee with only eighteen variants being identified. The most pronounced structural difference observed between human and chimpanzee P6 and P8 occurred at the outer boundaries. When human P6 was compared to its chimpanzee ortholog, the distal palindrome arm was observed to exhibit sequence similarity which extended beyond the outer boundary observed from the human sequence alignment alone. BLAST searches identified only one region in the human sequence which corresponds to the distal palindrome arm and suggests that a deletion may have occurred in the proximal arm which has shortened the palindrome in humans. Interestingly, the P8 sequence alignments also reveal a human-

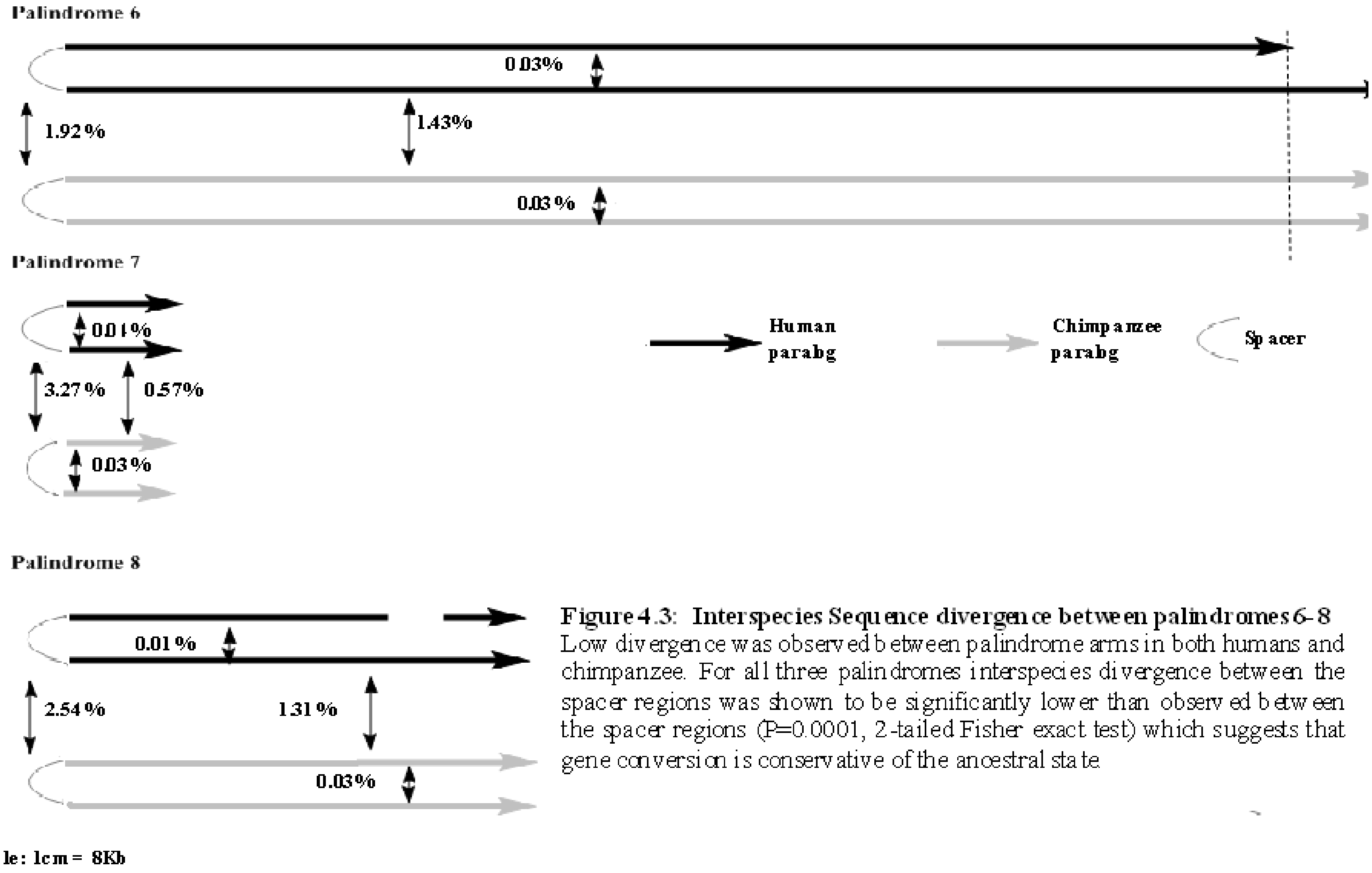


Figure 4.3: Interspecies Sequence divergence between palindromes 6-8 Low divergence was observed between palindrome arms in both humans and chimpanzee. For all three palindromes interspecies divergence between the spacer regions was shown to be significantly lower than observed between the spacer regions (P=0.0001, 2-tailed Fisher exact test) which suggests that gene conversion is conservative of the ancestral state.

chimpanzee difference near the outer boundary. A 270-bp retroviral insertion approximately 500bp from the outer boundary of the proximal arm of the human palindrome has led to the disruption of homology between human palindrome arms with the accumulation of 66 human-chimpanzee variants within this 500bp region.

4.2.1.4, Interspecies divergence

From comparison of the interspecies divergence between palindrome arms and the spacer regions Rozen et al. (2003) have previously suggested that gene conversion is conservative of the ancestral state. In this section similar comparisons for P6-P8 will be carried out: as only P8 contains genes, evidence of conservative gene conversion between the arms of non-genic P6 and P7 would suggest that conservative gene conversion is not limited to palindromes which contain genes. As previously reported by Rozen et al. (2003) interspecies divergence between human and chimpanzee P7 was very low at 0.57% while the divergence between the spacers was significantly higher higher at 3.20% (P=0.0001, 2-tailed Fisher exact test). Both P6 and P8 displayed higher interspecies sequence divergence between palindrome arms at 1.43% and 1.30% respectively, while divergence between the spacers was significantly higher at 1.92% for P6 and 2.54% for P8 (P=0.0001, 2 tailed Fisher exact test). This also suggests that gene conversion between the arms of P6 and P8 is conservative of the ancestral sequence.

4.2.1.5, Phylogenetic split network

Sequence analysis of P6-8 suggests that conservative gene conversion occurs between the arms of palindromes. Construction of a phylogenetic split network can provide a graphical view of the relationship between human and chimpanzee orthologs and can also indicate if gene conversion may be occurring. To gain further understanding of the evolutionary relationship between human and chimpanzee palindromes a network was constructed for palindromes 6 and 7 using the SplitsTree4 software (Figure 4.4).

Within the network the presence of a reticulation at the centre of the network can indicate a history of gene conversion, while the length of the edges represents the proportion of sites at which the sequences differ – as indicated by the scale bar. Interpretation of a split network can be drawn from the basic assumptions that gene conversion makes paralogs more similar while creating divergence between orthologs. For P6 the network and divergence calculations appear to support the basic assumptions of gene conversion whereby clustering is observed between paralogs at each end of the network while there is a greater distance between orthologs than between paralogs (Figure 4.4a). This suggests that since speciation the paralogs of P6 have become more similar to each other but more diverged from the ortholog and suggests that gene conversion is occurring. The reticulated structure of the network suggests that gene conversion is occurring between sequences. This network combined with low inter-arm divergence in each species and low interspecies divergence suggests that gene conversion has occurred in the human and chimpanzee lineages.

Interestingly, P7 produces a different network to that observed from P6 (Figure 4.4b). In this network clustering in observed between orthologs as opposed to paralogs which implies that the human and chimpanzee orthologous palindrome arms are more similar to each other while the reticulated structure indicates that conversion has occurred. This

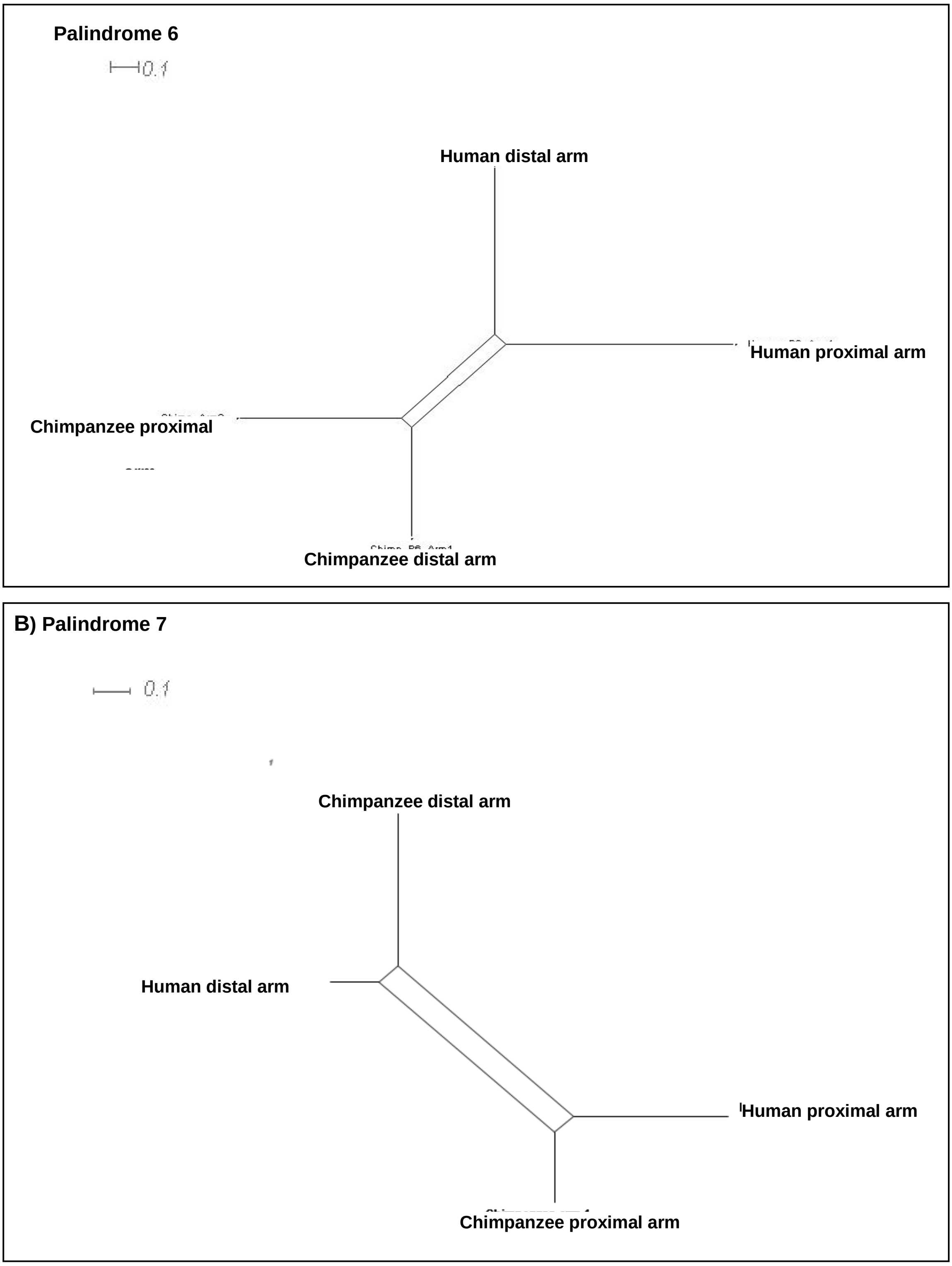


Figure 4.4: Phylogenetic split networks for palindromes 6 and & 7

a) For P6, clustering of paralogs and the reticulated network is suggestive of gene conversion.

b) For P7 the reticulated network is suggestive of gene conversion. Clustering of orthologs as opposed to paralogs could suggest that gene conversion is conservative of the ancestral state, making the orthologous palindrome arms more similar to each other than to the paralog.

could suggest that gene conversion is conservative of the ancestral state, as the orthologous palindrome arms are more similar to each other than either is to the paralog. If conversion is conservative of the ancestral state the orthologs would be more similar to each other while the presence of PSVs between paralogs in each species would create differences between the palindrome arms in both species.

4.2.2, Exploration of the *VCY* and *VCX* genes

Of the three palindromes analyzed in this chapter only P8 contains genes - these are the *VCY* genes which are located approximately 1.5kb from the outer boundaries of the palindrome. These genes are particularly interesting as they also have multiple *VCX* gametologs located on Xp21. The *VCX* genes are known to vary in copy number between individuals (Haussler et al. 2002) and are much larger than the *VCY* genes spanning over 1.6kb in comparison to 750bp for the *VCY* genes. As well as being variable in copy number the *VCX* genes also contain a 30-bp repeat unit which is variable between gene copies as well as between individuals (Haussler et al. 2002). In contrast, the *VCY* genes only contain one copy of this repeat (Haussler et al. 2002). In the past year gene conversion has been shown to occur between various X-Y gametologous genes, and in particular Trombetta et al. (2009) have claimed that the *VCY* genes act as a sequence acceptor from *VCX* during gene conversion. This suggestion is interesting as the significantly lower interspecies divergence between the P8 palindrome arms in comparisons to the spacer region (P=0.0001, 2-tailed Fisher exact test) suggests that gene conversion is conservative of the ancestral sequence. This section will carry out analysis of the *VCY* genes separately from the remainder of P8 to seek evidence of gene conversion between the *VCY* genes and its X-chromosome gametologs.

4.2.2.1, Sequence comparison

Sequences for the *VCX* and *VCY* genes were obtained from the UCSC genome browser and alignments carried out using VISTA lagan (Figure 4.5). The region shown in Figure 4.5a represents the region of chromosome X containing four isoforms of the *VCX* genes while the grey bars bellow the genes represent the gametologous regions of P8 where the *VCY* genes are situated. The grey bars within the browser show that the *VCX* genes display >95% sequence identity with the *VCY* genes.

VISTA outputs (Figure 4.5b) for the *VCY* genes demonstrate complete sequence identity between genes while the output for *VCY* and *VCX* genes identifies a block of approximately 600bp where sequence similarity is exhibited between genes. Alignments of both *VCY* genes with the four *VCX* isoforms from the reference sequence revealed variation of the 30bp repeat unit which disrupts the first 150bp of the alignment. Within this 600-bp block eight potential *VCY*-to-*VCY* conversion events were observed with two potential *VCX-VCY* conversion events also being observed.

4.2.2.2, Sequence divergence

The average sequence divergence between the *VCY* and *VCX* genes was calculated as 5.17% while no divergence was observed between the *VCY* genes. Given the apparent lack of divergence between *VCY* genes and that only 0.01% divergence was observed between the arms of P8 it is highly likely that gene conversion has been occurring between gene copies. Due to the different mutation rates of the X and Y chromosomes the true significance of divergence calculations is difficult to determine. Given the locations of the genes on different chromosomes the probability of interactions

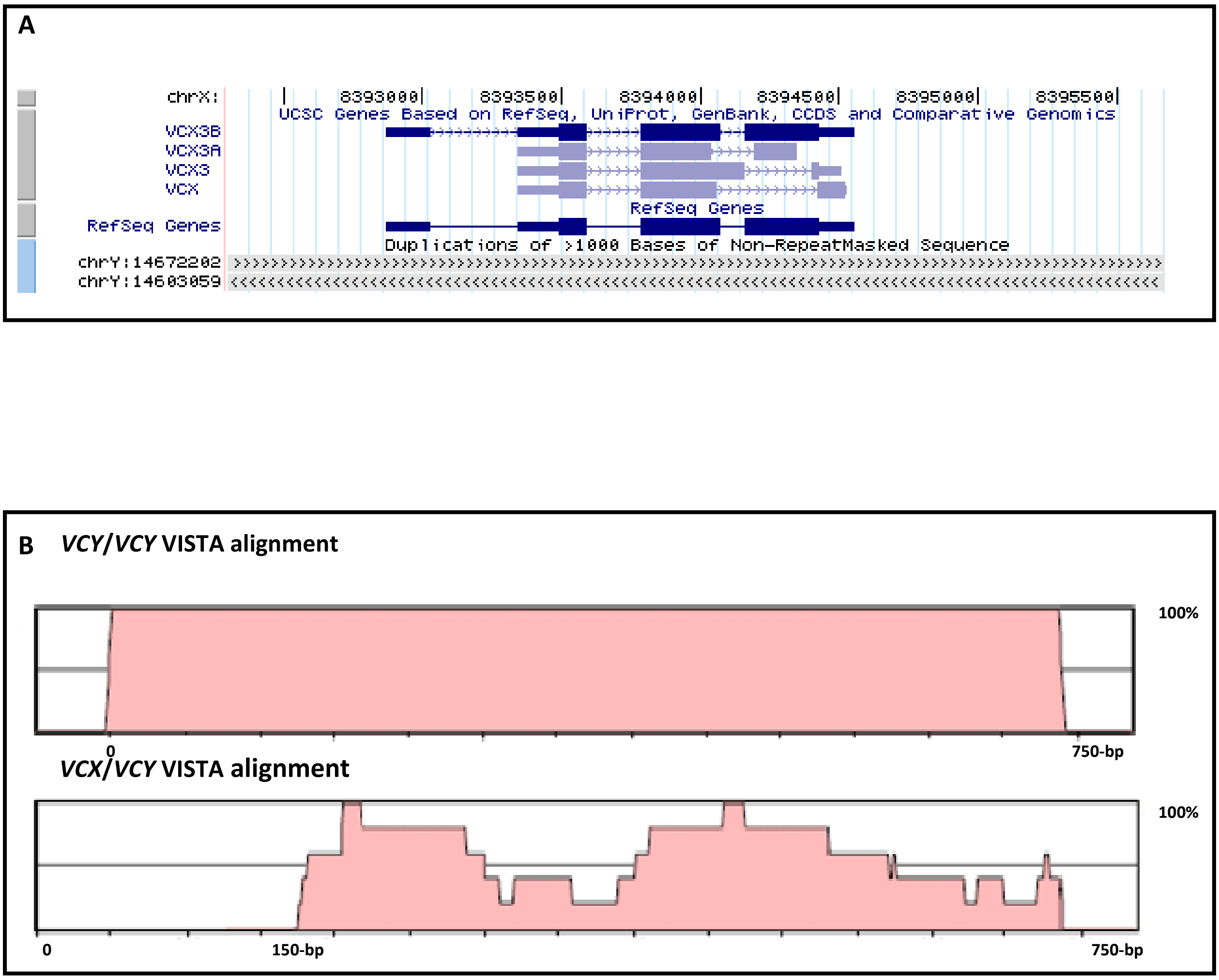


Figure 4.5: UCSC output for *VCX* - *VCY* region of palindrome 8

a) The VCY genes are situated within the arms of P8 with one copy located on each arm of the palindrome. The VCY genes also exhibit >95% sequence similarity with four isoforms of the VCX genes which are situated on Xq22.

b) 100% sequence identity is observed between *VCY* gene copies which is highly suggestive of gene conversion. Sequence alignments of *VCX* and *VCY* identified a 500bp block of sequence similarity between genes, while the first 150bp of sequence has been disrupted by a 30bp repeat which is polymorphic between *VCX* genes occurring between the *VCX* and *VCY* genes during meiosis is relatively low in comparison to interaction between Y chromosome paralogs and therefore gene conversion would be expected to be slower between the *VCX* and *VCY* genes which may make conversion events difficult to identify.

4.2.2.3, Human and chimpanzee sequence comparisons

Sequence comparisons with the chimpanzee genes may give a clearer picture regarding the history of gene conversion. Ideally, sequence comparisons would also be made with an independent primate species such as gorilla to provide the ancestral state: however, previous studies by Rozen et al. (2003) have revealed that P8 does not exist in gorilla, which suggests that the *VCY* genes have arisen on the Y chromosome prior to human and chimpanzee speciation. If this is the case then the gorilla *VCX* sequence can provide evidence of the ancestral state of the *VCX* and *VCY* genes in both human and chimpanzee. Chimpanzee *VCX* and *VCY* gene sequences were obtained from the UCSC database and alignments were carried out with the human gene sequences using VISTA lagan and clustalW. An attempt was made to obtain gorilla *VCX* sequence by carrying out BLAST searches of the human and chimpanzee *VCX* sequences against the gorilla trace archive files in NCBI; however, the sequence for these regions does not appear to be complete and reliable sequence could not be obtained.

The *VCY* genes

From sequence alignments no variation was observed between the human *VCY* genes, while the chimpanzee *VCY* gene copies only differed by a 3-bp indel. Interspecies sequence comparisons between the four *VCY* gene copies identified eight regions where there is a single base difference between the human and chimpanzee sequences which suggests that *VCY*-*VCY* gene conversion has occurred within one or both species. As an independent primate sequence is not available for comparison it is not possible to determine the ancestral state of these sites and it is therefore not possible to determine in which species the possible conversion events have occurred. To determine whether these variants could have arisen from a *VCX*-to-*VCY* conversion event and subsequent conversion into both *VCY* genes, sequence comparisons were carried out with the *VCX* genes in order to determine the ancestral state of the variants.

The *VCX* and *VCY* genes

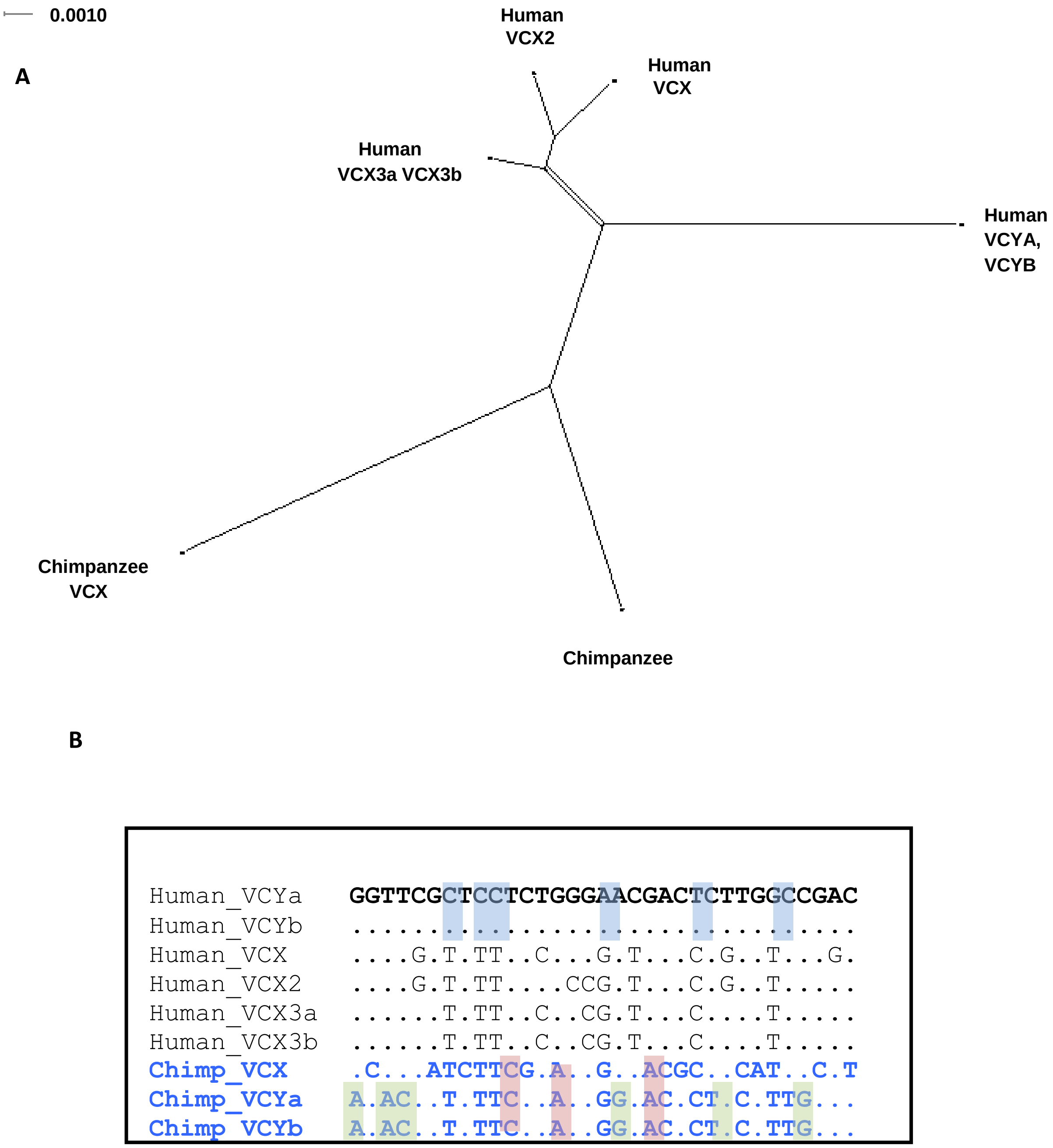
From sequence alignments of orthologous *VCX* and *VCY* genes a 600-bp block of sequence similarity was identified between all gene copies. Within this region seven nucleotides were identified between orthologous *VCX/VCY* gene copies, which appear to have undergone gene conversion in either the human or chimpanzee sequence. The *VCX* gene copy in both species does not carry the derived allele observed in the *VCY* sequences which does not suggest that these variants have arisen as the result of a *VCX*-to-*VCY* gene conversion event. Five nucleotides were identified where the *VCX* and both *VCY* genes all carry a derived allele suggesting that gene conversion may have occurred between all three genes. As it has not been possible to obtain reliable *VCX* sequence for an independent primate species it is not possible to determine in which species these potential conversion events have occurred. To assume that gene conversion is occurring between the *VCY* gene copies we must assume lack of conversion between the *VCX* gene copies if they are to be regarded as informative of the ancestral state of *VCY* PSVs.

4.2.2.4, Phylogenetic Split network

To gain further understanding of the ancestral relationship between the *VCX* and *VCY* orthologs a Split network was constructed based on the 550-bp block of sequence similarity observed between genes (Figure 4.6). The network shows clustering of human and chimpanzee orthologs at either side of the network while the reticulated structure between the human *VCX* and *VCY* genes is suggestive of gene conversion.

4.2.2.5, Interspecies divergence between the *VCY* genes

Sequence comparisons suggest that gene conversion has occurred between *VCY* gene copies in both humans and chimpanzee with many potential conversion events to the derived allele being observed in each species. A problem when identifying conversion events from a single reference sequence is that conversions to the ancestral state are "invisible" and will homogenise both sequences between species. Zero divergence was observed between the *VCY* gene copies in both human and chimpanzee which strongly suggests that gene conversion has been occurring in both species. Interestingly, a high interspecies divergence of 5.17% was observed between the human and chimpanzee *VCY* genes, which suggests that gene conversion is not conservative of the ancestral state (Figure 4.7). This is interesting as the significantly lower interspecies divergence between the non *VCY* segments of P8 suggests that gene conversion is conservative of the ancestral state. This observation could either be attributed to natural selection or the recent suggestion by Trombetta et al. (2009) that the *VCY* genes act as a sequence acceptor from *VCX* during gene conversion.



```
Human_VCYa    GGTTCGCTCCTCTGGGAACGACTCTTGGCCGAC
Human_VCYb    .................................
Human_VCX     ....G.T.TT..C...G.T...C.G..T...G.
Human_VCX2    ....G.T.TT....CCG.T...C.G..T.....
Human_VCX3a   ......T.TT..C..CG.T...C....T.....
Human_VCX3b   ......T.TT..C..CG.T...C....T.....
Chimp_VCX     .C...ATCTTCG.A..G..ACGC..CAT..C.T
Chimp_VCYa    A.AC..T.TTC..A..GG.AC.CT.C.TTG...
Chimp_VCYb    A.AC..T.TTC..A..GG.AC.CT.C.TTG...
```

Figure 4.6: Phylogenetic split network for the *VCX* and *VCY* genes

a) A splitstree network based on the 600bp region of gametology between all *VCX* and *VCY* orthologs. The network shows clustering of orthologs while the presence of a reticulated structure between the human *VCX* and *VCY* genes is suggestive of gene conversion.

b) Sequence comparisons identified 33 variants between sequences. Of these 6 are sites where conversion appears to have occurred between the human *VCY* genes (Blue), 7 are sites where gene conversion appears to have occurred between the Chimpanzee *VCY* genes (Green) and 3 are possible *VCX VCY* conversion events (Pink).

The *VCX* and *VCY* genes

As multiple copies of the *VCX* genes are known to exist in the human genome the divergence between each *VCX* gene and *VCY* was determined and the average divergence between *VCX* and *VCY* was calculated. Interspecies divergence between the *VCX* genes was calculated as 3.14% which is lower than the 5.17% divergence observed between *VCY* genes. This is not surprising as the X chromosome is known to have a lower rate of mutation than the Y chromosome (Nachmana and Crowella 2000). Divergence between chimpanzee *VCX* and *VCY* was calculated as 2.5% which is significantly lower (P=0.0001, 2-tailed Fisher exact test), than the 4.35% divergence between human *VCX* and *VCY* genes. This lower divergence between human *VCX* and *VCY* genes could be suggestive of the *VCY* genes acting as sequence acceptor from *VCX*, as gene conversion would lower divergence between *VCX* and *VCY* sequences in both species but creates greater divergence between orthologs.

4.2.3, Exploration of inverted repeats

Five sets of inverted repeats, termed IR1-IR5, are situated within the ampliconic regions of the Y chromosome. IRs are similar in structure to palindromes but contain much larger spacers which range in size from 240kb to 1.5Mb and display similarity of 98-99% between paralogs. In contrast to the palindromes which are located solely within the ampliconic regions of Yq, the IRs are situated on both arms of the chromosome and while IR2 and IR5 are both located on Yq and IR3 on Yp, IRs 1 and 4 both have one paralog located on Yq and the second on Yp. The presence or absence of palindromes has previously been determined in chimpanzee and other non-human primates: however, very little is known about the presence of IRs in primates. As the chimpanzee

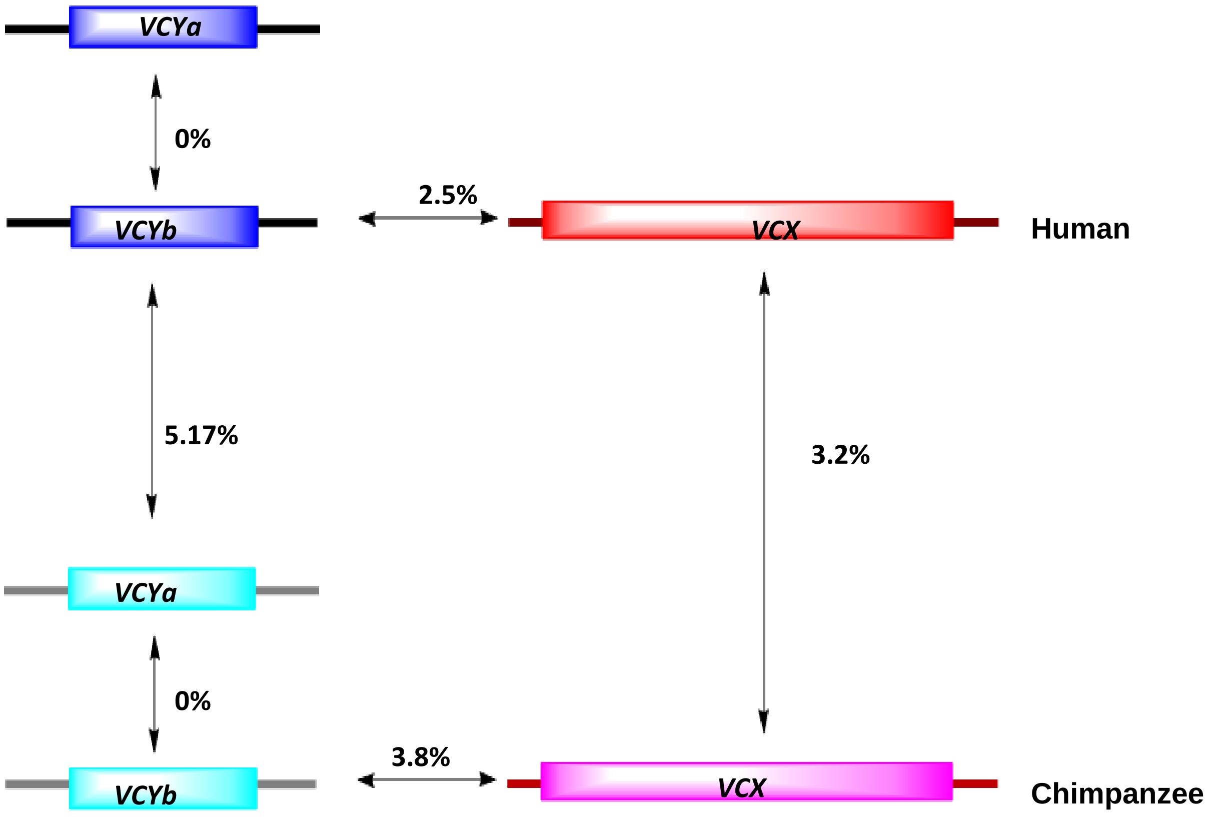


Figure 4.7: Divergence between *VCX* and *VCY* genes within and between humans and chimpanzee

While no divergence was observed between *VCY* genes in either species while divergence of 5.17% was observed between the *VCY* orthologs. Divergence between the *VCX* and *VCY* genes in chimpanzee was higher at 3.5% than the divergence of 2.5% observed between human *VCX and VCY* genes.

Y chromosome is smaller than that of the human it is possible that some IRs have arisen as the result of a human-specific duplication event. NAHR has previously been shown to occur between the paralogs of some IRs which results in chromosomal rearrangements such as translocations and inversions (Hurles and Jobling 2003). (Repping et al. 2006). Evidence of NAHR between IR3 paralogs has been well documented (Repping et al. 2006) and it has been hypothesized that NAHR occurring between the paralogs of IR1 or IR4 results in INV(Yp;Yq) (Hurles 2004). As NAHR has been shown to occur between the paralogs of some IRs it is also possible that gene conversion may also occur.

This section will explore IRs 1-4 which are not known to have previously been studied for evidence of gene conversion, but excludes IR5 which forms part of P1, which itself is known to undergo gene conversion. Of these IRs particular attention will be paid to IR1 and IR4, both of which have one paralog on Yp and a second on Yq. Evidence of gene conversion between either of the IRs would suggest that the Y chromosome can fold on itself allowing inter-arm recombination in the germline.

4.2.3.1, Sequence comparison

Sequences for each IR were obtained from the UCSC genome browser using landmark STSs described by Skaletsky et al. (2003), and sequences were aligned using VISTA lagan (Figure 4.8). In contrast to the previously studied palindromes, the regions in which the IRs are situated have multiple regions of identity within the Y chromosome. This makes it more difficult to study gene conversion in these regions, as one paralogous region cannot be easily distinguished from another in simple PCR-based approaches, and due to the high degree of similarity between sequences multiple sequence interactions have the potential to occur. As previously discussed in chapter 1, some IRs are located within a region of Yq which is prone to various duplications and deletions. These deletions have the potential to remove the whole of IR2 and the Yq repeat unit of IR1 and IR4. The mechanisms of these deletions will be discussed in more detail in Chapter 8. IR2 is further complicated by homology with P3, while IR3 is complicated by multiple testis-specific transcript (*TTTY)* clusters which are situated across the Y chromosome. Sequence alignments for each IR reveal that while a high degree of sequence similarity is exhibited between paralogs of IR2 and IR3, the copies of IR1 and IR4 are disrupted by multiple PSVs and indels. This high sequence similarity observed for IR2 and IR3 is suggestive of gene conversion; however, as it is not known whether these structures predate speciation, it is also possible that this high sequence similarity may be due to a human-specific duplication event.

4.2.3.2, Sequence divergence

Sequence divergence between the paralogs of each IR was calculated using DnaSP. As previously observed from the VISTA outputs, IR2 and IR3 showed the lowest average sequence divergence of 0.07% and 0.5% respectively while IR1 and IR4 both have a higher divergence of 1.56% and 4% respectively. Assuming that IR2 and IR3 have not arisen as the result of a duplication event, the low divergence observed between paralogs since speciation is suggestive of past gene conversion while the higher divergence between paralogs of IR1 suggests that conversion may have been occurring but at a slower rate. In the case of IR4 the amount of disruption between sequences and the high sequence divergence compared to other Y chromosome paralogs suggests that conversion is most likely not occurring within this region, or that conversion events are

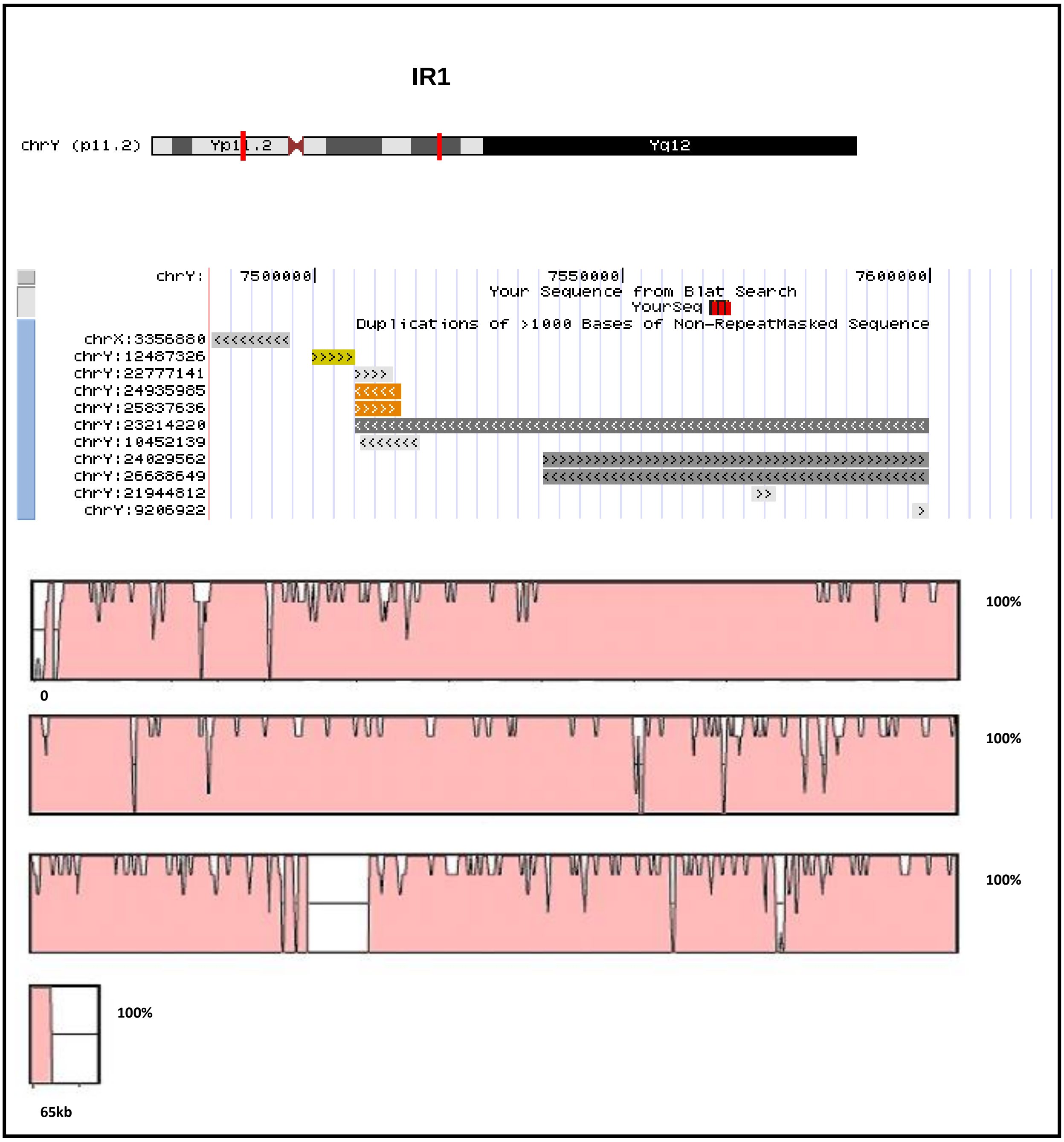


Figure 4.8a: Sequence analysis of human IR1

a) IR1consists of two 65kb paralogs one situated on Yp and the second situated on Yq. Paralogy was also observed between IR1 and Palindrome 1 (P1) towards the outer boundaries leaving only a 17-kb region of unique Yp-Yq identity.

b) Sequences alignments reveal the presence of many PSVs and indels between paralogs and does not suggest that gene conversion occurs rapidly between the IR1 paralogs.

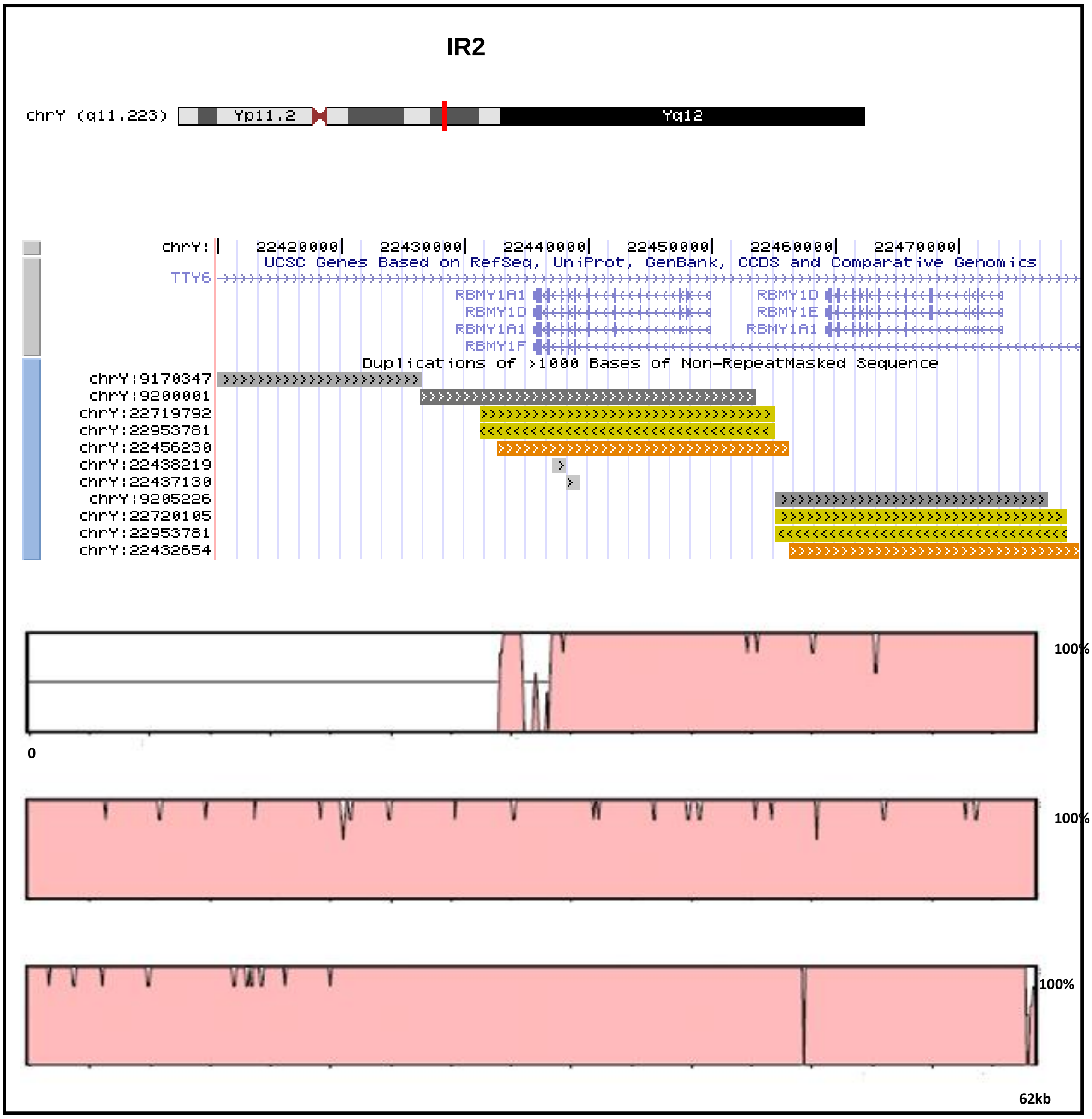


Figure 4.8b: Sequence analysis of human IR2

a) IR2 is situated solely on Yq and consists of two paralogs which each span 62-kb. The IR2 paralogs contain the *RBMY* genes of which there are four copies in IR2 and two copies located in Palindrome 3 (P3).

b) Although a high degree of similarity is observed between sequences which is suggestive of gene conversion, the study of gene conversion would be complicated by the presence of multiple *RBMY* genes and possible Yq deletions.

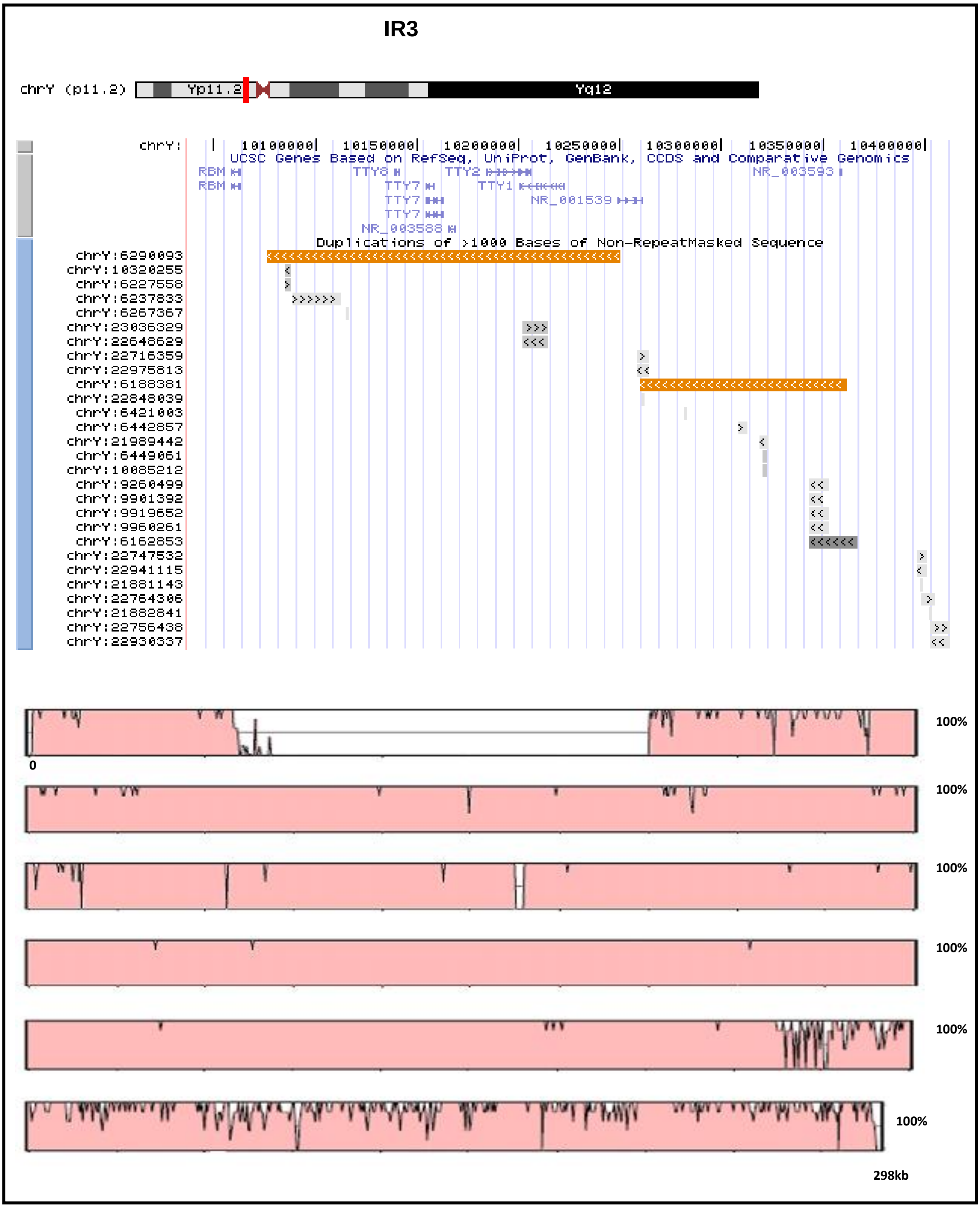


Figure 4.8c: Sequence analysis of human IR3

a) The paralogs of IR3 span 298-kb and are located solely on Yp. IR3 contains multiple copies of the *TTY* genes of which there are multiple copies distributed across the Y chromosome.

b) Although a high degree of similarity is observed between sequences, which is suggestive of gene conversion, the study of gene conversion would be complicated by the presence of multiple *TTY* genes.

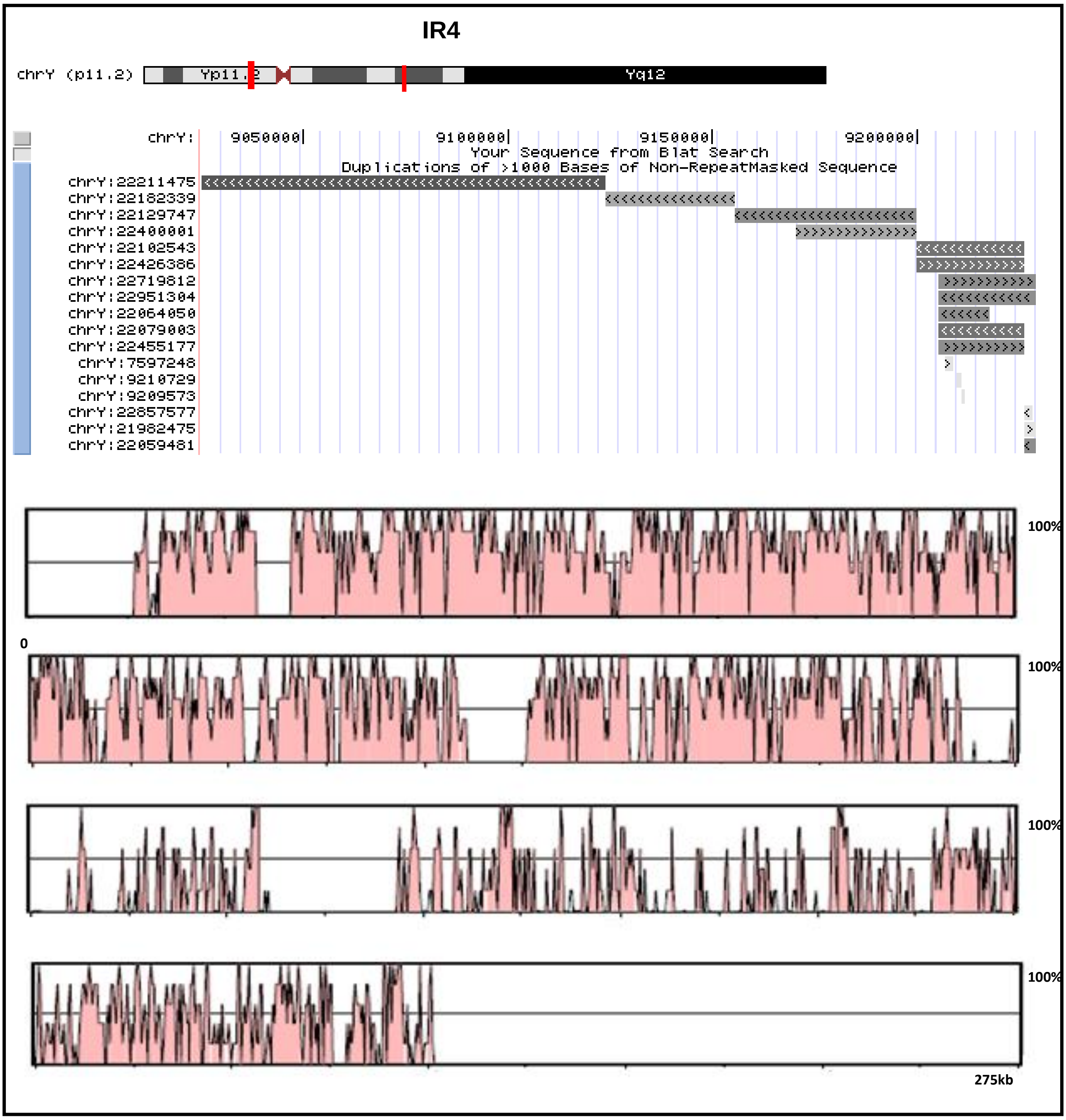


Figure 4.8d: Sequence analysis of human IR4

a) IR4 consists of two 275-kb paralogs one situated on Yp and the second situated on Yq.

b) Sequence alignments reveals the paralogs have become disrupted by the accumulation of multiple PSVs and large indels. This suggests that gene conversion does not occur frequently between the IR4 paralogs and therefore conversion events are likely to be difficult to identify.

rare.

4.2.3.3, Human and chimpanzee sequence comparison

While the palindromes have been relatively well characterized in the chimpanzee, little is known about the presence of IRs in the chimpanzee genome. Rozen et al. (2003) have previously sequenced the boundaries of all eight palindromes in chimpanzee to determine their presence or absence prior to speciation; however, no such study has been carried out for the IRs. Given the complexity of the regions in which the IRs reside in humans, re-sequencing of chimpanzee IRs may not be a simple task. To determine the presence of IRs in the chimpanzee genome, sequences were either obtained from the UCSC genome browser or from NCBI by carrying out BLAST searches of the human IR sequences against the chimpanzee reference sequence. Of the four IRs, full chimpanzee sequence could be obtained for IR1 (Figure 4.9) while only partial sequence for IR2 was obtained. No sequence could be found for IR3 or IR4 which suggests that these IRs may have arisen as the result of a duplication event in humans. However, as only partial sequence was obtained for chimpanzee IR2 it is possible that the chimpanzee reference sequence is incomplete for these regions. As full chimpanzee sequence could only be obtained for IR1 the remainder of this section will consider interspecies sequence comparisons for IR1 only. Chimpanzee sequences were obtained from the UCSC genome browser and sequence alignments carried out using VISTA lagan and clustalW. A higher degree of sequence similarity was observed between chimpanzee sequences than was observed between human sequences. While the human sequence alignment identified two large indels of 1.2kb and 2.9kb, only single nucleotide indels were observed in the chimpanzee alignment. Four-way sequence alignments carried out using clustalW identified multiple regions where gene conversion appears to have occurred but without an independent primate species for

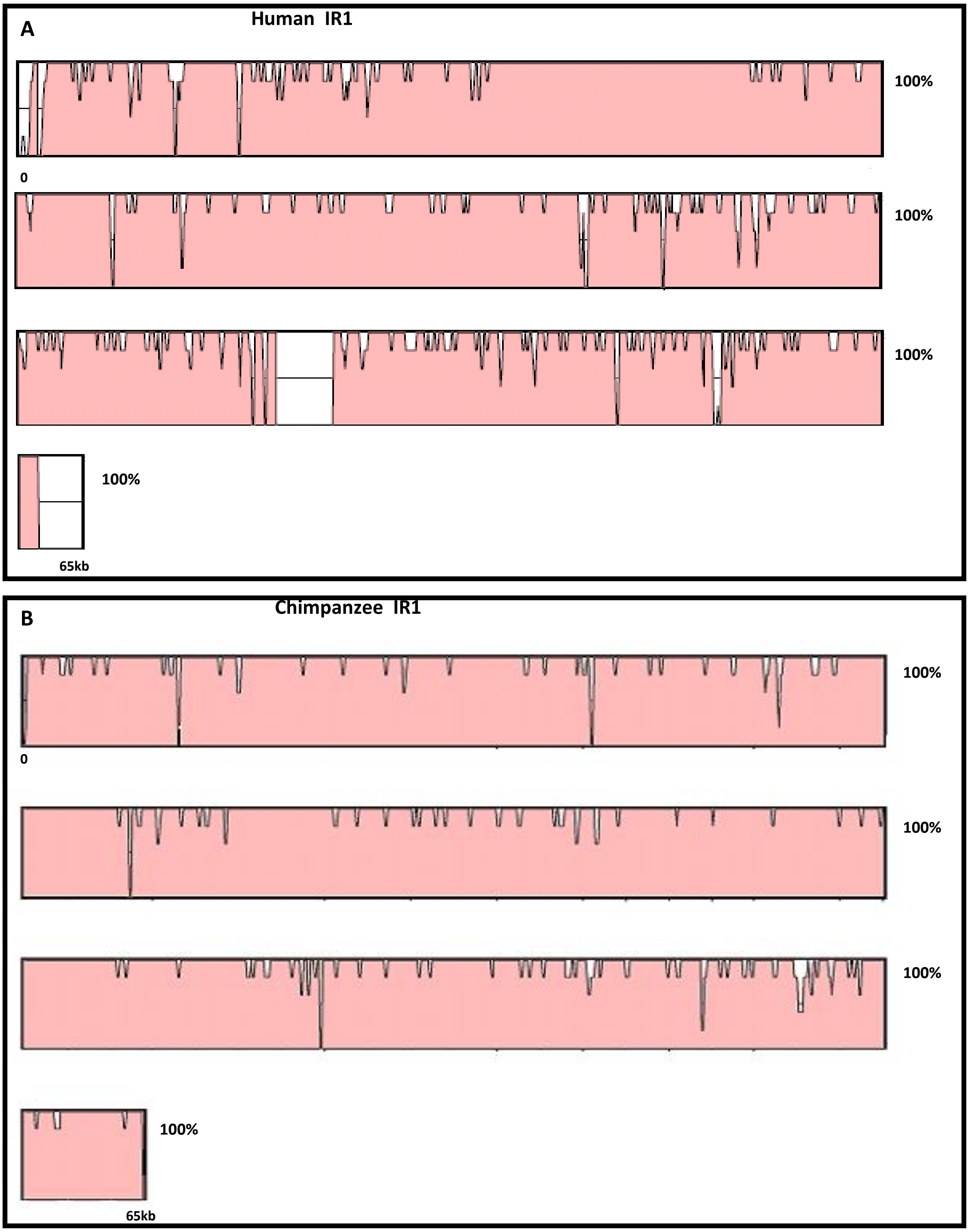


Figure 4.9: Human and chimpanzee sequence comparisons

a) Sequence comparisons reveal the human sequences to have become disrupted by the accumulation of multiple PSVs and several large indels. This could suggest that gene conversion is not a frequent occurrence between human paralogs or that the indels have disrupted gene conversion.

b) A higher degree of sequence similarity is displayed between the chimpanzee paralogs. This suggests that gene conversion may be occurring at a faster rate between the chimpanzee paralogs.

comparison it is not possible to determine in which species gene conversion may have occurred.

4.2.3.4, Interspecies sequence divergence

Interestingly interspecies sequence divergence between IR1 paralogs was calculated as 1.54% which is significantly higher (P=0.0001, 2-tailed fisher exact test) than the 0.12% divergence observed between the chimpanzee IR1 paralogs but lower than the 1.32% divergence observed between human IR1 paralogs. From comparisons of human and chimpanzee sequences it appears that gene conversion occurs between chimpanzee paralogs: however, due to the disruption of sequence alignments by two large indels and generally higher sequence divergence, it appears that if gene conversion does occur between human IR1 paralogs it is not likely to be a frequent occurrence.

4.2.3.5, Phylogenetic split network

To gain an understanding of the ancestral relationship between human and chimpanzee IR1 a phylogenetic Split network was constructed (Figure 4.10). The network shows clustering of paralogs at each end of the network which suggests that gene conversion is occurring in both species. The presence of a reticulation at the centre of the network also suggests that gene conversion occurs. This network and low divergence between IR1 paralogs since speciation suggests that gene conversion has been occurring.

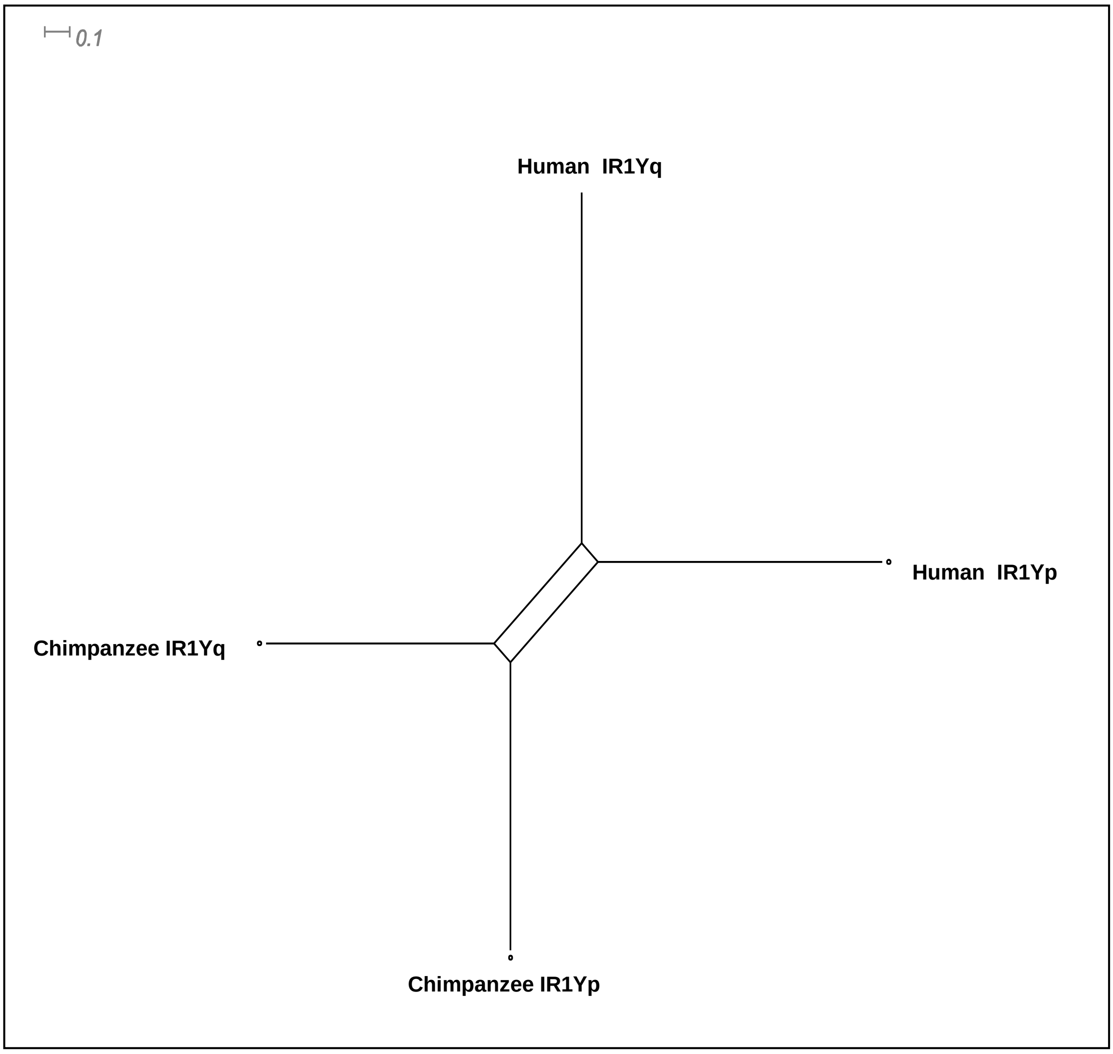


Figure 4.10: Phylogenetic split network of human and chimpanzee IR1 sequences Clustering of parlogs and the reticulated structure of the network suggests that gene conversion occurs between both human and chimpanzee IR1 sequences.

4.2.4, Exploration of the X-transposed region (XTR) of Yp

Approximately 4.7MYA (Ross et al. 2005) a transposition event between the X and Y chromosomes resulted in the transfer of a 3.8-Mb block of sequence containing three genes from Xq21 to Yp11. Since transposition a series of inversions and deletions on the Y chromosome have caused the block to split into two segments which span 3.38Mb and 200kb. Approximately 98% sequence similarity is exhibited between the X and Y sequences suggesting that gene conversion has the potential to occur; however, as the estimate of the time of transposition is itself based on the degree of divergence between sequences it is difficult to determine the full significance of this divergence. The three genes located within the XTR each have a functional copy on the X chromosome and as gene conversion has previously been reported between XY-homologous genes there is also the potential for gene conversion to occur between these genes. Given the location of the XTR on Yp11 and the homologous sequence on Xq21, gene conversion might be expected to occur at a slower rate than between the Y chromosome paralogs as the large physical distance separating the gametelogs means they are less likely to interact during meiosis.

The study of gene conversion between the X and Y chromosomes is more complicated than that of conversion between Y-chromosome paralogs. As the X chromosome is diploid in females, recombination can occur along the entire chromosome length during meiosis while on the Y chromosome recombination is suppressed for over 95% of the chromosome length and is restricted to the PARs. The diploid nature of the X chromosome introduces complications such as interallelic diversity which makes it difficult to distinguish gene conversion from crossover. Additionally, diploid chromosomes are known to contain recombination hotspots which have been shown to be focal points for allelic gene conversion (Jeffreys and May 2004) and therefore such gene conversion could also occur between X chromosomes. While the Y chromosome has a well-defined phylogeny which allows historical gene conversion events to be identified and the rate of conversion to be estimated, the lack of a phylogenetic framework for the X chromosome means that historical gene conversion is difficult to identify and the rate of gene conversion cannot readily be estimated. Another complicating factor is the larger effective population size of the X chromosome, which is (in principle) three times that of the Y chromosome. This means that Y-to-X conversion events may be passed to multiple descendant chromosomes leading to over estimation of a conversion event or its lost from the population during meiosis.

4.2.4.1, Sequence comparisons between the XTR and Xq21

The XTR of the Y chromosome was identified in the UCSC genome browser (Figure 4.11a) and the X and Y sequences were aligned using VISTA lagan (Figure 4.11b). A large number of GSVs were observed between X and Y sequences and sequence alignments were disrupted by fifteen indels ranging in size from 3-bp – 2.5kb. As rapid gene conversion would be expected to homogenise the two sequences this high number of GSVs observed between the chromosomes suggests that if gene conversion does occur in this region is likely to be at a slow rate.

4.2.4.2, Sequence divergence in the X transposed region

As the XTR spans approximatly 4Mb and contains small regions which exhibit multiple paralogy, the XTR will be divided into five sequence blocks to exclude regions where multiple sequence interactions may potentially occur. Divergence for each segment of

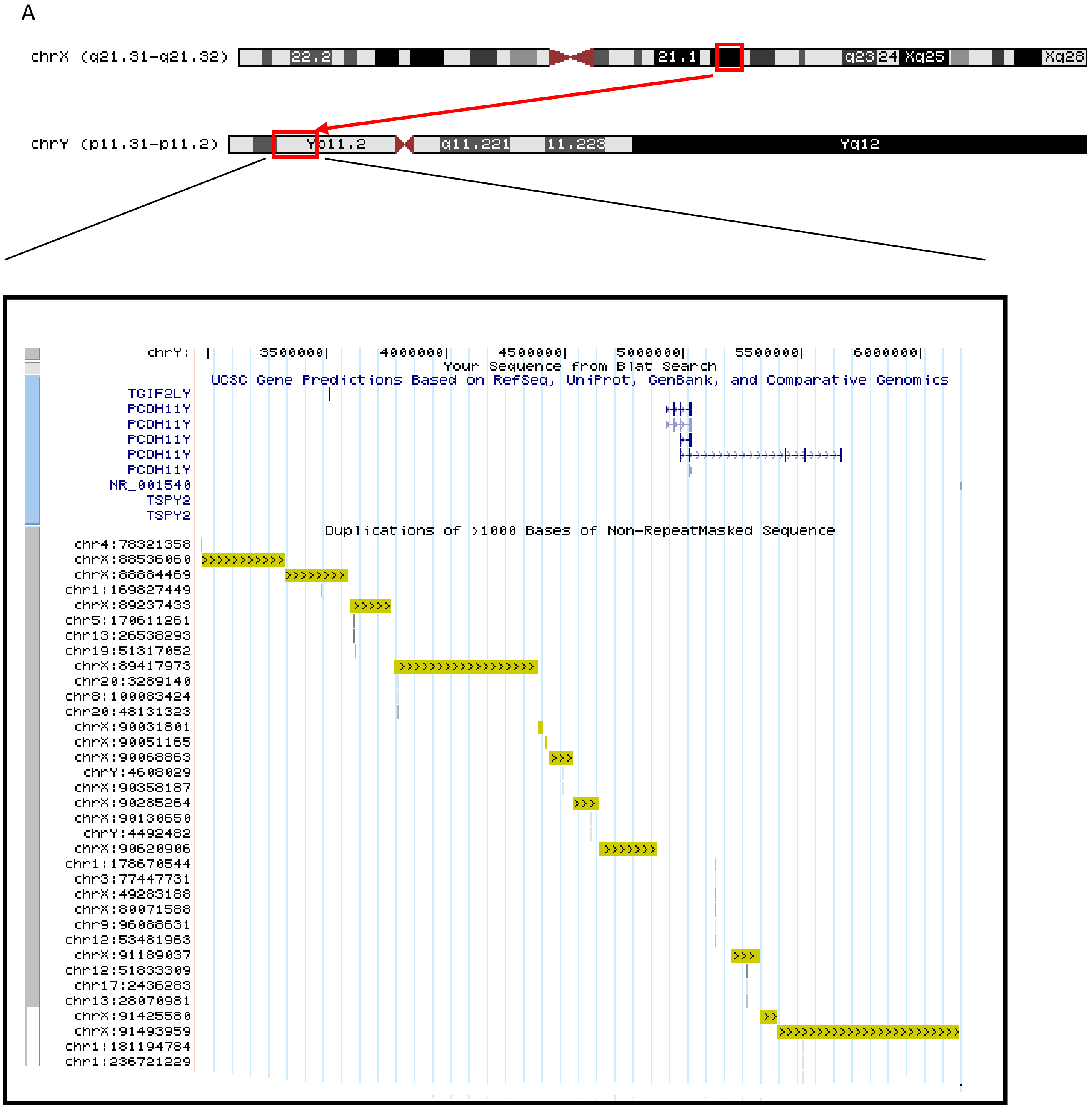
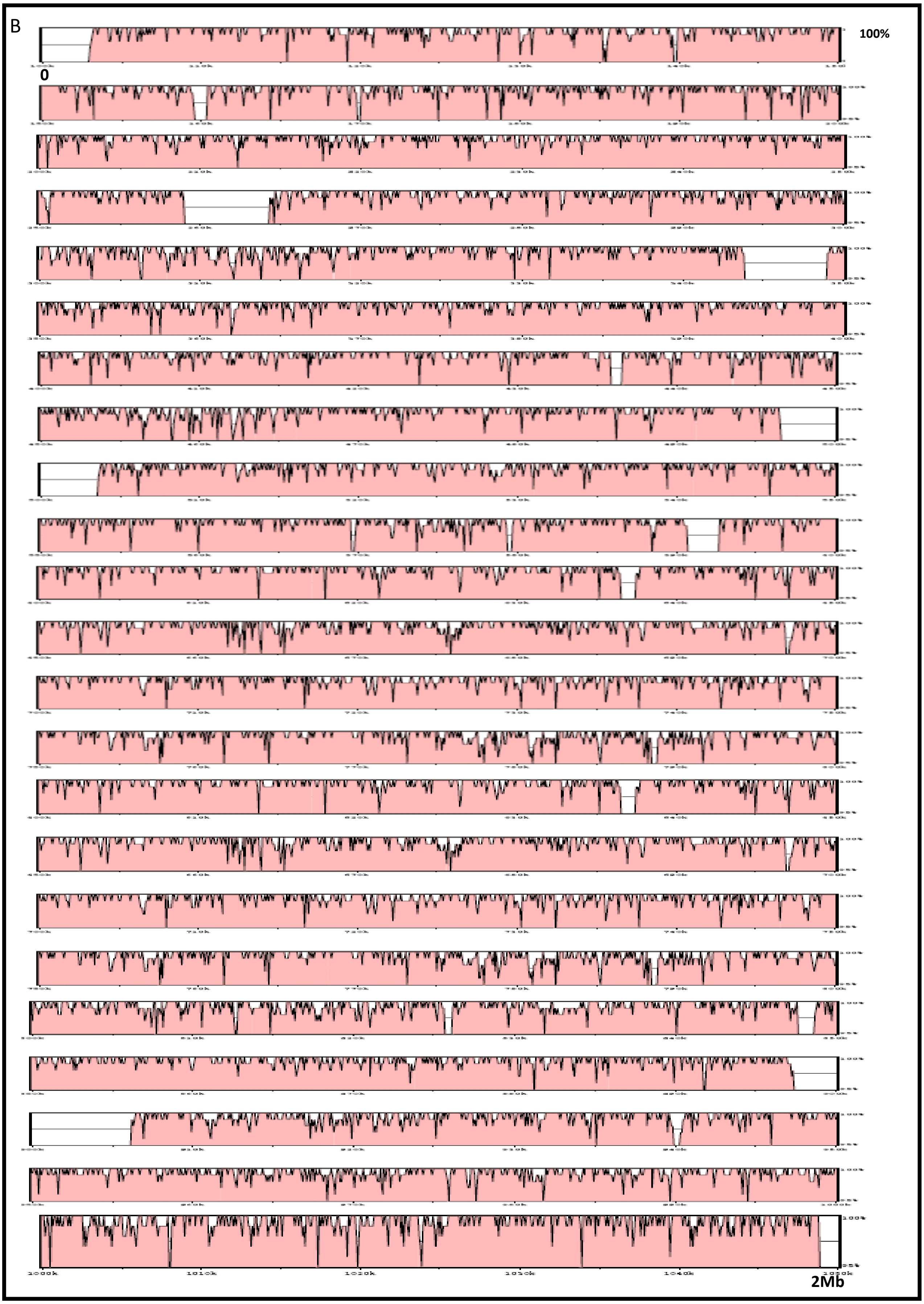


Figure 4.11a: UCSC genome browser for the XTR of the human Y chromosome

a) The regions corresponding to the XTR of the Y chromosome and Xq22 were located in the UCSC genome browser.

b) Sequence alignments reveal that since transposition the gametologous sequences have become disrupted by the accumulation of multiple GSVs and INDELS which suggests that gene conversion between these regions it is not likely to be a frequent occurrence.


B
100%
0
2Mb

the XTR was calculated using DnaSP, and the average divergence across the whole region was determined to be 1.21%. When comparing sequence divergence between the X and Ychromosomes it must be taken into account that the two chromosomes have different mutation rates, with the Y chromosome mutating faster than the X. It could be argued that if rapid gene conversion were occurring between the X and Y chromosomes then the sequences will be homogenized despite the very different mutation rates. However, if gene conversion is slow or rare it is unlikely to influence sequence divergence.

4.2.4.3, Human and chimpanzee sequence comparison

To gain an understanding of how the sequences have evolved since transposition, comparisons were made with the chimpanzee reference sequence. As transposition occurred after speciation, only sequence for chimpanzee Xq21 is available for sequence comparisons to be made. This is useful when carrying out analysis of human Xq21 and the XTR as X-Y gene conversion cannot be occurring in the chimpanzee this will give an idea of how the sequences have evolved since speciation, and allow the ancestral states of human GSVs to be determined.

The region corresponding to human Xq21 was identified in the chimpanzee UCSC reference sequence and alignments with human Xq21 and the XTR were carried out. Sequence comparisons revealed that the chimpanzee sequence for Xq21 was not complete, creating multiple gaps in the sequence alignment, and sequence comparisons for the entire XTR were not possible. To gain some idea of how the XTR has evolved since speciation comparisons with the available portions of the chimpanzee sequence were made.

4.2.4.4, Human and chimpanzee sequence divergence

Divergence between the human and chimpanzee Xq21 sequences was calculated as 0.94% which is similar to the 1% average divergence previously reported for the X chromosome. Interspecies divergence between chimpanzee Xq21 and the XTR of the Y chromosome was determined to be 1.53% which is significantly higher (P=0.0001, 2-tailed Fisher exact test) than the 1.23% observed between human Xq21 and the XTR. While this lower divergence between human sequences could suggest that gene conversion has been occurring it may also reflect the presence of variations between the human and chimpanzee Xq21 sequences at the time of transposition (Figure 4.12).

As transposition time has been calculated based on sequence divergence between the two chromosomes it is not possible to determine whether sequence divergence is lower than expected: however 98.79% sequence similarity observed between Xq21 and the XTR suggests that gene conversion has the potential to occur and could be detectable among extant chromosomes.

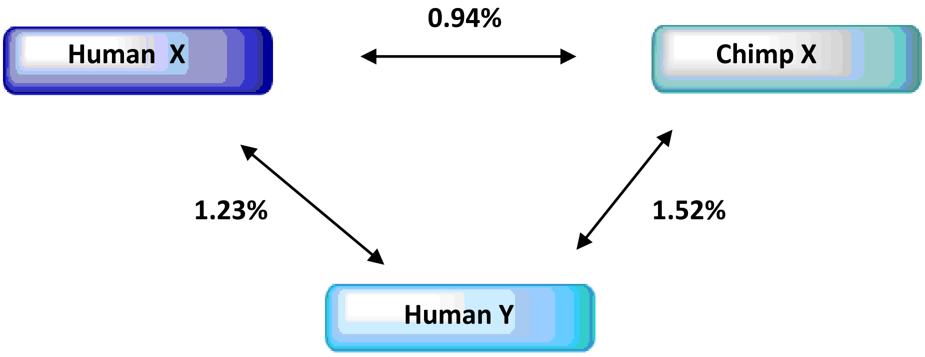


Figure 4.12: Sequence divergence between the human and chimpanzee XTR
From the comparison of 2Mb of sequence, .94% divergence is observed between the human and chimpanzee X chromosome sequences which is approximately the same as the average X chromosome divergence of 1%. Lower divergence is observed between the human X and Y chromosome sequences (1.23%) than between the chimpanzee X and human Y chromosome sequences (1.52%).

4.3, Discussion

The Y chromosome is rich in paralogous repeats which exhibit >95% identity between sequences and could potentially be undergoing gene conversion. While a high degree of sequence similarity is suggestive of gene conversion it may also indicate that a paralog has arisen as the result of a more recent duplication event. When seeking evidence of gene conversion on the Y chromosome it is difficult to distinguish the paralogs which exhibit high sequence similarity due to gene conversion from those that have arisen as the result of a more recent duplication event since the Y chromosome is prone to multiple duplications and deletions.

A bioinformatic analysis of available sequences can give some indication as to which paralogs are undergoing gene conversion and to some extent can indicate the rate and direction of gene conversion. The use of the chimpanzee reference sequence allows the presence or absence of paralogs prior to speciation to be determined. Determining presence of a paralog in the chimpanzee genome prior to speciation indicates that low sequence divergence between human paralogs cannot be due to a recent duplication event and could be the result of gene conversion. While the use of bioinformatics can indicate which paralogs may be undergoing gene conversion, direct evidence of gene conversion cannot be gained without experimental analysis being carried out. Over the past five years several additional Y chromosome sequences have been published but the majority of these represent chromosome from the same Y chromosome haplogroup and offer only limited additional information in the study of gene conversion. As more Y chromosome sequences become available it may become increasingly possible to identify gene conversion events without the need for experimental analysis. Publication of the 1000 genomes project would allow the identification of additional PSVs and is expected to make identification of conversion events *in silico*, increasingly possible.

From the bioinformatics analysis carried out in this chapter four regions were identified for further analysis, these are P6, P8, IR1 and the XTR. Further analysis of P6 will be carried out as it contains a higher number of PSVs than P7 and P8, and offers greater potential to identify gene conversion events. Also, conservative gene conversion is hypothesized to protect the function of spermatogenic genes and as P6 is non-genic, evidence of conservative gene conversion would suggest that this is not a phenomenon which is associated solely with palindromes that contain genes. P8 is also interesting as it appears that conservative gene conversion occurs rapidly between palindrome arms: in contrast, interspecies sequence comparisons between the *VCY* genes is not suggestive of conservative gene conversion. As the *VCY* genes have also been claimed to act as a sequence acceptor from *VCX* during gene conversion (Trombetta et al. 2009) direct evidence of *VCX*-to-*VCY* gene conversion events will be sought. The high degree of sequence similarity observed between human and chimpanzee IR1 sequences suggests that gene conversion has been occurring in both species. Observing evidence of gene conversion between the paralogs of IR1 would also be interesting as this would provide evidence of inter-arm recombination occurring between Yp and Yq, which would suggest that the Y chromosome can fold on itself across the centromere allowing recombination during meiosis. While it is difficult to determine the significance of divergence calculations between Xq21 and the XTR, evidence of gene conversion will also be sought. As the estimation of transposition time has been based on divergence between the two sequences evidence of gene conversion between sequences could potentially alter estimations of transposition time.

Chapter 5: Identifying conversion events between duplicated regions of the Y chromosome

Despite the Y chromosome being a useful tool for studying non-allelic gene conversion, identifying conversion events between duplicated regions of the Y chromosome is not always easy. The difficulty varies depending on the class of variant being typed as well as the region in which it is located. Following a duplication event which creates a paralog, both sequences will be identical until a mutation occurs within one paralog creating a variant between the two sequences. Over time additional mutations will occur, creating additional variants in descendant chromosomes. This study will involve typing two classes of variant single nucleotide PSVs (snPSVs) and microsatellites, identified between duplicated regions of the Y chromosome as well as snGSVs identified between gametologous regions of the X and Y chromosomes. This Chapter will provide a general preface to the following Chapters, by outlining how conversion events can be identified, the assumptions which must be made and the problems associated with typing different variants.

5.1, Single nucleotide PSVs (snPSVs)

When looking for evidence of gene conversion snPSVs, can be typed in chromosomes from diverse Y chromosome haplogroups, and the ancestral states for individual haplogroups can be inferred by maximum parsimony. In the study of gene conversion, haplogroups with a pseudoheterozygous ancestor are most informative, as given the low probability of SNP reversion/recurrence via point mutation the observation of pseudohomozygosity within descendant chromosomes provide evidence of gene conversion. Comparisons with the chimpanzee and other non-human primates such as gorilla and macaque will provide evidence of the deep-rooting ancestral state of a PSV, allowing conversion events towards and away from the ancestral allele to easily be identified (Figure 5.1).

5.1.1, Problems associated with typing snPSVs

Although identifying conversion events through typing snPSVs is relatively straightforward, there are several factors which can complicate interpretation. Firstly, although the probability of SNP recurrence is generally low, mutations occur much more frequently at the site of a CpG dinucleotide than at non CpG sites (Ehrlich and Wang 1981). In snPSVs which occur at the site of a highly mutable CpG dinucleotide, gene conversion cannot easily be distinguished from CpG hypermutation and therefore these PSVs cannot be included in analysis. Secondly, identifying snPSVs from a single reference sequence creates an ascertainment bias, as only sites which are variable within that sequence will be identified. PSVs which have been homogenised through gene conversion will not be identified from a single chromosome sequence, therefore gene conversion events will be missed leading to underestimation of the rate of gene conversion. Some PSVs may also be private to an individual chromosome or a subclade of the Y phylogeny, and will not provide any information on gene conversion. From the analysis of a single reference sequence, distinguishing the potentially informative PSVs from non-informative PSVs is not possible and valuable time may be wasted typing PSVs which are non-informative. These problems can be overcome by sequencing chromosomes from diverse Y chromosome haplogroups in order to identify additional PSVs and to distinguish potentially informative PSVs from those which are non-informative.

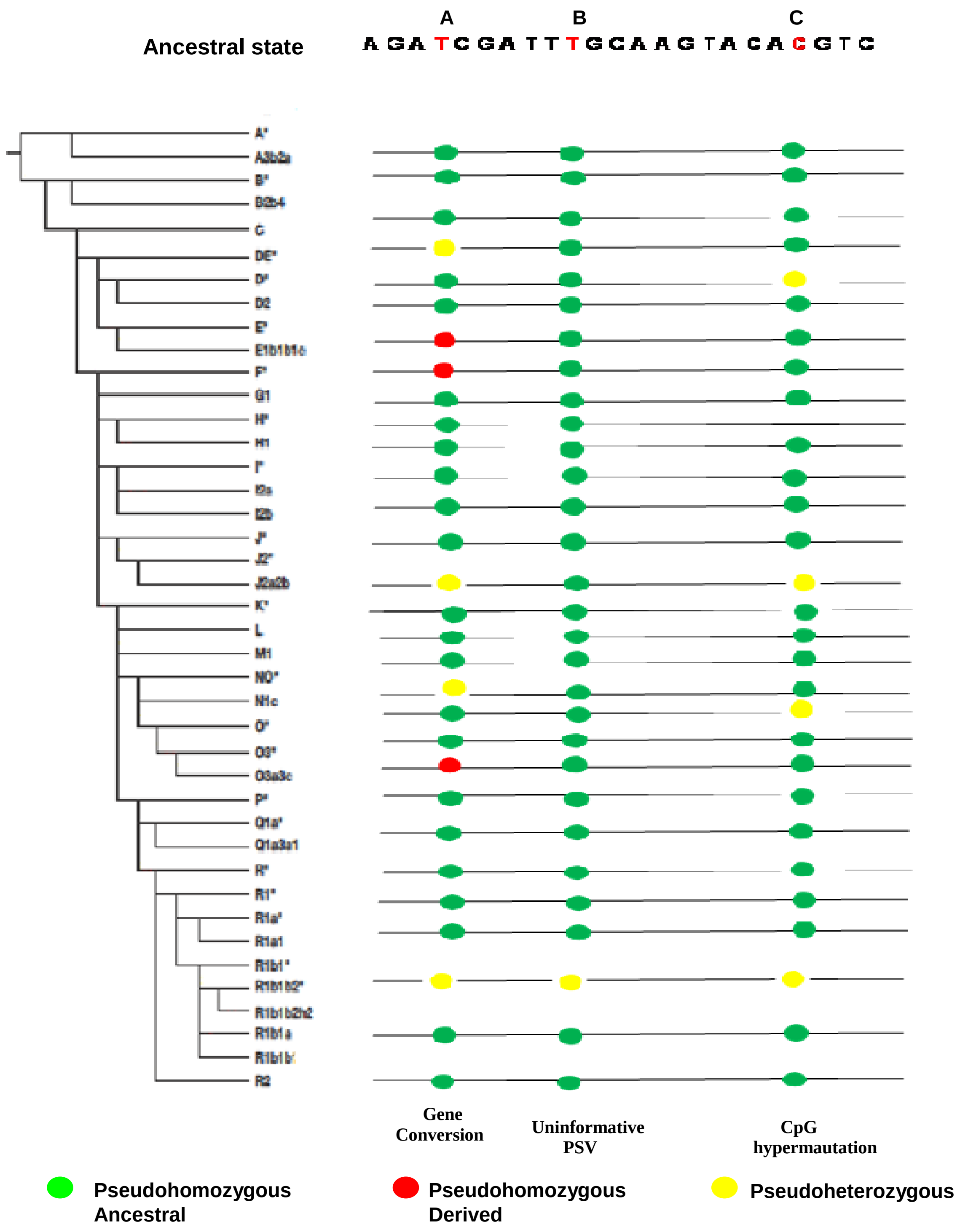


Figure 5.1: Identifying conversion events through phylogenetic analysis of snPSVs

a) Comparisons with the chimpanzee reference sequence allows the ancestral state of a PSV to be determined and conversions events to and away from the ancestral allele can easily be identified

b) Phylogentic analysis also allows the identification of PSVs which may be specific to an individual sequence and will offer no information on gene conversion

c) Gene conversion at the site of a CpG dinucleotide cannot be distinguished from hyper mutation especially if conversion is unidirectional.

5.2, Identifying conversion events between duplicated microsatellites

Typing duplicated microsatellites overcomes the ascertainment bias associated with typing snPSVs as all duplicated microsatellites within a paralog can easily be identified from a single reference sequence. However, identifying conversion events through typing duplicated microsatellites is more complex than for snPSVs. Duplicated microsatellites are subject not only to the same evolutionary mechanisms operating at single-copy microsatellites, but also to mechanisms specific to duplicons (Balaresque et al. 2007). When looking for evidence of gene conversion occurring between duplicated microsatellites it is useful to establish a model of the expected outcomes that mutation alone may have on microsatellite diversity. This model can then be used to determine the ways in which gene conversion may influence the diversity of duplicated microsatellites.

5.2.1, Model of mutation in duplicated microsatellites

At the time of initial duplication, both microsatellites would be expected to have the same number of repeat units. Under mutation alone, following duplication each microsatellite would be expected to mutate independently with each microsatellite potentially gaining or losing 1-3 repeats, through replication slippage (Carvalho-Silva et al. 1999). 2-step mutations are approximatly 10-fold lower in frequency than 1-step mutations (Brinkmann et al. 1998b). 3-step mutations are very rarely observed making it difficult to estimate a rate and they may be ~10x less frequent still.

As both microsatellite copies have the potential to gain or lose up to 3 repeats in any one mutation event (Brinkmann et al. 1998a), large length differences can potentially accumulate between duplicated microsatellite copies within an individual chromosome.

While mutation alone can create large differences between duplicated microsatellite copies there is also the potential for both copies to be homogenized through convergent mutation. Mutation alone can therefore produce many possible combinations of haplotype, and within a population the variability will depend on the mutation rate (Figure 5.2).

5.2.2, Problems in identifying conversion events between duplicated microsatellites

As gene conversion acts to homogenize two sequences, conversion events between duplicated microsatellites would be expected to result in each copy having the same repeat number (pseudohomozygosity). It could therefore be assumed that the observation of pseudohomozygous microsatellites within a chromosome where multiple chromosomes within the same haplogroup show a large difference in repeat number indicates that gene conversion has occurred.

As duplicated microsatellites are also subjected to the same mutation process as single-copy microsatellites, identifying conversion events is not simple. As for single-copy microsatellites, the mutation rate of duplicated microsatellites is expected to vary according to repeat number, array homogeneity, location and repeat size which will have different effects on microsatellite diversity both within a chromosome and across the Y phylogeny. The observation of a high proportion of pseudohomozygous chromosomes within a haplogroup could be the result of a low microsatellite mutation rate as well as gene conversion and distinguishing the two will be difficult. A high mutation rate will also complicate the identification of gene conversion events, as gene conversion would initially homogenize the two copies but mutation occurring soon after would effectively erase the evidence of gene conversion events. Similarly to mutation,

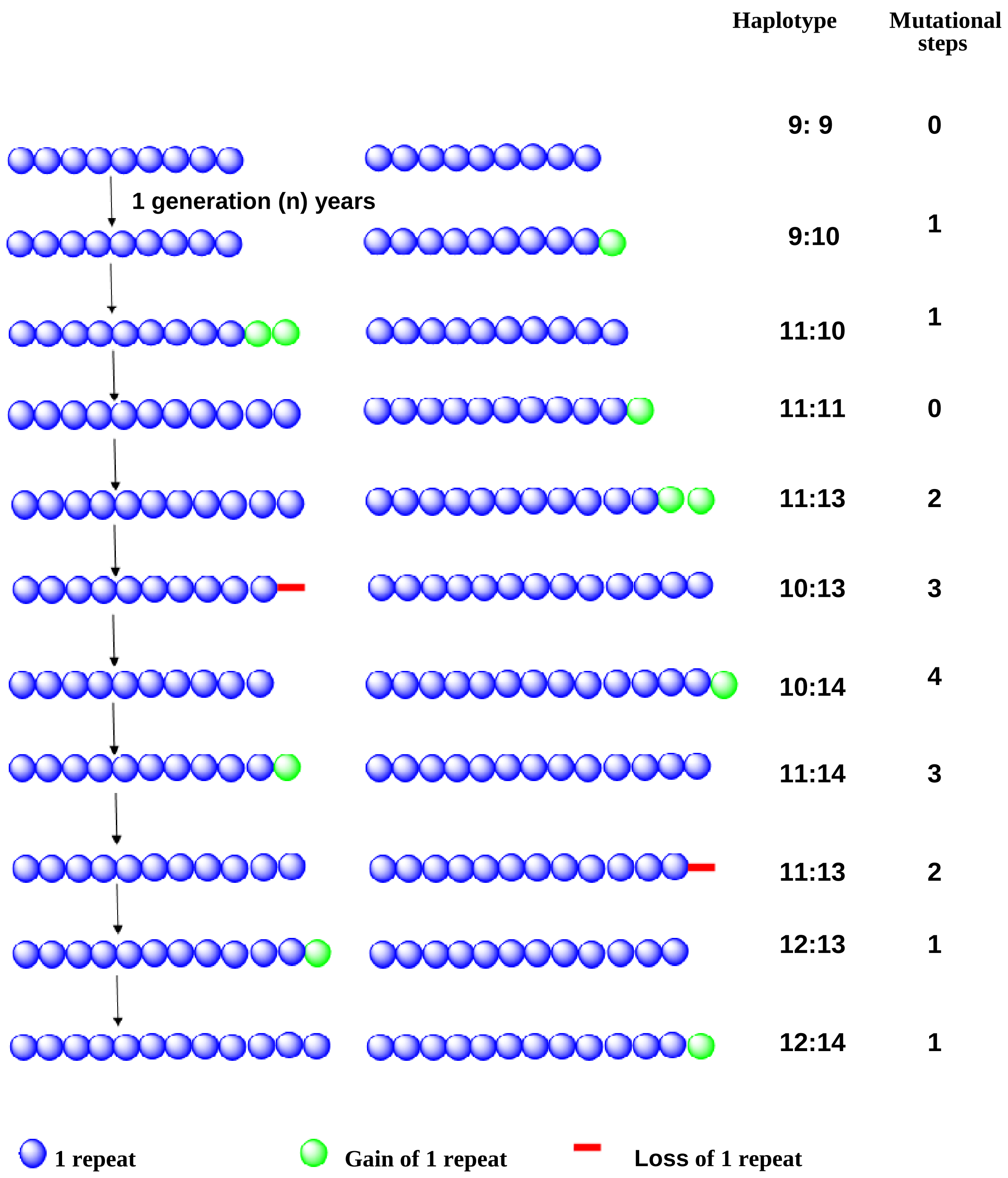


Figure 5.2: Hypothetical model of microsatellite mutation Following duplication each microsatellite is expected to mutate independently with the gain and loss of 1-3 repeats potentially occurring at each microsatellite. While mutation alone may create large differences between microsatellite copies there is also the potential for mutation to homogenise copies.

the rate of gene conversion can also have a number of effects on microsatellite diversity which may be difficult to distinguish from mutation. If conversion is very rapid, then allele size differences between copies will tend to be small and cannot easily be distinguished from the simple effect of low mutation.

5.2.2, Identifying conversion events between duplicated microsatellites

As gene conversion events are much harder to observe through typing duplicated microsatellites a set of criteria will be set out in which pseudohomozygosity can only reasonably be explained by gene conversion.

Unlike snPSVs which can show gene conversion when typed in chromosomes from across the Y phylogeny, duplicated microsatellites are more informative when typed in a large number of chromosomes from an individual haplogroup. In a set of chromosomes which show a large difference in repeat number, pseudohomozygosity is less easily explained by stepwise mutation alone, especially in the absence of the intermediate haplotypes. For example, in Figure 5.3A, in a set of chromosomes the CA(9:14) haplotype is observed along with the CA(9:9) and the CA(14:14) haplotypes. This distribution of haplotypes cannot be explained by mutation alone as the intermediate haplotypes such as CA(9:10) or CA(9:12) are not observed. In Figure 5.3B the intermediate haplotypes CA(9:11) and CA(9:12) are also observed and this distribution can now be explained by mutation. The ability to detect all possible haplotypes will depend on the number of chromosomes available for analysis and ideally a very large sample set would be analyzed in order to ensure that all haplotypes are observed. In very small sample sets it is possible that intermediate haplotypes may not be observed leading to the false identification of gene conversion events.

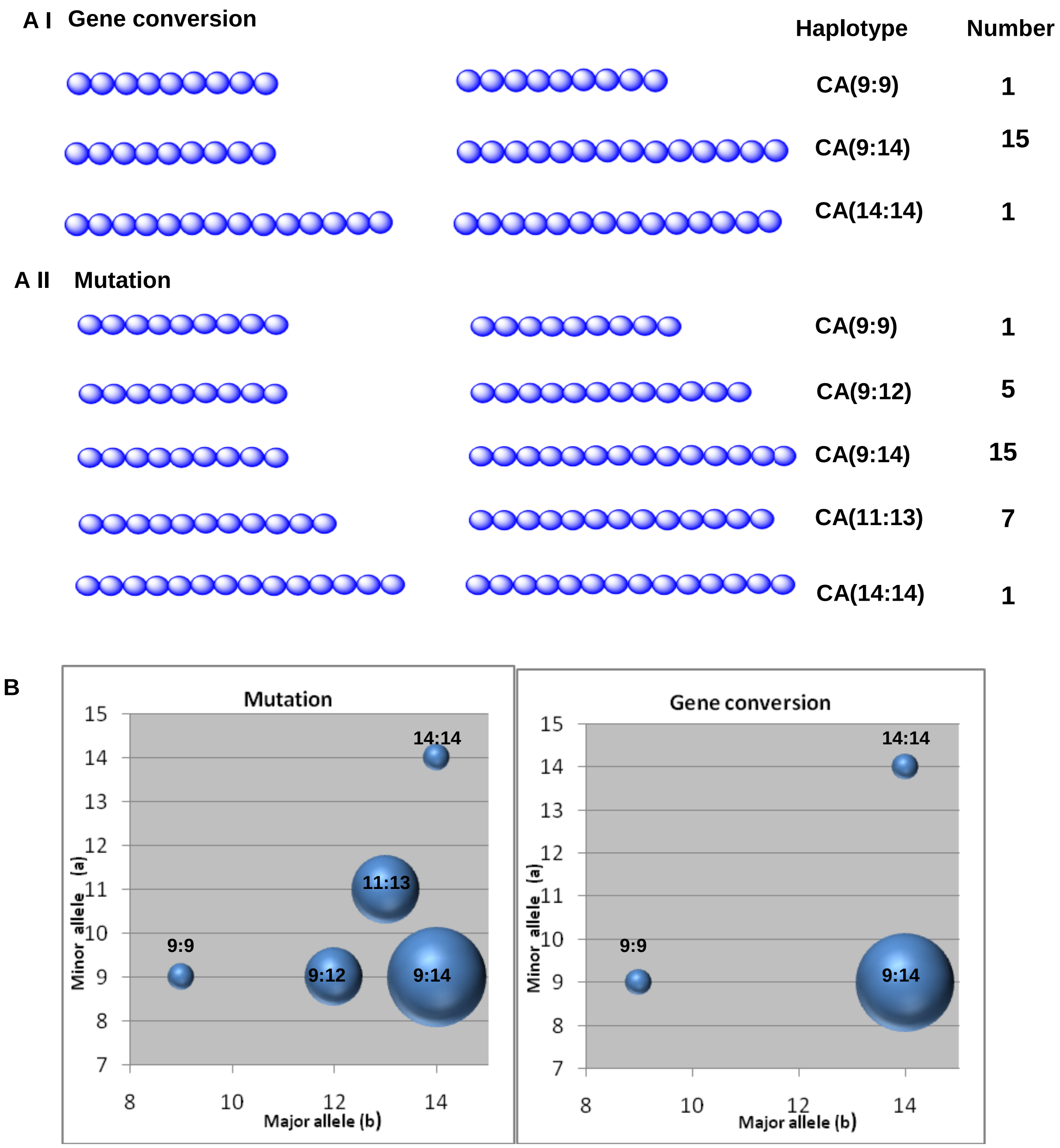


Figure 5.3: Identifying conversion events through typing duplicated microsatellites.
Gene conversion can be identified when pseudoheterozygous microsatellites CA(9:14) are observed along with pseudohomozygotes for the minor CA(9:9) or major CA(14:14) haplotypes **(AI).** In the presence of the intermediate haplotypes [CA(9:12), CA(9:13), CA(11:12)], pseudohomozygosity from gene conversion cannot be distinguished from pseudohomozygosity from mutation **(AII).** Microsatellite data can be displayed as bubble plots **(B)** where the X-axis represents the Major allele and the Y-axis represents the minor allele. The area of the bubble is proportional to the number of chromosomes carrying that genotype.

In order to identify haplogroups which have a large difference between copies, each duplicated microsatellite was initially typed in a panel of chromosomes from across the phylogeny. This not only allows the identification of haplogroups which may be of further interest but can also identify the microsatellites which have the most potential to show conversion. For example, a microsatellite which is largely pseudohomozygous across the phylogeny is likely to have a low mutation rate, and identifying conversion events would be difficult. Also, identifying conversion events from duplicated microsatellites which show only 1-2 repeats difference across the phylogerny would be difficult as pseudohomozygosity due to gene conversion could not be distinguished from pseudohomozygosity that is part of the normal allele distribution.

The ability to identify conversion events using microsatellites will also rely on having a large number of chromosomes from a well defined haplogroup available for analysis. In this study DNAs from the CEPH-HGDP panel (Cann et al. 2002) will be used for analysis; however, some haplogroups are not well represented or defined within the panel. For example, the haplogroup O(xO3e) could potentially contain chromosomes from haplogroups O*, O1, O2, O3 and their subclades which may reflect higher diversity within the sample set.

In this study duplicated microsatellites will be typed in a large set of chromosomes from a single well-defined haplogroup. When looking for evidence of gene conversion between duplicated microsatellites the observation of pseudohomozygous microsatellite copies without the intermediate haplotypes being observed will be taken as evidence of gene conversion.

5.3, Problems identifying conversion events between X-Y gametologous regions

While the same principles which apply to identifying gene conversion events between snPSVs and microsatellites on the Y chromosome also apply to X-Y homologous regions, identifying conversion events is complicated further by several factors. When typing snGSVs ascertainment bias is more of a problem, as many variant sites between the X and Y chromosomes appear to have become fixed during evolution. This means that many GSVs identified from an individual sequence will be uninformative while those that have undergone gene conversion will remain unidentified. For this reason resequencing of X-Y homologous regions will be carried out to distinguish uninformative GSVs from those that have the potential to show gene conversion.

While the haploid nature and well-defined evolutionary history of the Y chromosome make it an ideal tool for studying gene conversion, the diploid nature of the X chromosome introduces additional complications when identifying conversion events between the X and Y chromosomes. As the X chromosome is diploid, conversion events from the Y to the X chromosome may be lost during meiosis which can lead to conversion events being unidentified and lead to underestimation of the rate of gene conversion. Similarly, X-to-Y gene conversion events can also be lost from the Y chromosome due to genetic drift which is particularly strong on the Y chromosome. The presence of a polymorphic SNP on the X chromosome also complicates the study of gene conversion and may lead to the false identification of gene conversion events. When sequence alignments are carried out an X-chromosome SNP may falsely be identified as a snGSV with a derived allele of the SNP creating a snGSV between the X and Y chromosomes. During phylogenetic analysis the presence of a SNP which carries the ancestral allele may be mistaken as a gene conversion event. To address this problem dbSNP will be used to identify X chromosome SNPs and variation at these sites will be taken into account when looking for evidence of gene conversion. However it must be taken into account that some GSVs are known to be wrongly annotated as X-chromosome SNPs (Rosser et al. 2009; Cruciani et al. 2010). The use of the Y phylogeny is also limited in the study of gene conversion between the X and Y chromosome. During meiosis in females the X chromosomes have the ability to cross over and recombine along the entire length of the chromosome. While it is possible to establish an evolutionary phylogeny for segments of the X chromosome which lie between recombination points, the X chromosome as a whole cannot have a single defined phylogeny. The lack of defined evolutionary phylogeny for the X chromosome means that historical gene conversion events cannot be identified on the X chromosome and determining a direction of gene conversion is difficult.

When looking for evidence of gene conversion, identifying conversions from the X to the Y chromosome is relatively easy. Due to the known phylogeny of the Y chromosome, phylogenetic analysis can easily identify a variant which has arisen due to a conversion event from the X chromosome. The known time depth of the phylogeny also shows when the conversion event has occurred. On the X chromosome it is difficult to determine when a Y-to-X conversion event occurred. Due to the larger population size of the X chromosome which is three times that of the Y it is possible for a single conversion event to be observed in multiple descendant chromosomes leading to over estimation of a conversion event.

Chapter 6: Seeking evidence of gene conversion between the arms of Palindrome 6

6.1, Introduction

During meiosis, recombination between the X and Y chromosomes is restricted to the pseudoautosomal regions (PARs) while the remainder of the Y chromosome is considered to be non-recombining and male-specific. For many years lack of recombination within the male-specific region of the Y chromosome (MSY) led to the assumption that the Y chromosome would eventually become devoid of all genes and genetically inert (Graves, Koina, and Sankovic 2006). While gene conversion has previously been shown to occur between the *CDY* genes located in the arms of P1 (Rozen et al. 2003), evidence of gene conversion occurring between the other Yq palindromes has not been examined. Comparison of human and chimpanzee palindrome sequences has revealed lower interspecies divergence between palindrome arms than between the spacer regions leading to suggestions that gene conversion is conservative of the ancestral state (Rozen et al. 2003). As it is hypothesised that conservative gene conversion occurring between palindrome arms acts to maintain the function of spermatogenic genes, this raises the question as to whether gene conversion occurs in all palindromes or is restricted to those which contain genes. To address this question this Chapter will carry out an analysis of non-genic P6 in order to determine whether gene conversion is a phenomenon associated with Y chromosome palindromes in general and also determine whether gene conversion is conservative of the ancestral state.

From analysis of palindromes carried out in Chapter 4, over fifty PSVs were identified between the arms of P6 making it a good candidate to study gene conversion, while low overall divergence of 0.03% between palindrome arms is suggestive of a history of gene conversion. As P6 has previously been shown to predate human-chimpanzee speciation, the low divergence between palindrome arms could not be due to a recent duplication event in humans.

6.1.1, Chapter Aims

This Chapter will look for evidence of gene conversion occurring between the arms of P6. A phylogenetic analysis of snPSVs and microsatellites will be carried out in order to identify historical conversion events and to determine whether gene conversion is conservative of the ancestral state as previously suggested by Rozen et al. (2003).

6.2, Materials and methods

Analysis was carried out as described in Chapter 2, with the following exceptions.

6.2.1, Oligonucleotides

Oligonucleotides used in the analysis of P6 are detailed in table 6.1

6.2.2, DNAs used for C39 Microsatellite typing

Twenty genomic male DNAs from haplogroup N* were selected from a Bhutanese population sample (Parkin et al. 2006; de Knijff et al. 2009) (Supplementary table S6.1).

6.2.3, Microsatellite typing

PCR was carried out using 1-2µl of WGA DNA, the buffer of Jeffreys et al. (1990) and 1U Kappa Taq. PCR conditions included an initial denaturation step at 95°C for 3 minutes followed by 94°C 30s, 58°C 20s, 65°C 30s, for 20 cycles.

Table 6.1: Primers used in the analysis of palindrome 6.1

PCR Primers			
Primer name	**Forward primer 5´- 3´**	**Reverse primer 5´- 3´**	**Tm**
C6-7	AGACTGTCCACATAAGCTGG	CCCCAGACGACTGTCTGC	60
C10-11	GGCAAGTGATAGTGACATGG	CCCATCACTAGCTTCTGCG	60
C12-13	AGACAAAGAATCAGTAGAAGG	CATCTAGATGGCCTGCAGG	60
C24	TTGCTTCTAGTATTGTATTTGAAG	AATCTTAGACTGGATTAGTTTCC	60
C28	AGCTCATCTCCTATCTTCAACATATG	TCAGCCTATAGTCTCTCTATTCTGTGAC	60
C36	GAGAACAAGGCTGTGAAAATCTG	AGGTATAAAATGAGCAAATGAGGTG	60
C40	CAAAGGTAGACAAGATATATATCAATATCTCAG	CACAATGACTAATGTGTGAGAAAAGTC	60
Snapshot primers			
Primer name	**Forward primer 5´- 3´**		
P6_C6	AAAAA AGGGATATTGACTTTGATAA		
P6_C7	AAAAA AAAAA TTGTCCATCCAAGGACCAGA		
P6_C10	AAAAA GCTGGTCACAGAAAAGTGGA		
P6_C11	AAAAA AAAAA GACAGTGCTTAACAAGGTGG		
P6_C12	AAAAA AAAAA AAAAA TTTACTGAGATAAATGCATA		
P6_C13	AAAAA AAAAA AAAAA AAAAA TTGGTGGAATGTCATGAGGT		
P6_C24	ATATATACAATGTATACATA		
P6_C28	GTAGATTTAGTGTCCCGTGGG		
P6_C36	TATTCATTTTAATTTAATTT		
P6_C40	GTGTCCTCATCTCCTATTCT		
Microsatellite primers			
Primer name	**Forward primer 5´- 3´**	**Reverse primer 5´- 3´**	**Tm**
C25	8TTGCTTCTAGTATTGTATTTGAAG	AATCTTAGACTGGATTAGTTTCC	58
C34	6ATAAGTTAAGGAGCGTTTGTAC	TGTTCATATAAATGTATGTATTGG	58
C39	6CAAGTTCAAATCTGTATGAGAAC	ATATCCATTTCTAACTTCAGATTG	58
C47	6ACACATATCGTTAATTGTATACG	AGTGAGCCTTACTCAAGACC	58
C52	8TCCAACATGAGCAACACAGTG	CTCATCTGGGATTGTAATTCTC	58
C53	6TCTACGTTAATATTTCCATGTTAC	CTAGATTCTGTAAATATTAGGTAG	58
Sequencing primers			
Primer name	**Forward primer 5´- 3´**	**Reverse primer 5´- 3´**	**Tm**
P6_A1inner	acatcagttgttgcctcctc	ctcctatctctgcatggcctg	60
P6_A1outer	acatcagttgttgcctcctc	ctcctatctctgcatggcctg	60
P6_A2inner	acatcagttgttgcctcctc	ttgctgataaagacatatccaacac	60
P6_A2outer	ATATGGTATTTTCTATATATTTTTGGAATC	TTGGAGATGTCCTAAGTGGTTAAC	60

6.3 Results

6.3.1, Re-sequencing of palindrome boundaries

Interspecies sequence comparisons carried out in Chapter 4 revealed that homology observed between the distal arm of the human palindrome and both arms of the chimpanzee palindrome continued past the outer boundary observed in the human sequence alignment. This suggests that a deletion may have occurred in the proximal arm of the human sequence which has shortened the palindrome in humans. (Figure 6.1A). As previously discussed in Chapter 3, various rearrangements and deletions are known to occur on the Y chromosome which do not affect male fertility, and therefore it is possible that this deletion may be specific to the Y chromosome reference sequence, or to a particular phylogenetically related set of chromosomes. Rozen et al. (2003) also observed that outer palindrome boundaries are less well conserved between species than the inner boundaries which raises the question as to whether the boundaries of P6 are variable across the Y phylogeny. To address this question both the inner and outer boundaries were sequenced in eight individuals representing the major haplogroups of the Y phylogeny and were also compared to the corresponding regions in the CV, JW and Yh database sequences.

Sequence comparisons revealed both the inner and outer boundaries to be fixed in all haplogroups across the Y phylogeny, while no evidence of gene conversion at the boundaries was observed. This suggests that the deletion occurred following speciation but before the MRCA of the Y phylogeny and has persisted in the human lineage.

6.3.2, Sequence divergence

Divergence calculations carried out in Chapter 4 suggest that gene conversion has been occurring between the arms of P6 in both humans and chimpanzees (Figure 6.1B). Sequence divergence between palindrome arms for both species was calculated as 0.03%, despite the palindromes predating speciation. Interspecies divergence between the arms of P6 was calculated as 1.43% while divergence between the spacers was higher at 1.92%. The significantly lower interspecies divergence between the palindrome arms in comparison to the spacers (P=0.0001, 2 tailed Fisher exact test), which are non-duplicated and therefore cannot undergo gene conversion, suggests that conversion is conservative of the ancestral state. Conversions to the derived allele in either species would be expected to increase divergence between orthologs while conversions to the ancestral allele in either species would homogenise both sequences lowering the divergence between orthologs. These findings are similar to those of Rozen et al. (2003), which showed significantly lower interspecies divergence between palindrome arms in comparison to the spacers.

6.3.3, Seeking evidence of gene conversion through typing snPSVs

Direct evidence of historical gene conversion was sought through typing snPSVs identified between the arms of human P6. From comparison of all available sequences thirty snPSVs were identified and the ancestral state of each PSV was determined through comparisons with the chimpanzee reference sequence.

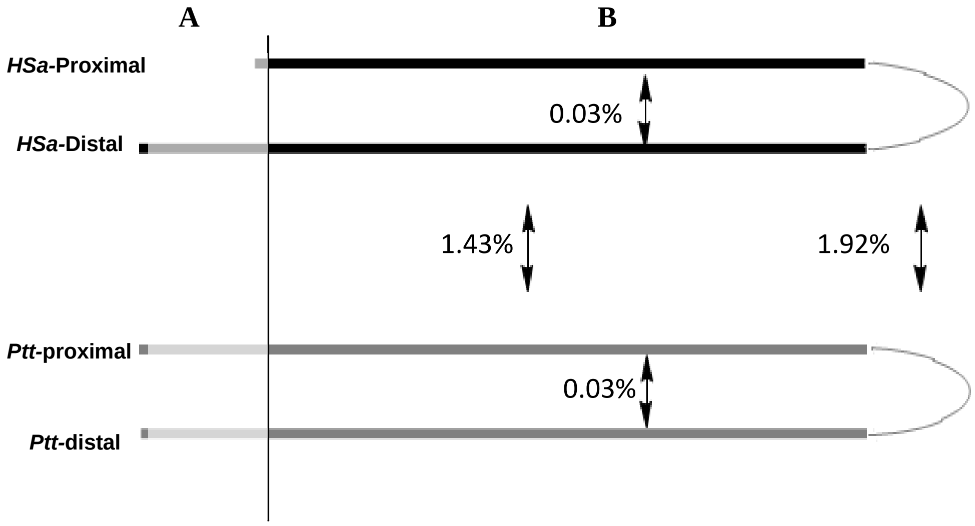


Figure 6.1: Interspecies sequence comparison of human and chimpanzee P6

a) Sequencing revealed both boundaries to be fixed across the phylogeny suggesting that the deletion which shortened the palindrome in humans occurred after speciation but before the MRCA of the Y phylogeny.

b) Low divergence of 0.03% between palindrome arms in both human (HSa) and chimpanzee (Ptt) suggests that gene conversion occurs in both species. Significantly lower interspecies sequence divergence of 1.43% is observed between palindrome arms in comparison to 1.92% observed between the spacer regions (P=0.000, 2-tailed Fisher exact test) which suggests that gene conversion is conservative of the ancestral state.

6.3.4, SnPSV typing using the SNaPshot assay

Of the thirty snPSVs identified, ten were typed using the SNaPshot (Applied Biosystems) mini-sequencing protocol in a panel of 64 males representing thirty haplogroups of the Y phylogeny (Figure 6.2). The remaining PSVs could not be typed as they were situated in highly repetitive regions and specific PCR amplification was unsuccessful.

Of the ten PSVs typed two (C28 & C36) appear to be confined to the reference sequence as private variants, and were hence uninformative across the Y phylogeny. One PSV (C13) represnts a CpG-to-TpG transition and as only the pseudoheterozygous and pseudohomozygous ancestral states were observed, unidirectional gene conversion cannot be distinguished from CpG hypermutation. In order to take a conservative approach to identifying gene conversion events, this PSV was not included in further analysis. Of the remaining seven PSVs, unidirectional conversion to the ancestral allele was observed at two (C11 and C40) while at the remaining five PSVs (C6, C7, C10, C12, C24) bidirectional conversion events to both the ancestral and derived alleles were observed. Two PSVs, C6 and C7, which are separated by 81bp in the reference sequence alignment, appear to have undergone co-conversion in eleven chromosomes with both PSVs simultaneously converting to either the ancestral or derived allele. For example, in Figure 6.2 in the haplogroup H1 chromosome C6 and C7 both undergo conversion to the derived allele while in the haplogroup L chromosomes both C6 and C7 convert to the ancestral allele. Of the eleven chromosomes which appear to undergo co-conversion, in eight chromosomes both C6 and C7 co-convert to the ancestral allele while in the remaining three chromosomes both PSVs converted to the derived allele suggesting that C6 and C7 may have been contained within the same

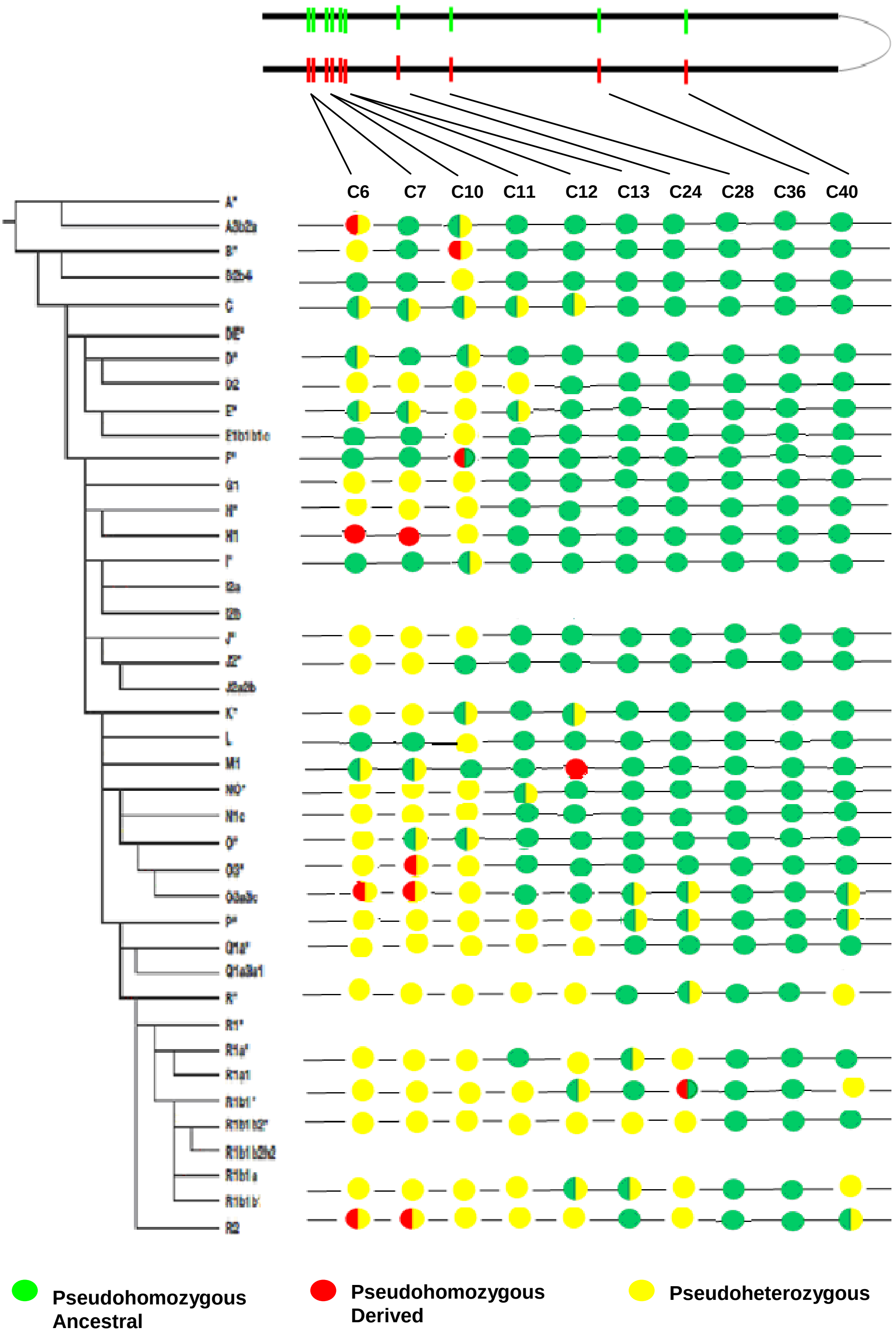


Figure 6.2: Phylogenetic analysis of snPSVs identified between the arms of P6
From phylogentic analysis 62 conversion events were identified, 53 of which returned a PSV to the ancestral allele while 9 converted to the derived allele. These data provides significant evidence (P=0.0001 chi square test) of conservative gene conversion occurring between the arms of P6.

historical conversion tract. However this may not necessarily be the case as in two chromosomes one from haplogroup O* and one from haplogroup O3* C6 remains in the pseudoheterozygous state while C7 undergoes gene conversion which suggests that the two PSVs may not always lie within the same conversion tract.

As C6 and C7 are suspected to be contained within the same conversion tract, chromosomes where both PSVs convert to the same allele will be counted as one conversion event. In total sixty-two separate conversion events were identified, fifty-two of which returned the PSV to the ancestral allele while ten converted the PSV to the derived allele. These data support the interspecies divergence calculations and provide evidence that significantly conservative gene conversion occurs between the arms of P6 ($P>0.0001$ using the Chi square test).

6.3.5, Determining gene conversion tract length

As the C6 and C7 PSVs appear to lie within the same conversion tract, re-sequencing of a 1.2-kb region surrounding both PSVs was carried out to see if additional PSVs could be identified and an estimation of tract length made. Measuring conversion tracts from snPSVs is difficult especially when typing PSVs identified from a single sequence, due to the ascertainment bias. It is possible that additional PSVs may lie within the regions surrounding C6 and C7 which have not been identified from the reference sequence alignment, and if this is the case it is possible that C6 and C7 might be shown to be contained within two separate conversion tracts which would increase the number of conversion events observed. Due to the complete sequence homology surrounding snPSVs when a conversion event is observed at the site of a snPSV it is not known how much of the flanking sequence is involved in the conversion tract. While a tract could include only a single nucleotide, it could also contain stretches of sequence which span several hundred nucleotides; however, a conversion event will only be observed at the PSV site. This means that tract length cannot be definitively determined and only estimations of tract length can be made (Jeffreys and May 2004). In this study maximum tract length will be estimated by observing a conversion event at the site of a PSV which is flanked by two PSVs which remain in the pseudoheterozygous state. The maximum conversion tract will be the distance between the two pseudohomozygous PSVs, while the minimum conversion tract will be the single base where the conversion event is observed. The two PSVs which are closest to C6 and C7 in the reference sequence alignment are situated >1kb either side of C6 and C7 and as maximum conversion tracts have been reported to be less than 1kb in length (Chen et al. 2007), these two PSVs are unlikely to be informative when determining tract length. As it is also possible that other PSVs are located within this region which would be useful in determining tract length, a 1.2-kb segment surrounding C6 and C7 was sequenced in twenty chromosomes representing fourteen different haplogroups, including the eleven chromosomes in which C6 and C7 appear to have undergone co-conversion.

Sequencing of the 1.2-kb region surrounding C6 and C7 identified three additional snPSVs which were situated in a 500bp region encompassing C6 and C7, while no additional PSVs were identified in the 81-bp region separating C6 and C7 (Figure 6.3). Two of the additional PSVs were located 71bp and 88bp either side of the region bounded by C6 and C7, while the third PSV was located 327bp distal to C6. In fourteen chromosomes a maximum conversion tract could not be estimated as a conversion event

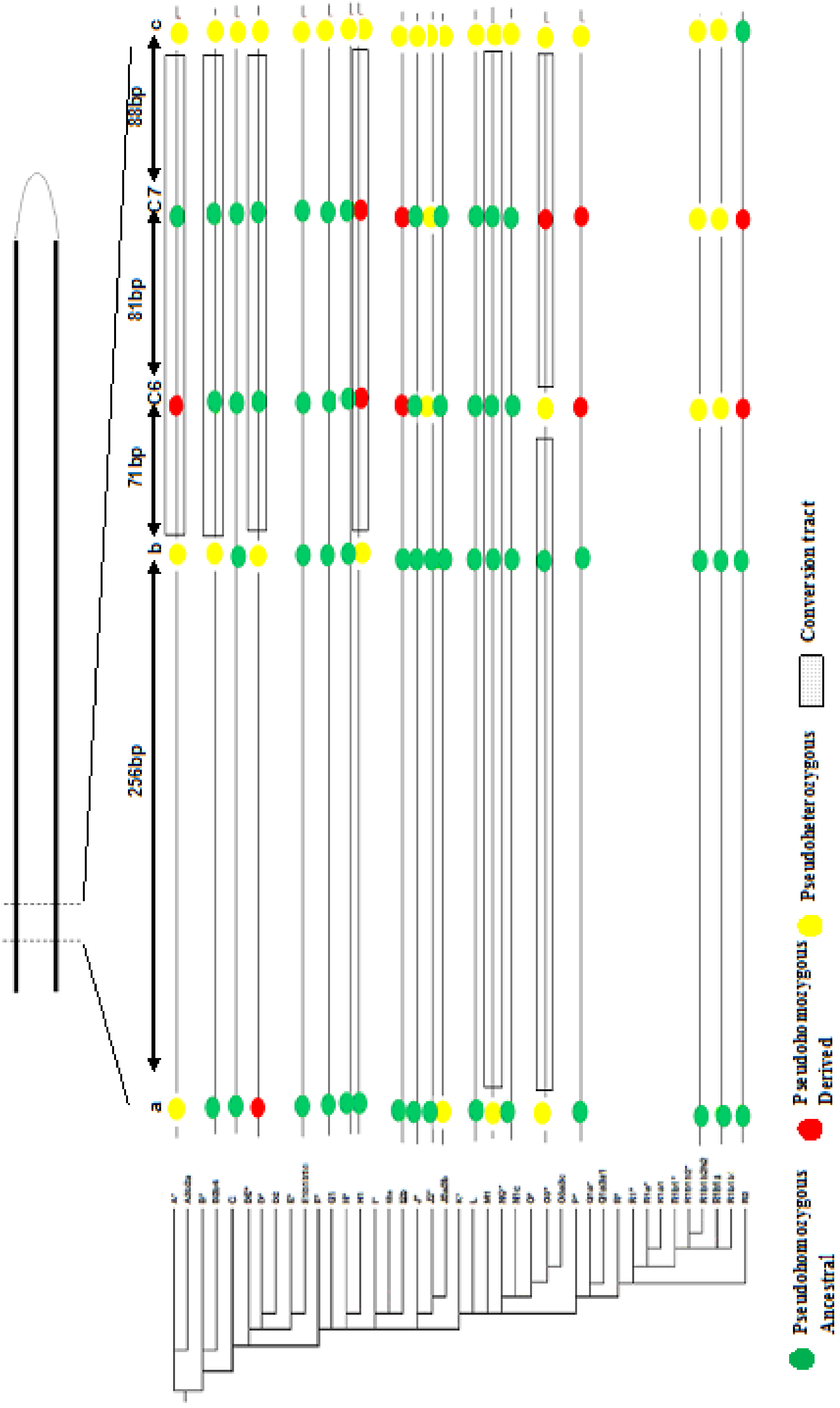

a
b
c
C6
C7
256bp
71bp
81bp
88bp
Pseudohomozygous Ancestral
Pseudohomozygous Derived
Pseudoheterozygous
Conversion tract

Figure 6.3: Estimation of gene conversion tract length in the regions surrounding C6 and C7
Re-seaquencing of a 1.2-kb region in 20 chromosomes identified three additional PSVs (a-c) within 500bp encompassing C6 and C7, all of which appear to be undergoing gene conversion. From this extended sequencing conversion tracts appear to vary between haplogroups as well as between individuals within a haaplogroup. The maximum estimated conversion tracts ranged between 163bp and 496bp.

was not flanked by two pseudoheterozyogus PSVs. For five chromosomes a single conversion tract was identified, which in four cases ranged from 1-240bp (haplogroup A(xA3b2a), B2b4, D and H1) and in one case from 1-496bp (haplogroup M1). In the remaining chromosome (Haplogroup O3*) two conversion tracts ranging from 1-160bp and 1-327bp were identified.

Re-sequencing of this region overcomes some of the ascertainment bias associated with typing snPSVs, and although the major branches of the phylogeny have been covered it is possible that there are additional PSVs which have not been identified through sequencing. This analysis suggests that conversion tracts in this region of the palindrome are not of a fixed length; however, analysis of more chromosomes may be needed in order to identify additional PSVs and further define tract length.

6.3.6, Seeking evidence of gene conversion through microsatellite typing

Analysis of snPSVs has revealed that frequent gene conversion has occurred between the arms of P6 in recent human evolution and provided significant evidence that gene conversion is conservative of the ancestral state. However, typing snPSVs identified from a single reference sequence creates an ascertainment bias because only sites which are variable within the sequence will be identified as regions which will potentially show conversion events. Since gene conversion acts to homogenize two sequences, nucleotides which have undergone historical conversion will not be identified from alignment of a single chromosome sequence and many conversion events will be missed. Typing microsatellites provides an additional source of variation, and alleviates the ascertainment bias associated with typing snPSV. As microsatellites have a higher mutation rate and are multiallelic they are therefore expected to be independently variable between haplogroups as well as between individuals within a haplogroup. This section will seek evidence of gene conversion through typing duplicated microsatellites which lie within the arms of P6.

6.3.7, Analysis of microsatellite variability

From alignment of the reference sequence, nine duplicated microsatellites which are variable in length between copies were identified, while an additional three which were not variable in the reference sequence but likely to be informative in additional haplogroups, were identified using Tandem Repeats Finder (Benson 1999) using default settings.

To gain an idea of the variability, six of the twelve P6 microsatellites identified from the reference sequence were successfully typed in a panel of 64 Y chromosomes representing thirty haplogroups of the Y phylogeny (Figure 6.4). The remaining six microsatellites could not be typed as they were either complex, consisting of more than one type of repeat, or located in repetitive regions and specific PCR amplification was not possible. Following preliminary analysis, microsatellites that were considered most likely to be informative were subjected to further typing. Of the six microsatellites typed, C34 and RM52 showed only 1-3 repeat differences between copies across the Y phylogeny with approximately 40% of chromosomes for C34 and 64% for RM52 carrying pseudohomozygous haplotypes. Considering the high frequency of pseudohomozygotes and the small numbers of repeat differences between copies, identifying conversion events would be difficult as pseudohomozygosity due to gene conversion could not be distinguished from pseudohomozygosity which exist as part of

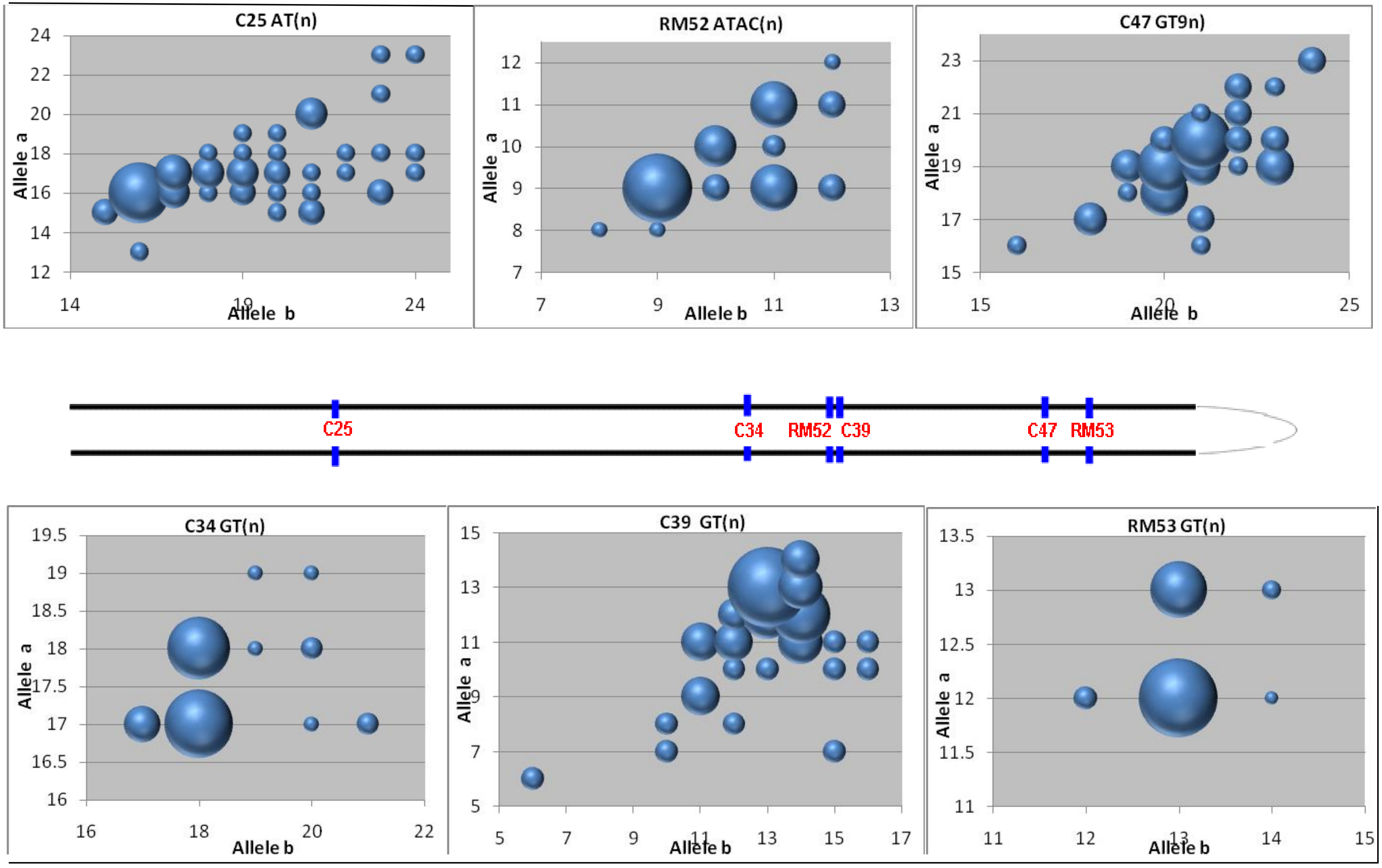

C25 AT(n)
Allele a
Allele b
24
22
20
18
16
14
12
14
19
24
RM52 ATAC(n)
Allele a
Allele b
12
11
10
9
8
7
7
9
11
13
C47 GT9n)
Allele a
Allele b
23
21
19
17
15
15
20
25
C25
C34
RM52
C39
C47
RM53
C34 GT(n)
Allele a
Allele b
19.5
19
18.5
18
17.5
17
16.5
16
16
18
20
22
C39 GT(n)
Allele a
Allele b
15
13
11
9
7
5
5
7
9
11
13
15
17
RM53 GT(n)
Allele a
Allele b
13.5
13
12.5
12
11.5
11
11
12
13
14
15

Figure 6.4: Phylogenetic analysis of P6 microsatellites.

Microsatellite data has been displayed as bubble plots where the X-axis represents the major allele and the Y-axis represents the minor allele. The area of the bubble is proportional to the number of chromosomes carrying that genotype. The C34 and RM53 microsatellites displayed only 1-3 mutational differences between copies across the phylogeny and as gene conversion cannot easily be distinguished from mutation, these microsatellites were excluded from further typing. Microsatellites C25, RM52, C39 and C47 were more variable across the phylogeny with up to 7 repeat unit differences being identified between microsatellite copies.

the natural distribution of alleles. For these reasons C34 and RM52 were excluded from further analysis. Of the remaining microsatellites, C25, C39, C47 and RM53 were more variable across the Y phylogeny with 3-7 repeat differences being observed between microsatellite copies. As these microsatellites are generally more variable and display large differences in repeat number in some haplogroups, conversion events could potentially be identified.

Analysis of phylogenetic data for the C25 microsatellite showed one individual from haplogroup B2b4 to carry the AT(18:24) haplotype giving a 6-repeat difference between copies, while one individual from haplogroup P* carried the AT(16:23) haplotype giving a 7-repeat difference between copies. Haplogroups B2b4 and P* are not well represented in the CEPH-HGDP panel with each haplogroup being represented by only six chromosomes. This small sample set would not be sufficient to distinguish pseudohomozygosity resulting from gene conversion from pseudohomozygosity which is part of the natural distribution of alleles and therefore C25 was excluded from further analysis. RM53 was also excluded from further analysis, as despite showing a four-repeat difference between copies in chromosomes from haplogroup M1, only eight chromosomes are available in the CEPH-HGDP panel which would not be sufficient to identify gene conversion events. Phylogenetic data for the C47 microsatellite showed haplogroup O3* to carry the TG(19:23) haplotype, giving a 4-repeat difference between copies. Although haplogroup O* is represented by forty chromosomes in the CEPH-HGDP panel the haplogroup is not well defined and could potentially contain a mixture of chromosomes from multiple haplogroup O sub-clades. As the founder of each haplogroup O sub-clade could potentially carry different microsatellite haplotypes, this could lead to increased diversity between alleles and complicate the interpretation of conversion events. For this reason C47 was also excluded from further analysis.

Of the six microsatellites typed, only C39 was selected for further typing. From preliminary analysis of 64 chromosomes, two chromosomes from haplogroup N(xN1c) were shown to carry the TG(13:18) haplotype giving a 5-repeat difference between copies which offers the potential to identify gene conversion events. Although haplogroup N(xN1c) is not well defined or represented in the CEPH-HGDP panel and 45 chromosomes from the Bhutanese population belonging to haplogroup N* (Emma Parkin, unpublished observations) were available for analysis. Although it is not known from preliminary typing which haplogroup N subhaplogroup carries the TG(13:18) haplotype evidence of gene conversion was sought in these chromosomes.

Of the twenty haplogroup N* chromosomes analysed (Figure 6.5), 8 were observed to carry the TG(13:18) haplotype, as observed in preliminary typing, while 1 chromosome were observed to carry the TG(13:17) haplotype leaving a four-repeat difference between copies. However, a more interesting finding was 2 chromosomes carrying the pseudohomozygous TG(13:13) haplotype. This is especially striking, as intermediate haplotypes (i.e. 13,16, 13,15, 13,14) were not observed within the sample set. In the absence of the intermediate haplotypes this 4-repeat difference cannot easily be explained by mutation alone and suggests that gene conversion could have resulted in the production of the TG(13:13) haplotype. As 8 chromosomes were also observed

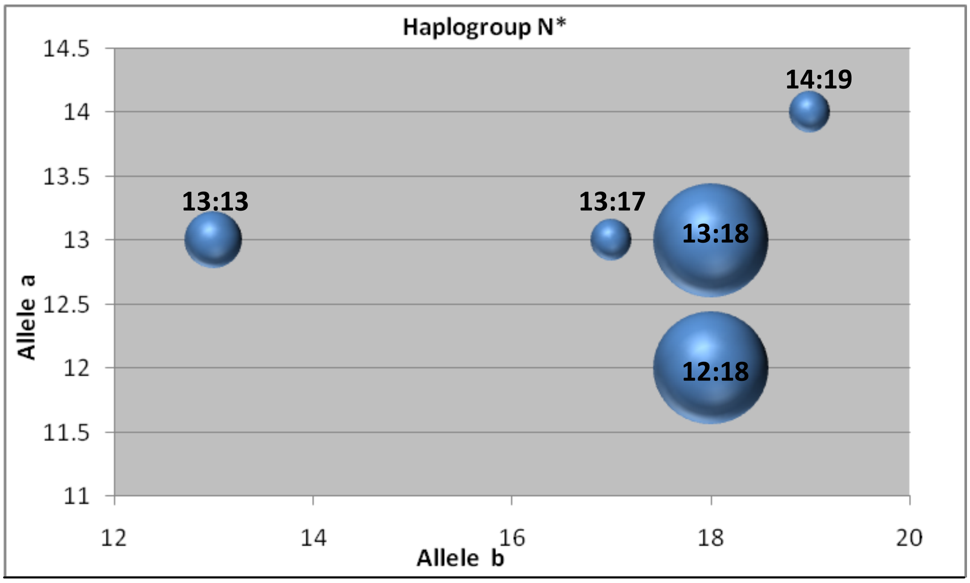


Figure 6.5: Typing of the C39 microsatellite in hg N* chromosomes.
8 individuals were observed to carry the TG(13:18) haplotype while 1 individual carries the TG(13:17) haplotype, leaving a gap of 4 repeats in the haplotype distribution. Two individuals were observed to carry the pseudohomozygous TG(13:13) haplotype which cannot readily be explained by mutation alone.

to carry the TG(12:18) haplotype and 1 chromosome to carry the TG(14:19) haplotype it appears that mutation is also influencing microsatellite diversity. From typing duplicated microsatellites within the arms of P6 only two possible conversion events have been identified; however, as the two chromosomes which carry the TG(13:13) haplotype are from the same population this haplotype could be the result of a single conversion event, carried identical by descent. To test this further, two phylogenetic networks were constructed based on the haplotypes of each chromosome. One network was constructed based on data obtained for 22 non-duplicated microsatellites which had previously been typed for each haplogroup N* chromosome (Figure 6.6A) while the second network also included data from the duplicated C39 microsatellite which was obtained in this study (Figure 6.6B). A weighting scheme was employed as described by Qamar et al. (2002) with specific weights assigned to each microsatellite based on the variance observed among chromosomes within the sample set.

In the first network which was constructed based on the haplotye of 22 non-duplicated microsatellites, the two chromosomes which carry the C39 TG(13:13) haplotype are both contained within the same node (red) and are separated from the nearest node (or haplotype) by only two mutational steps. As both chromosomes carrying the TG(13:13) haplotype also carry the same haplotype based on the 22 non-duplicated microsatellites it is likely that both share a recent common ancestor. When the C39 microsatellite is also included in the network, the two chromosomes carrying the TG(13:13) haplotype are still situated within the same node; however, this node is now separated from the nearest node by six mutational steps, with the four additional mutational steps coming from the C39 microsatellite. This cannot be explained by mutation alone and suggests

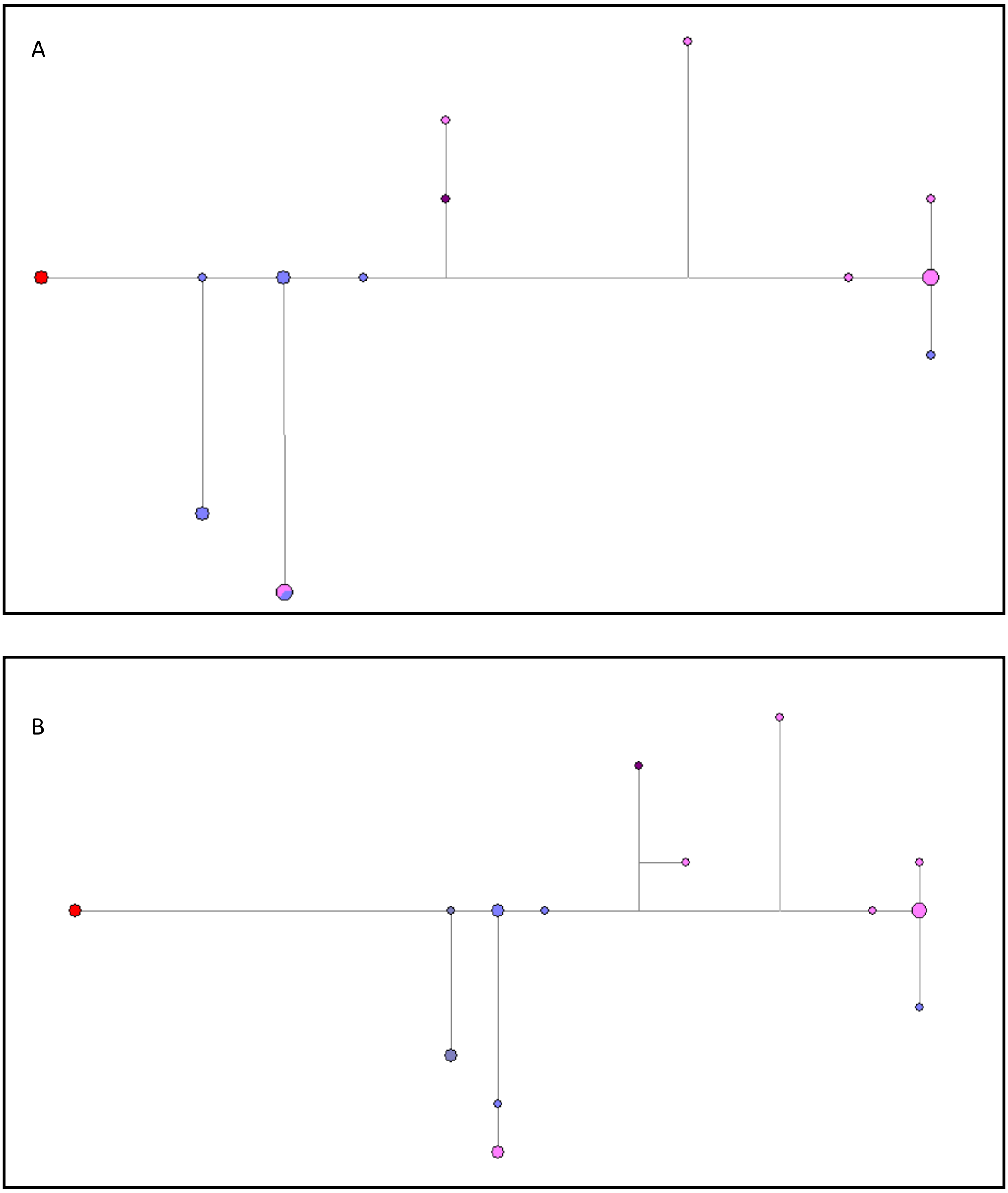


Figure 6.6: Phylogentic networks of C39 haplotypes

a) Network based on microsatellite data from 22 non-duplicated microsatellites and nodes coloured based on the C39 haplotypes observed for each individual (Red = 13:13, Pink = 13:18 Blue = 12:18 and purple = 14:19). The two individuals which carry the TG(13:13) haplotype are contained within the same node which shows that the two individuals carry the same haploype based on the 22 microsatellites typed.

b) Network including data for the C39 haplotype shows the two individuals carrying the TG(13:13) haplotype to now be separated from the remainder of the network by 6 mutational steps which cannot be explained by mutation alone. As the two individuals which carry the TG(13:13) haplotype are contined within the same node it appears that the TG(13:13) haplotype has arisen as the result of a single gene conversion event.

that the two chromosomes carrying the TG(13:13) haplotype probably arose from a single gene conversion event. Typing six duplicated microsatellites situated within the arms of P6 has identified only one gene conversion event; however, it has already been established through typing snPSVs that gene conversion occurs much more frequently. The variable mutation rate of microsatellites greatly complicates the identification of conversion events with the majority of microsatellites being excluded from further analysis as conversion could not be distinguished from mutation. It is highly likely that gene conversion occurs much more frequently but many conversion events have been masked by subsequent step-wise mutation

6.3.8, Estimating the rate of gene conversion

Determining the rate of gene conversion occurring between two paralogs is not straight forward and the true rate of gene conversion cannot easily be determined. The known time depth of the Y phylogeny allows the period of time over which gene conversion events have occurred to be determined and the rate of gene conversion to be estimated.

From analysis of snPSVs carried out in this study >70 conversion events have been identified: however due to the ascertainment bias associated with indentifying PSVs it is highly likely that gene conversion is occurring more frequently than observed in this study. These ascertainment biases have led to problems with estimating the rate of gene conversion occurring between the paralogs of P6. As all of the snPSVs typed in this study have been identified from a single reference sequence, PSVs which have arisen within a different haplogroup but undergone gene conversion in the reference sequence chromosome will have remained unidentified which will lead to underestimation of the rate of gene conversion. This ascertainment bias has been overcome by reseqencing of a

1-kb segment encompassing the C6 and C7 PSVs; however, this has also created an additional bias as only a small region of the palindrome which has been shown to be undergoing frequent gene conversion has been analyzed. Analysis of such a small segment of the palindrome would not necessarily be sufficient to determine the rate of gene conversion across the entire palindrome. An additional problem when trying to determine the rate of gene conversion is that each PSV has arisen within different branches of the Y phylogeny so the number of generations over which conversion events have occurred varies between PSVs.

Due to the various ascertainment biases involved in this study it has not been possible to determine the rate of gene conversion between the arms of P6. To determine the rate of gene conversion further analysis of P6 would need to be carried out to identify PSVs from additional Y chromosomal haplogroups. The rate of gene conversion could only be estimated by sequencing either the whole palindrome, or a large segment, in diverse Y chromosomes representing haplogroups from across the Y phylogeny. Ideally the "phase" of each PSV would also be determined to establish any biases in the direction of gene conversion and the rate of gene conversion for each direction.

6.4 Discussion

Rozen et al. (2003) have previously suggested that gene conversion occurring between the arms of palindromes is conservative of the ancestral state, a mechanism which they suggest maintains the function of spermatogenic genes. This was determined from the observation that interspecies divergence between palindrome arms is significantly lower than that between the spacer regions, which are non-duplicated and cannot undergo gene conversion. While conservative gene conversion has been hypothesized no evidence has previously been reported to show that this bias has been active during evolution.

From analysis of the reference sequence, low sequence divergence of 0.03% between the arms of human P6 is suggestive of gene conversion while interspecies divergence was shown to be significantly lower (P=0.0001, 2 tailed Fishers exact test) between palindrome arms (1.43%) than between the spacer regions (1.92%) suggesting that gene conversion in P6 is conservative of the ancestral state. Direct evidence of gene conversion was observed through typing snPSVs identified from alignment of the available database sequences. Phylogenetic analysis of ten snPSVs identified sixty-two separate conversion events, with significantly more conversion events returning a PSV to the ancestral allele (52) than to the derived allele (10). This difference was shown to be statistically significant (P=0.0001, Chi square test) providing direct evidence of conservative gene conversion occurring between the arms of P6.

Despite frequent gene conversion being identified through typing snPSVs only one conversion event was identified from typing duplicated microsatellites. Identifying gene conversion events in this way is much more complicated than from typing snPSVs.

While the mutation rate at single nucleotides is relatively constant the mutation rate of microsatellites is much more variable and is dependent on factors such as repeat size, length and location. For example, larger microsatellites mutating faster than shorter microsatellites. While only one conversion event has been identified here, gene conversion is likely to be occurring much more frequently than observed and the variable mutation rate has complicated the identification of conversion events. While the observation of many pseudohomozygous haplotypes, within a sample set is suggestive of gene conversion it may also be due to a slow or non-mutating microsatellite. Additionally microsatellites which display only 1-2 repeat difference in haplotype distribution between copies could be influenced solely by stepwise mutation, but it is also possible that gene conversion has produced pseudohomozygous haplotypes which have undergone subsequent mutation. This study provides a single observation in which gene conversion between duplicated microsatellites has propagated the smaller allele; however more analysis would need to be carried out to determine whether there is any systematic bias in this process. As the mutability of microsatellites increases with increased repeat number, gene conversion to the smaller allele could in principle act to stabilise large microsatellites.

This study supports previous work carried out by Rozen et al. (2003) and provides direct evidence of conservative gene conversion occurring between palindrome arms during human evolution. This study also provides evidence that gene conversion is a feature of Y chromosome palindromes in general and is not restricted to regions that contain genes. While conservative gene conversion occurring in palindromes which contain genes would preserve gene function, in palindromes which do not contain genes the process may preserve the ancestral sequence and maintain a high degree of sequence similarity preventing the Y chromosome from decay.

Chapter 7: Seeking evidence of gene conversion between the arms of Palindrome 8

7.1, Introduction

Analysis of P6 carried out in chapter 6 has provided significant evidence of conservative gene conversion occurring between the arms of P6 indicating that conservative gene conversion is not limited to palindromes that contain genes. It has previously been suggested that gene conversion maintains the function of spermatogenic genes (Rozen et al. 2003) and therefore it is possible that conversion may occur at a faster rate in palindromes which contain genes. P8 is a useful tool to seek evidence of differences in the rate of gene conversion between genic and non-genic palindromes as it contains the *VCY* genes which are located approximately 1.5-kb from the outer boundaries. While gene conversion has previously been reported between the *VCY* genes and its X-gametoolog *VCX*, evidence of gene conversion is not known to have previously been sought between the *VCY* genes themselves or the remainder of P8.

Analysis of P8 sequences carried out in Chapter 4 shows low overall divergence of 0.03% between the palindrome arms, which is suggestive of gene conversion. Using PCR assays targeted at the inner and outer boundaries, Rozen et al. (2003) have previously shown P8 to be absent in bonobo and gorilla but present in chimpanzee and therefore to predate human-chimpanzee speciation, demonstrating that this low sequence divergence cannot be attributed to a more recent duplication event.

7.1.1, Evidence of *VCX- VCY* gene conversion

It has recently been reported that *VCY* acts as a sequence acceptor from *VCX* during gene conversion (Trombetta et al. 2009) and gene conversion was estimated to occur at a rate of 3.6x10^{-6} X-to-Y conversion events per base per generation. This is slower than the rate of gene conversion observed between Y chromosome palindromes (Rozen et al. 2003) but similar to the rate of 3.8x10^{-8} to 1.7x10^{-6} observed between the *PRKX-PRKY* genes (Rosser et al. 2009). Analysis of the *VCX* and *VCY* genes carried out in Chapter 4 suggests that gene conversion may occur between the *VCY* gene copies as well as between *VCY* and *VCX*.

7.1.2, The Myers motif

In 2008 Myers et al. reported evidence of NAHR occurring within hotspots adjacent to the *VCX* genes. These hotspots contain one copy of a *VCX* gene and a 13bp tandem repeat of CTCCCTCCCCAC which has become known as the Myers motif. NAHR has been shown to occur between directly oriented *VCX3A* and *VCX* resulting in deletion of the *STS* gene (Myers et al. 2008) which has been shown to cause X-linked ichthyosis (Van Esch et al. 2005) Fine mapping of four breakpoints has revealed that all occurred precisely within the motif-rich tandem repeats (Myers et al. 2008). Evidence of NAHR occurring between the *VCX* genes raises the possibility that gene conversion, which is also mediated by NAHR, may occur between the *VCX* genes and the surrounding regions. As the region containing the *VCX* genes exhibits high sequence similarity with P8 it is possible that this sequence motif is also present in the P8 sequence and that NAHR may occur between the arms of P8 as well as the X and Y chromosomes.

7.1.3, Xp-Yq translocations

Chromosomal rearrangements such as translocations which are mediated by NAHR are known to occur between the X and the Y chromosomes. X-Y chromosome translocations are not a common occurrence in humans with only approximately fifty cases having been identified by 1991 (Yen et al. 1991). Not all Xp-Yq translocations are detrimental and while some have been shown to be sporadic events others appear to have been inherited (Jacobs et al. 2004). When analyzed cytogenetically the majority of these translocations have been shown to contain breakpoints at Xp22 and Yq11 producing a monocentric X chromosome which lacks the region distal to Xp22 and is fused to the region distal to Yq11 (Ferguson-Smith et al. 1992). While some translocations do not produce a phenotype other translocations are known to be responsible for conditions such as sex reversal and mental retardation (Devriendt et al. 2001). X-Y translocations in females produce a less severe phenotype than in males. Translocations between Xp22 and Yq11 in females tend to produce only short stature due to haploinsufficiency of the SHOX genes (Hattori et al. 2002). In contrast, translocations in males produce a more severe phenotype due to absence of X-specific genes between the breakpoint and pseudo-autosomal boundary which results in haploinsufficency of the PAR1 genes (Devriendt et al. 2001). The severity of the phenotype depends on the extent of the Xp deletion (Van Esch et al. 2005), which can result in growth retardation, developmental delay, partial ichthyosis and facial dysmorphism, (Van Esch et al. 2005).

As translocations are known to result from NAHR between the regions which contain the *VCX* and *VCY* genes it is possible that NAHR may also mediate gene conversion between the *VCX* and *VCY* genes and the surrounding regions. Gene conversion has previously been observed at other known Y chromosome translocation hotspots with Rosser et al. (2009) providing evidence of gene conversion between the *PRKY* and *PRKX* genes at the site of a known translocation hot spot.

7.1.4, Chapter aims

This chapter aims to identify gene conversion events between the arms of P8 through phylogenetic analysis of snPSVs. It will also aim to determine whether gene conversion occurs more rapidly in palindromes that contain genes. Evidence of gene conversion between the *VCY* genes and between the X/Y homologous *VCX* and *VCY* genes will also be sought.

7.2, Materials and methods

Analysis was carried out as described in chapter 2, with the following exceptions.

7.2.1, Oligonucleotides

Oligonucleitides detailed in table 7.1 were used in the analysis of P8.

7.2.1.1, *VCX-VCY* sequencing

Twenty-six genomic DNAs representing 30 haplogroups of the Y phylogeny were selected for sequencing.

7.2.1.2, Control DNAs

DNAs from human-rodent hybrid cell-lines containing the X or Y chromosome as the only human DNA content were used to test X- and Y-specificity. To test for X-chromosome specificity the X-only human-rodent hybrid DNAs ThyB-X (Dahlberg et al. 1983) and MOG13.9 (Povey et al. 1980) were used. To test for Y chromosome specificity the Y-only human-rodent hybrid DNAs 853 (Burk and Smith 1985), Q988-8 (Emrie et al. 1988) and 3E7 (Marcus et al. 1976) were used.

7.2.2, Chromosome-specific PCR

Chromosome-specific primers were designed based on sequence differences between the X and Y chromosomes which have been shown to have become fixed between chromosomes. Cycling conditions included an initial denaturation step at 95°C for 3mins followed by 25 cycles at 94°C for 30s, 62°C 30s, 70°C 30s. To ensure PCR specificity X- and Y-only hybrid DNAs were included as controls and products were run on a 2.5% (w/v) agarose gel for verification.

Table 7.1, Table of primers used in analysis of P8.

PCR Primers			
Primer name	**Forward primer 5´- 3´**	**Reverse primer 5´- 3´**	**Tm**
C1	GGCCATGATTTCAGGAACAC	CCTCTGCCAGGAAGTAGCAC	60
C2	AGGTTTGTTTTGCATTGGCG	AGACTCAGTGGGCGTAGGACC	60
Arm1_outer	TCTGGTACCCAGCCATCACG	TTAAAGAGCTAAGCATGAGATGATG	60
Arm1_inner	ACATGGATAAAATAGGGTGCAGAC	CTCAAGCTATCAGAGAAAATCTTGG	60
Arm2_outer	ATTTGCTATCTGAGACAGATTGTGAC	TTAAAGAGCTAAGCATGAGATGATG	60
Arm2_inner	ACATGGATAAAATAGGGTGCAGAC	gagccaggaggatggtatga	60
Arm1_deletion	TGATAATTTCCTTTCTCCTTTTTCC	CACACTTTAGTAATACACAGGTCTTTTC	60
Snapshot primers			
Primer name	**Primer sequence 5´- 3´**		
C1	TATTAAGCCTCAGGCCTGCC		
C2	AAAAAAAAAAGCGAGCCGAAGCAGGGCGAG		
X/Y1	CCTTCCTTCCCACCCAGGGC		
X/Y2	TCACAGCTCAGGGGCGTGAT		
X/Y3	AAAAA GGGATCGCGAGAGGGGTATA		
X/Y4	AAAAA GCCAGGCAGCCTGGAGTTAG		
X/Y5	AAAAA AAAAA TGCGAGACGTTGAGCTGCGG		
X/Y6	AAAAA AAAAA CTCTCAGCTGAGCCCCAGTG		
X/Y7	AAAAA AAAAA AAAAA CCTCGTCTTCCCCTCGCCTC		
X/Y8	AAAAA AAAAA AAAAA CACAAGAAGCCTCTCCTGTC		
Sequencing Primers			
Primer name	**Forward primer 5´- 3´**	**Reverse primer 5´- 3´**	**Tm**
VCX	CTCACACAAATATCCTGTTGGC	GAATTGTACTTTCTGTCTCCTG	60
VCX_seq1	CTCACACAAATATCCTGTTGGC	TTCTGTTCCTCCACACCTGC	60
VCX_seq2	TATCCCAGTTAGCATGGAG	AGATCACAGAAGGGCTCCG	60
VCX_Seq3	AGGACACGTCCCTGTTCCC	CACACCACCTCTTCCTCCC	60
VCY	TCCCCACACACCTCTTACC	TTCCCACCCAGGGCTACC	60
VCYA	GTAGATTACATATATGCACAATAG	TTCCCACCCAGGGCTACC	60
VCYB	CTAACAGATATATATCATCTATATC	TTCCCACCCAGGGCTACC	60

7.3, Results

7.3.1, Human-chimpanzee sequence divergence

To gain some understanding of how P8 has evolved since human-chimpanzee speciation, divergence calculations were carried out using the 2009 version of the chimpanzee reference sequence. Divergence of 0.01% was observed between human palindrome arms while divergence between the chimpanzee palindrome arms was slightly higher at 0.03% this suggests that gene conversion has been occurring in both species (Figure 7.1). Interspecies divergence between palindrome arms was calculated as 2.31% while divergence between spacers was higher at 2.54%: interestingly, and in contrast to palindromes 6 and 7, this does not suggest that gene conversion is conservative of the ancestral state (P=0.45, 2-tailed Fisher exact test). Interspecies divergence between the 780bp of sequence containing the *VCY* genes was calculated as 5.17% which is significantly higher than observed between the palindrome and the spacer region (P=0.0001, 2-tailed Fisher exact test). This does not suggest that gene conversion between the *VCY* genes is conservative of the ancestral state, as conservative gene conversion would lower the interspecies divergence between the genes compared to regions which cannot undergo gene conversion. Trombetta et al. (2009) also observed higher divergence between the *VCY* genes in comparison to the remainder of the palindrome sequence and this could well be due to *VCY* acting as a sequence acceptor from *VCX* during gene conversion. As the *VCY* genes appear to be under a different influence to the remainder of P8, interspecies divergence was recalculated between regions of P8 excluding the *VCY* genes. This revealed the divergence between the orthologous palindrome arms to be 1.3% which is significantly lower than the 2.54% divergence observed between the spacers (P=0.0001, 2-tailed Fisher exact test)

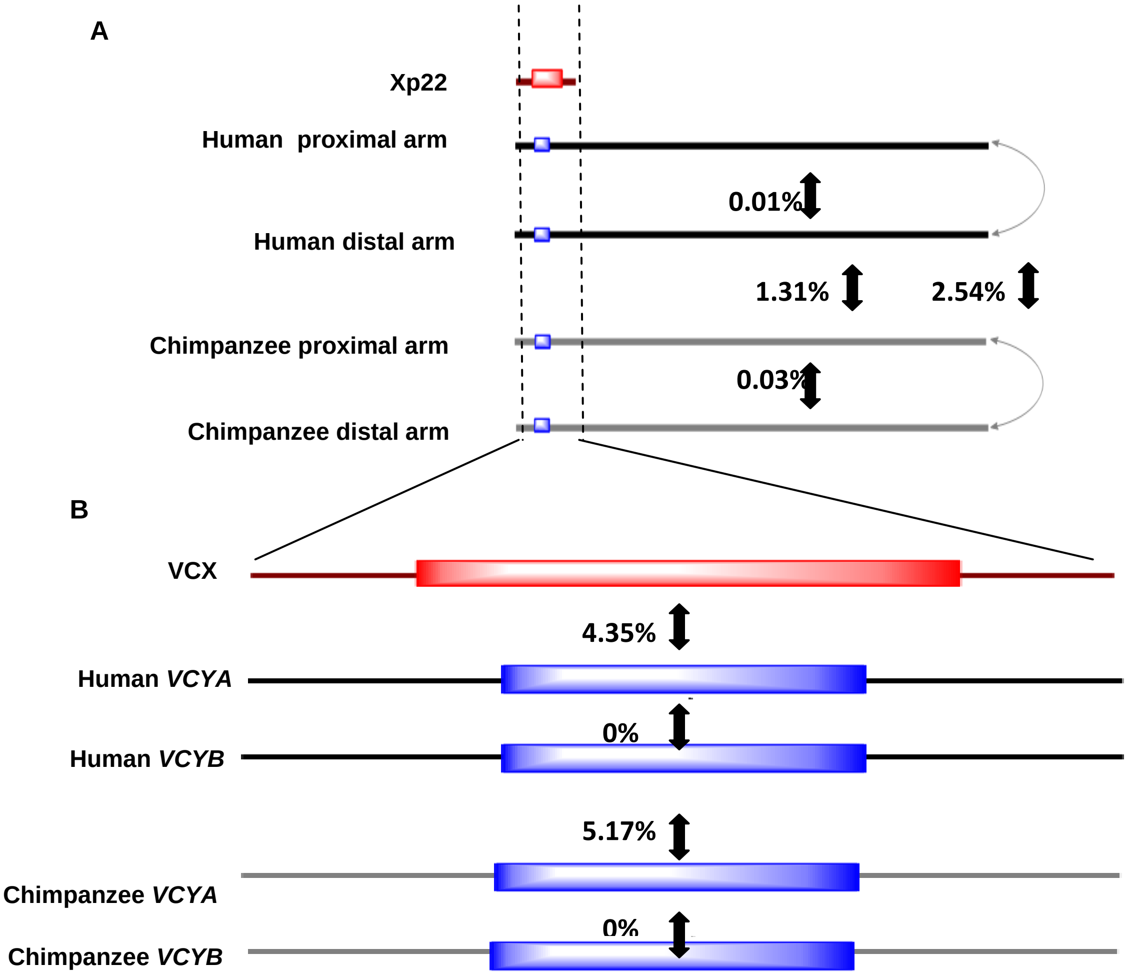


Figure 7.1: Human and Chimpanzee sequence divergence for P8

a) Divergence of 0.01% was observed between human palindrome arms while 0.03% divergence was observed between chimpanzee palindrome arms. Interspecies divergence shows significantly lower (P=0.0001, 2-tailed Fisher exact test) divergence of 1.31% between palindrome arms in comparison to 2.54% divergence observed between the spacers.

b) Zero divergence is observed between the *VCY* gene copies in both human and chimpanzee, while 5.17% divergence was observed between orthologous genes. The significantly higher interspecies divergence (P=0.001, 2-tailed Fisher exact test) in comparison to 1.31% divergence observed between orthlogous palindrome arms does not suggest that gene conversion between the VCY genes is conservative of the ancestral state, suggesting that gene conversion between the arms of P8, but not the *VCY* genes in particular is conservative of the ancestral state.

7.3.2, Re-sequencing of palindrome boundaries

Human and chimpanzee sequence comparisons carried out in Chapter 4 revealed disruption of the outer palindrome boundary due to a 340-bp insertion in the proximal arm of the human palindrome. Sequence similarity within this region was disrupted by the accumulation of multiple mutations on the proximal arm of the palindrome. Divergence between these regions was >5% compared to 0.01% observed across the remainder of the palindrome. As previously discussed in Chapter 3, various rearrangements and deletions are known to occur on the human Y chromosome which do not affect male fertility (Reviewed by Jobling, 2008), therefore it is possible that the insertion may be specific to the Y chromosome reference sequence. Rozen et al. (2003) also observed that outer palindrome boundaries are less well conserved between species than the inner boundaries, which raises the question as to whether the boundaries of P8 are variable across the Y phylogeny in humans. To address this question both the inner and outer boundaries were sequenced in eight individuals representing the major branches of the Y phylogeny and were also compared to the corresponding regions in the CV, JW, and Yh database sequences. Sequencing revealed both inner and outer boundaries to be fixed in all chromosomes, and no evidence of conversion was observed. This suggests that the insertion occurred following speciation but before the MRCA of the Y phylogeny.

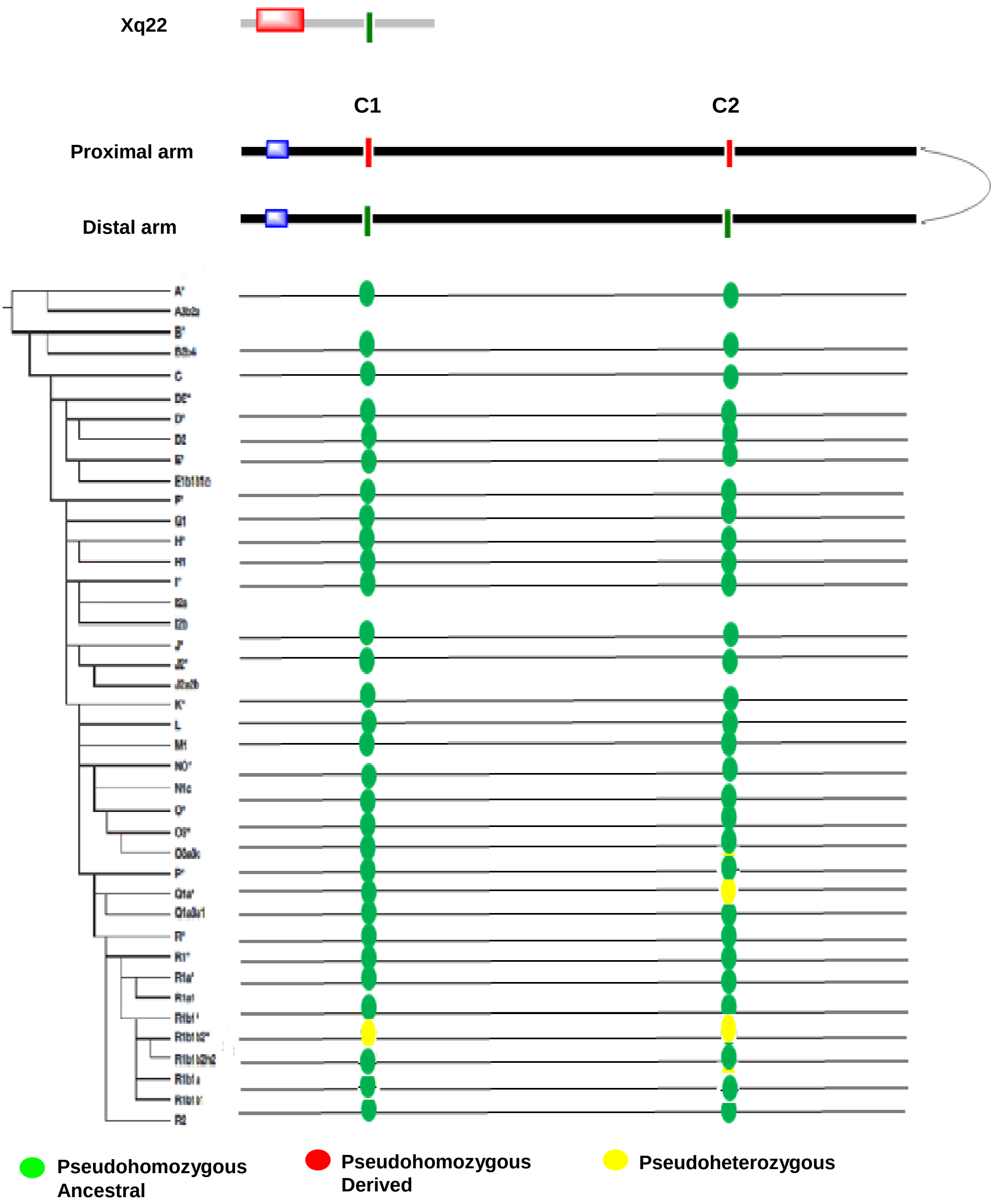
7.3.3, Seeking evidence of gene conversion through typing snPSVs

From analysis of all available database sequences only two PSVs (C1-C2) were identified between the arms of P8. The C1 G/C PSV was only observed in the CV, JW and reference sequence while the C2 G/- PSV was present in all four sequences (Figure 7.2a). In an attempt to identify historical gene conversion events, phylogenetic analysis of both PSVs was carried out. Each PSV was typed in a panel of 64 male individuals representing 30 different haplogroups of the Y phylogeny using the SNaPshot minsequencing protocol (Figure 7.2b).

7.3.3.1, Phylogenetic analysis of the C1 PSV

The C1 C/G PSV is located approximately 7kb from the outer boundary of the palindrome, in the region which exhibits >95% sequence similarity with Xp22. Comparisons with the chimpanzee reference sequence reveals the derived G-allele to have arisen on the proximal arm of the palindrome. Comparisons with the corresponding regions of the X-chromosome reveals the gamtologous region of the X-chromosome to carry the ancestral C-allele suggesting that this PSV is not likely to have arisen from an X-to-Y conversion event and has probably arisen from a mutation on the proximal arm of the palindrome.

Phylogenetic analysis of C1 shows all chromosomes analysed to be pseudohomozygous for the ancestral C-allele with the PSV only being observed in the haplogroup R1b1b$_2$ database sequences. As this variant is also observed in the CV and JW sequences this suggests the PSV has arisen in the R1b1b2 clade of the Y phylogeny and has not undergone subsequent gene conversion in the sequences that are available.



(Figure not drawn to scale)

Figure 7.2: Phylogenetic analysis of snPSVs identified between the arms of P8. From phylogenetic analysis the C1 PSV was only observed in haplogroup R1b1b2 chromosomes and no conversion events were identified. The C2 PSV was observed in the O1a and R1b1b2 reference sequences which are both located within superhaplogroup K*. This suggests that the founder of superhaplogroup K carries the PSV and rapid conversion to the ancestral allele has occurred in all descendant chromosomes other than O1a and R1b1b2. However, this could also represent a sequencing error in the Yh sequence.

7.3.3.2, Phylogenetic analysis of the C2 PSV

The C2 G/- PSV is located approximately 26kb from the outer boundary of the palindrome and lies outside the region of X-Y gametology. Human and chimpanzee sequence comparisons show the single G-deletion to have arisen on the proximal arm of the palindrome. Phylogenetic analysis of C2 did not identify the deletion in any additional chromosomes with all chromosomes being pseudohomozygous for the ancestral allele. This initially suggested that the G-deletion represents a private mutation in the reference sequence chromosome; however, subsequent analysis of the additional reference sequences which were made available after analysis of this section had been completed revealed all three sequences to carry the G-deletion. This is interesting, as the Yh sequence represents a haplogroup O1a chromosome and as haplogroup O1a and R1b1b2 both lie within the K super-clade of the Y phylogeny. There are several possible explainations for this observation. Firstly it could suggest that the founder chromosome of the K super-clade carries the G-deletion with rapid conversion back to the ancestral allele occurring in all descendant clades other than haplogroups O1a and R1b1b2b. It is also possible that two independent deletions have occurred at the O1a and R1b1b2 branches of the phylogeny; however, as the PSV is not located in a poly G tract hypermutation seems unlikely. As the Yh sequence has been assembled based on the reference sequence alignment it is possible that a sequencing error has occurred in the YH sequence.

With so few PSVs being available for analysis, identification of conversion events between the arms of P8 is difficult. The low divergence and lack of variation between palindrome arms, despite the palindrome being known to predate speciation suggests that rapid gene conversion has been occurring in recent human evolution while analysis of the C2 PSV also suggests that conservative gene conversion occurs rapidly between the arms of P8; however, with so few PSVs available for analysis it is not possible to provide statistically significant evidence of this.

7.3.4, Seeking evidence of gene conversion between the *VCY* genes

This section aims to seek phylogenetic evidence of gene conversion occurring between the *VCY* genes, and will also seek evidence of *VCX* acting as a sequence donor to *VCY* during gene conversion as previously suggested by Trombetta et al. (2009).

From analysis of all available database sequences, no PSVs were identified between the *VCY* gene copies. Given that the *VCY* genes are known to predate humn-chimpanzee speciation, this high degree of sequence similarity and lack of variants between gene copies strongly suggests that gene conversion occurs rapidly between *VCY* genes.

In order to identify PSVs which may have the potential to show gene conversion events, sequencing was carried out in 26 male individuals representing the major haplogroups of the Y phylogeny. Both *VCY* gene copies were simultaneously amplified using Y-chromosome-specific PCR, and arm-specific 'phase' was determined through arm-specific sequencing when necessary. Arm phase was determined by carrying out a primary PCR using primer pairs which span the outer boundary of each palindrome arm and sequencing of the *VCY* gene was carried out using the primary PCR as a template. The ancestral state of human PSVs was determined through comparison with the chimpanzee reference sequence.

Sequencing of the 800bp region encompassing the *VCY* genes in 26 individuals identified two PSVs, in two unrelated individuals each carrying a haplogroup J2*(xJ2a2b) Y chromosome. The first PSV (V1 C/G) was located at nucleotide 46 of the *VCY* gene sequence. Comparisons with the chimpanzee reference sequence and arm-specific sequencing revealed the variant G-allele arose within the *VCYA* gene which is located within the proximal arm of the palindrome. The second PSV (V2 G/C) was identified at nucleotide 497 of the *VCY* gene sequence with the variant C-allele arising within the *VCYB* gene located within the distal arm (Figure 7.3). In order to identify historical conversion events both PSVs were typed in a panel of 164 males representing 30 haplogroups of the Y phylogeny. From phylogenetic analysis both PSVs were only observed to be in the pseudoheterozygous state in the haplogroup J2(xJ2a2b) chromosome from which the PSVs were originally identified while all 34 chromosomes descending from haplogroup J2(xJ2a2b) were pseudohomozygous for the ancestral allele. While it is possible that the two PSVs identified between the *VCY* genes copies are rapidly converted back to the ancestral state it is more probable that the PSVs represent a private mutation in these chromosome and therefore it is not possible to determine that gene conversion is occurring.

7.3.4.1, Seeking evidence of conversion between the *VCX* and *VCY* genes

Phylogenetic analysis of snPSVs identified between the *VCY* genes has not provided direct evidence of gene conversion occurring between the *VCY* genes in humans. However, the low sequence divergence of 0.01% since the MRCA of the Y phylogeny

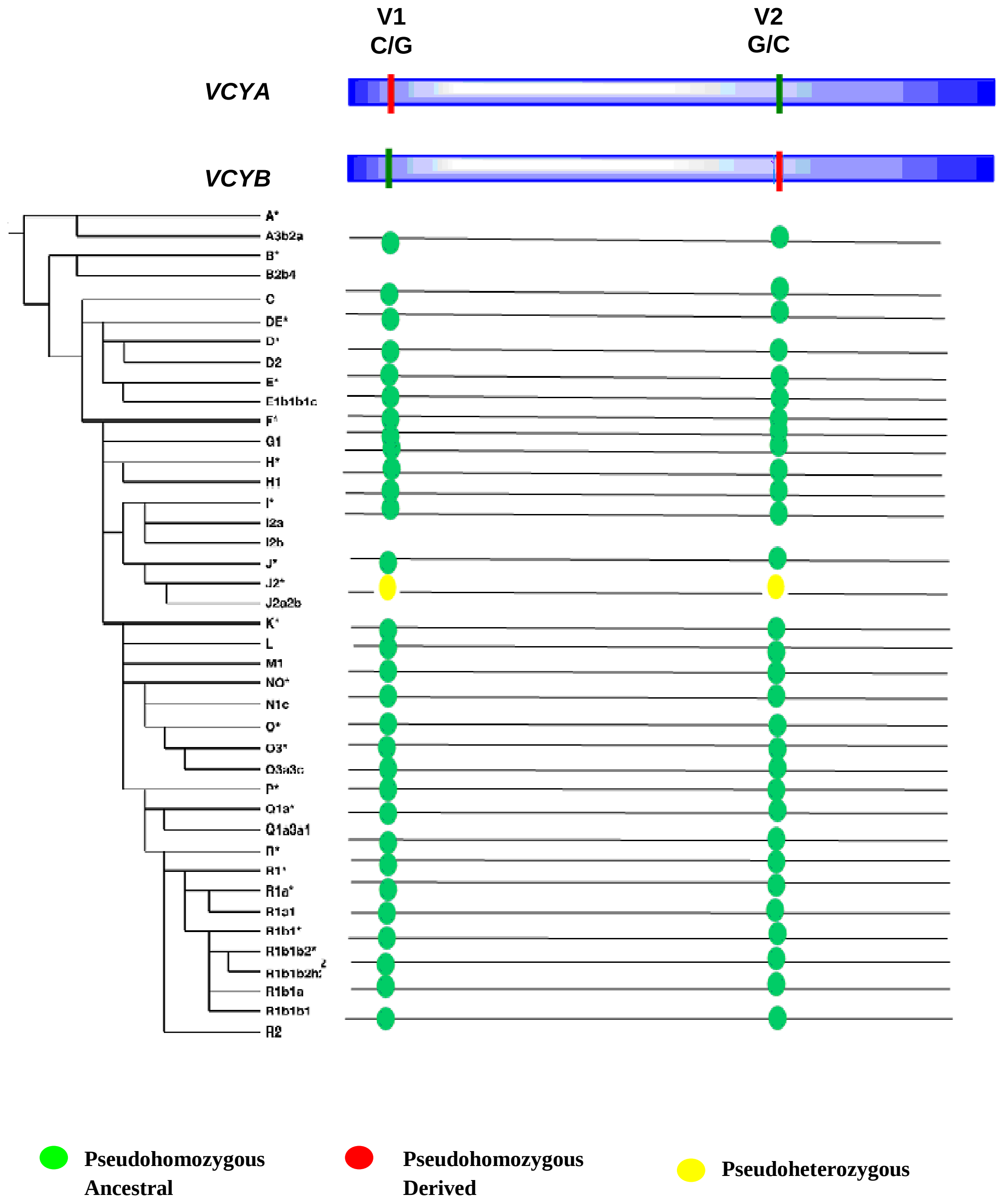


Figure not drawn to scale

Figure 7.3: Phylogenetic analysis of snPSVs identified between human *VCY* genes
Phylogenetic analysis revealed both variants to be limited to the haplogroup J2*(xJ2a2b) chromosome from which the PSVs were originally identified. This suggests that the PSV either represents a private mutation in each individual or that rapid gene conversion back to the ancestral allele has occurred in descendant chromosomes of haplgroup J2*(xJ2a2b).

suggests that gene conversion is occurring between genes. Also significantly higher interspecies divergence between orthologous genes which is also significantly higher than that observed between the spacer regions does not suggest that gene conversion is conservative of the ancestral state. Trombetta et al. (2009) have previously shown that gene conversion occurs between the *VCX* and *VCY* genes and given the location of the genes on different chromosomes the rate of conversion between genes would be expected to be much lower than between *VCY* gene copies and therefore *VCX*/*VCY* conversion events may be more easy to identify.

To seek evidence of gene conversion occurring between the *VCX* and *VCY* genes the *VCX* genes were simultaneously sequenced in the same 26 individuals in which *VCY* has previously been sequenced. Firstly, the V1 & V2 PSVs which were identified from sequencing of the *VCY* genes were compared to the corresponding regions of the *VCX* genes to determine if they have arisen as the result of a *VCX*-to-*VCY* gene conversion event. Secondly, sequence comparisons of the full *VCX* and *VCY* sequences were carried out in order to identify additional gene conversion events.

7.3.4. 2, Sequence analysis of VCY PSVs

Comparison of the V1 C/G PSV where the derived G-allele arises within the *VCYA* gene, with the corresponding region of the *VCX* genes revealed that the *VCX* gene copies in all chromosomes carry the ancestral C-allele. This suggests that the derived G-allele has most likely arisen as the result of a mutation occurring within the *VCYA* gene and not as the result of a conversion event between the *VCX* and *VCY* genes (Figure 7.4).

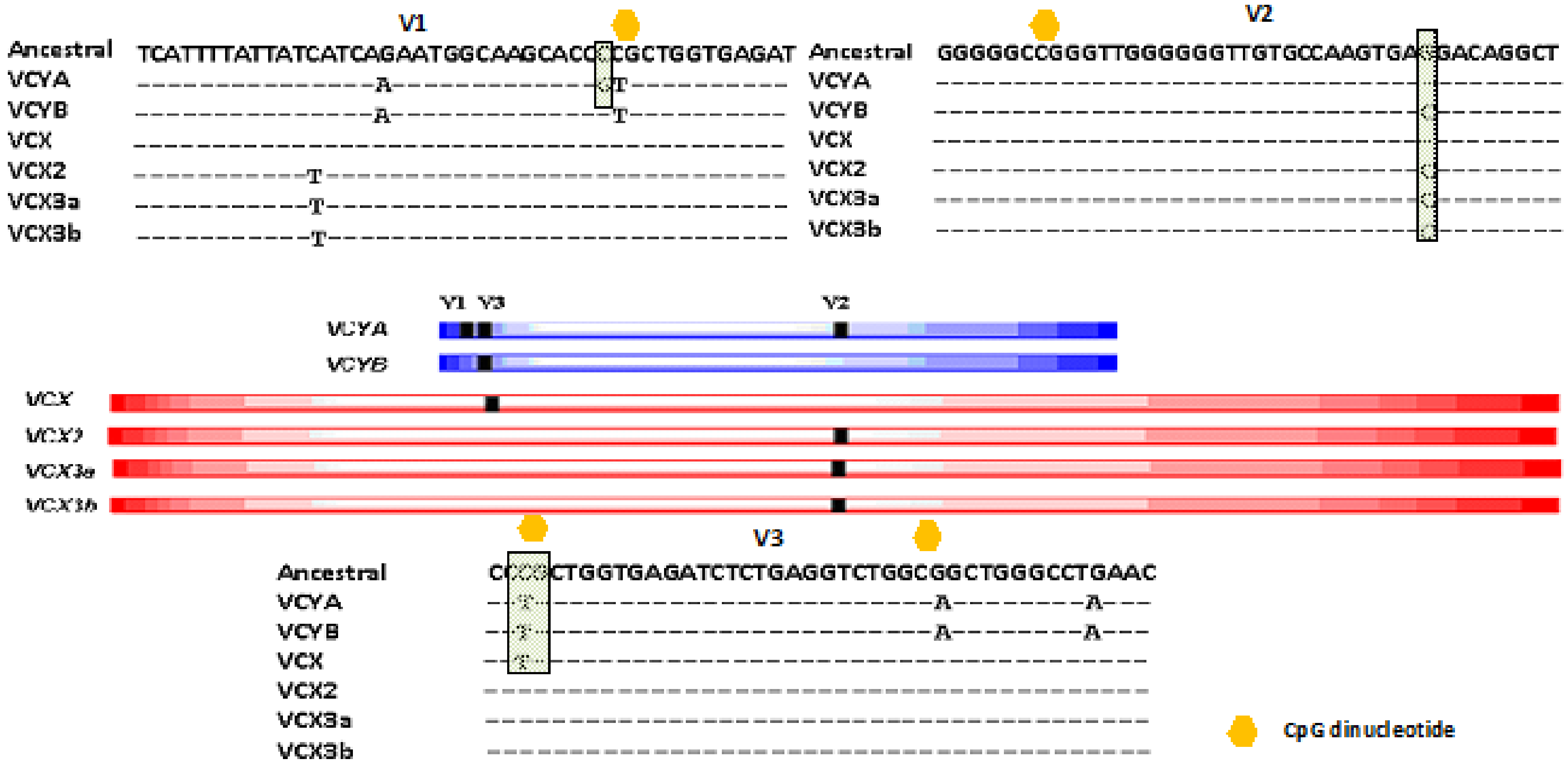

V1
Ancestral
VCYA
VCYB
VCX
VCX2
VCX3a
VCX3b
TCATTTTATTATCATCAGAATGGCAAGCACCCCGCTGGTGAGAT
A
T
T
V2
GGGGGCCGGGTTGGGGGGTTGTGCCAAGTGACGACAGGCT
V1 V3
V2
VCYA
VCYB
VCX
VCX2
VCX3a
VCX3b
V3
CCCGCTGGTGAGATCTCTGAGGTCTGGCGGCTGGGCCTGAAC
CpG dinucleotide

Figure 7.4: Sequencing of the human *VCX* and *VCY* genes
Sequencing of 26 individuals suggests that the V1 PSV has arisen as the result of a mutation on the *VCYA* gene while V2 PSV appears to have arisen as the result of a *VCX*-to-*VCY* conversion event. An additional *VCX-VCY* gene conversion event was also identified (V3) however this was complicated by occurring at the site of a CpG dinucleotide and therefore gene conversion cannot be distinguished from CpG hypermutation

Comparison of the V2 G/C PSV where the derived C-allele arises within the *VCYB* gene, shows the corresponding regions of the *VCX2, VCX3a and VCX3b* genes within this individual to all carry the derived C-allele with only the *VCX* gene carrying the ancestral G-allele. The most plausible explanation for this observation is that one of the *VCX* genes which carries the derived C-allele (*VCX2, VCX3a or VCX3b)* has undergone gene conversion with *VCYB* resulting in a *VCX*-to-VCY gene conversion event which has produced the V2 G/C PSV. Another possibility is that the derived C-allele has arisen within the *VCYB* gene and has been converted in to one of the three *VCX* genes with subsequent gene conversion occurring between the *VCX* gene copies. A third explanation could be that mutation has occurred at the same nucleotide of *VCYB* and three of the *VCX* genes; however, as this variant has not arisen at the site of a highly mutable CpG dinucleotide the chances of the same mutation occurring at the same nucleotide within four different genes is vanishingly small.

From this analysis there is evidence that a gene conversion event has occurred between the *VCX* and *VCYB* genes. Although it is not possible to definitively determine the direction of gene conversion, it seems most plausible that a *VCX* gene has converted the *VCYB* gene. Once converted into the *VCY* sequence the PSV has either remained as a private mutation or been converted back to the ancestral allele by the *VCYA* gene.

7.3.4.3, Full gene sequence comparisons between the VCX and VCY genes

Full analysis of the twenty-six *VCX* and *VCY* gene sequences also identified one possible *VCY* to *VCX* conversion event. However, the interpretation of this possible conversion event is complicated as it occurs at the site of a CpG dinucleotide. This variant represents a transition at a CpG dinucleotide to a TpG which occurs as at rate of ~1.6 x 10^{-7} (Nachmana and Crowella 2000) and while it is possible that gene conversion has occurred, it cannot be distinguished from CpG hypermutation.

In an attempt to identify additional *VCX* to *VCY* conversion events a SNaPshot assay was designed to determine whether *VCX* SNPs are converted into the *VCY* sequence. A total of 29 *VCX* SNPs were identified in the region which exhibits gametology with *VCY* using the UCSC genome browser. Of these, eight which were observed to be polymorphic within the 26 chromosomes sequenced were successfully typed in 200 male chromosomes representing the major haplogroups of the Y phylogeny. The region of *VCY* corresponding to a *VCX* SNP was amplified using chromosome-specific PCR and both *VCY* genes were typed simultaneously. For each X-chromosome SNP all 200 chromosomes were pseudohomozygous for the ancestral allele and no evidence of a derived allele from a *VCX* SNP being converted into the *VCY* sequence was observed.

From this analysis only one possible conversion event between the and *VCY* genes has been identified. Although it has not been possible to definitively determine the direction of gene conversion it seems most plausible that a *VCX* gene which carries the derived allele has converted the *VCYB* gene. This supports findings of Trombetta et al. (2009) who have recently shown the *VCY* genes to act as a sequence acceptor from *VCX* during gene conversion and this may explain the higher interspecies divergence observed between the orthologous *VCY* genes. Despite gene conversion being expected to occur rapidly between *VCY* gene copies, no evidence of a derived allele being converted into the *VCY* gene sequences of descendant chromosomes has been observed.

7.3.4 4, Analysis of the Myers motif

In 2008 Myers et al. reported evidence of NAHR occurring between a 13bp CTCCCTCCCCAC sequence motif located within hotspots adjacent to the *VCX* genes. As gene conversion is known to be mediated by NAHR it is possible that these motifs may also mediate allelic-gene conversion between these regions of the X-chromosome. As the region of the Y chromosome containing the *VCY* genes shares >95% sequence similarity with Xp22 in which the *VCX* genes are situated it is possible that these sequence motifs are also present within the P8 sequence. If this is the case then it is also possible that NAHR occurs between the X and Y chromosomes. Analysis of the P8 sequence revels the CTCCCTCCCCAC sequence motif to be present 12 times within the 1.3kb region between the *VCY* genes and the outer boundaries of the palindrome. The GC content of this region was determined to be 57% which is higher than than the 44% GC content observed in the 1.3kb of sequence downstream of the *VCY* genes which does not contain the sequence motif. While this higher GC content could explain the high density of the sequence motif in this region, as the motif is rich in C alleses its presence could also explain the high GC content of this region. As these motifs are known to mediate NAHR the observation of the Myers motif in the P8 sequence raises the possibility that NAHR and therefore gene conversion may also occur between these regions of the X and Y chromosomes and also between the arms of P8. Analysis of snPSVs identified from the reference sequence alignment shows the two PSVs to be located 7kb and 25kb away from the Myers motif and as conversion tracts are believed to be less than 1kb in length (Chen et al. 2007) these PSVs are not likely to be contained in a conversion tract which is mediated by NAHR between the CTCCCTCCCCAC sequence motif.

Analysis of the P8 reference sequence suggests that the *VCYA* gene is in the same orientation as *VCX* while the *VCYB* gene is in the same orientation as *VCX2*, *VCX3a* and *VCX3b*. As one conversion event has been identified between directly orientated *VCX* and *VCYB* genes it seems possible that NAHR mediated by the Myers motif may have resulted in gene conversion between the *VCX* and *VCY* genes. As the Myers paper was published after work on this chapter had been completed, sequencing of the 800-bp regions containing the Myers motif was not carried out in this study. However, analysis of all available reference sequences for this region did not provide evidence of gene conversion in this region.

7.4, Discussion

Significantly lower interspecies divergence between the arms of P8, (excluding the regions which contain the *VCY* genes), suggests that gene conversion between the arms of P8 is conservative of the ancestral sate. From analysis of the P8 reference sequence only two PSVs were identified between the palindrome arms since the MRCA of the Y phylogeny, this in itself is highly suggestive of gene conversion; however, phylogenetic analysis of these PSVs did not provide significant evidence of gene conversion. This highlights the problem of ascertainment bias associated with typing snPSVs identified from a single reference sequence as additional PSVs may have arisen in ancestral chromosomes but undergone gene conversion within the reference sequence chromosome and therefore these sites will remain unidentified. From data obtained in this study it would appear that gene conversion occurs rapidly between the arms of palindrome 8, which also poses a problem when trying to identify gene conversion events and determining the rate of gene conversion. As gene conversion acts to homogenise two sequences conversion events are effectively invisible, and this poses even more of a problem if gene conversion is conservative of the ancestral state. Although it appears that gene conversion is occurring rapidly it has not been possible to identify conversion events in recent human evolution.

Interestingly, interspecies sequence divergence between the *VCY* genes showed significantly higher divergence between orthologous *VCY* genes, with divergence being approximately twice that observed between the spacers. While 100% sequence identity was observed between *VCY* genes, which is suggestive of rapid gene conversion, interspecies divergence of 5.17% does not suggest that gene conversion is conservative of the ancestral state, as appears to be the case for the remainder of the palindrome. In the study by Trombetta et al. (2009) it was suggested that the *VCY* genes act as a sequence acceptor from *VCX* during gene conversions and this may well be the cause of the higher interspecies divergence observed between the *VCY* genes. From this study only one example of gene conversion between the *VCX* and *VCY* genes was observed and while this seems most likely to represent a *VCX*-to-*VCY* conversion event it has not be possible to definitively determine the direction of gene conversion. These data appear to support the findings of Trombetta et al. (2009) which shows *VCY* acting as sequence acceptor from *VCX*; however, as very few gene conversion events have been identified, further sequencing of these regions would need to be carried out.

While relatively few gene conversion events have been identified between P8 and its X-chromosome gametologs, there is growing evidence to suggest that NAHR occurs. Identification of Myers motifs and chromosomal rearrangements mediated by NAHR suggest that gene conversion has the potential to occur, and sequencing of these regions and known breakpoints may identify additional conversion events. While this chapter provides evidence which is highly suggestive of gene conversion more detailed analysis of these regions may have the potential to identify additional conversion events.

Chapter 8: Seeking evidence of gene conversion between paralogs of IR1

8.1, Introduction

Identifying gene conversion events between the paralogs of inverted repeats (IRs) is more complicated than between the arms of palindromes. Not only are IRs located in regions which display paralogy with other regions of the Y chromosome, some IRs such as IR1 and IR2 are located in regions which are prone to deletion and duplication. From the bioinformatic exploration carried out in Chapter 4, IR2 and IR3 displayed the highest degree of inter-paralog sequence similarity at 99.93% and 99.95% respectively, while IR1 was lower at 99.65% and IR4 only displayed 96% similarity between paralogs. Despite the high sequence similarity observed between paralogs of IR2 and IR3, due to the complexity of the regions in which they reside it would not be possible to study gene conversion phylogenetically. Also, as it could not be determined whether these IRs predate speciation, low divergence due to gene conversion could not be distinguished from low divergence due to recent duplication. For this reason only IR1 was selected for further analysis. IR1 is also complicated by paralogy with P1 towards the boundaries; however, a 17kb region of unique Yp-Yq sequence identity exists which could be studied phylogenetically for evidence of inter-arm gene conversion.

8.1.1, Y chromosome rearrangements mediated by NAHR

As some IRs are known to be located in regions which are prone to deletion, a potential problem when seeking evidence of gene conversion among IRs is deletion of one paralog of the IR. When typing PSVs identified between the paralogs of IRs, this will result in pseudohemizygosity being misinterpreted as pseudohomozygosity, which could lead to an over-ascertainment of gene conversion events. As well as exhibiting paralogy with P1, the IR1 paralog situated on Yq (IR1Yq) is located within a region which is prone to duplication and deletion. The most common deletions occurring in this region are the *AZF* deletions of which there are three different types termed *AZFa, AZFb* and *AZFc.* Of these the *AZFb* and *AZFc* deletions have been shown to result from NAHR between large repeats contained within the arms of palindromes (Navarro-Costa, Plancha, and Gonçalves 2010). NAHR between P1 and P5 is known to cause the AZFb deletion which removes a 6.2-Mb segment of sequence containing 32 genes (Repping et al. 2003), while the *AZFc* deletion resulting from NAHR between P1 and P3 causing a 3.3-Mb deletion which removes 22 genes (Page 1986). Both the *AZFb* and *AZFc* deletions result in the deletion of genes which are essential for spermatogenesis and result in male infertility (Page 1986; Repping et al. 2003). As previously discussed in Chapter 3, not all Y chromosome deletions result in infertility and many can be passed on unnoticed from father to son. The *b2/b3* deletion is one such deletion which occurs in the same region as *AZFc* and results in the loss of the portion of Yq which contains the IR1Yq paralog. The *b2/b3* deletion removes only half of the genes deleted in an *AZFc* deletion and as a result its effect is milder and it does not result in male infertility (Repping et al. 2003). The founder of the haplogroup N branch of the Y phylogeny is known to carry a *b2/b3* deletion resulting in all haplogroup N subclades also carrying the deletion (Kamp et al. 2000). Independent *b2/b3* deletions also have the potential to arise in individuals from any haplogroup of the Y phylogeny and this must be taken into account in this study.

Other Y chromosome rearrangements are also known to occur which are mediated by NAHR between the paralogs of IRs; for example, NAHR between IR3 paralogs causes a paracentric inversion on Yp (Repping et al. 2006) and it has been hypothesised that NAHR between the paralogs of IR1 or IR4 may sponsor a pericentric inversion which causes inv(Y)(p11.2q11.23) (Hurles and Jobling 2003). Similarly to deletions, inversions of the Y chromosome do not necessarily have detrimental consequences on male fertility and pericentric inversions of the Y chromosome such as inv(Y)(p11.2q11.23) which occur frequently in the Gujarati Muslim Indian population have been observed cytogenetically and the inverted Y is apparently not associated with any reproductive disadvantages (Bernstein et al. 1986).

As some Y chromosome rearrangements are known to be caused by NAHR between palindromes and IRs it is also possible that gene conversion occurs in these regions. This Chapter will look for evidence of gene conversion between the paralogs of IR1; such evidence would be interesting as it would suggest that the Y chromosome can fold on itself to allow intrachromosomal recombination during meiosis.

8.1.2, Problems identifying gene conversion events between IR1 paralogs

Paralogy with P1 towards the boundaries of IR1 makes only 17kb of unique Yp-Yq sequence available for phylogenetic analysis. Phylogenetic analysis of the remaining sequence would not be possible as co-amplification during PCR would make it difficult to determine whether gene conversion is occurring and between which regions a conversion event has occurred. As previously discussed the study of gene conversion between IR1 paralogs could potentially be complicated by several deletions which occur on Yq which may result in the false identification of gene conversion events.

8.1.3, Chapter aims

The aims of this chapter are to establish whether gene conversion occurs between paralogs of IR1 using a phylogenetic analysis of snPSVs and microsatellites located within a 17-kb region of unique IR1 sequence identity. Sequence analysis of the regions of IR1, which cannot be studied phylogenetically due to paralogy with P1, will also be carried out in order to determine whether gene conversion occurs within these regions of IR1.

8.2, Materials and methods

Analysis was carried out as described in chapter 2, with the following exceptions.

8.2.1, DNA samples

8.2.1.1, Phylogenetic analysis

58 male DNA samples representing 28 Y chromosome haplogroups from 17 populations were selected from the CEPH–HGDP diversity panel (Cann et al. 2002). Chromosomes from haplogroup N which are known to carry the *b2/b3* deletion were excluded from analysis. All samples were subject to whole-genome amplification (WGA) by the multiple-displacement amplification method (Dean et al. 2002) using the RepliG midi Kit (Qiagen).

8.2.1.2, Microsatellite typing

Forty-five genomic male DNAs from haplogroup R1a1 representing four populations were selected from the CEPH–HGDP diversity panel (Cann et al. 2002)

Table 8.1, Primers used in the analysis off IR1

Sequencing Primers		
Primer name	**Primer sequence 5´- 3´**	**Tm**
240Bp.F	CTGAGACACTATGAGACAAAG	60
240Bp.R	AAAGTAGTCATAACAAAACAGAG	60
360Bp1.F	TGAATCATTAAGGGGACCATG	60
360Bp1.R	GCCTGAGCGAGGTCACAG	60
360Bp2.F	ATCTGTGACCTCGCTCAGG	60
360Bp2.R	TTGTCCTAGCTTGAGTTGCC	60
540Bp1.F	GAAGAGCAGGAAAAACCTATG	60
540Bp1.R	CAGTGAACCTGGGAGAAGC	60
540Bp2.F	TTCTCCCAGGTTCACTGTAC	60
540Bp2.R	CTCTCCTTATTGACTCTCAAG	60
180Bp.F	AAGGCTCACTAGCCACCAG	60
180Bp.R	TTGTCTCAACCAATCAGGCC	60
60Bp.F	TTCACTAAAGGAGAGCATACC	60
60Bp.R	GGTATTTACAGACAATGGTTAC	60
600Bp1.F	CAGTACTGGTACATCTCAGC	60
600Bp1.R	ACTTGCCAGAAATCCATCTTG	60
600Bp2.F	ATTTCCTGCAATAATGAGAGTG	60
600Bp2.R	TTCTCAGGTTAACGGTCCTC	60
PCR Primers 5´- 3´		
Ir1_C19_23.F	CGTTCTCTGAGGTGGAGTG	60
Ir1_C19_23.R	CCTGGCAGGGTGGCTCAC	60
IR1_C24_C26.F	TACCACATTCTATGGACTCAC	60
IR1_C24_C26.R	CAAAGAGGGCTTGTGTCAAG	60
C1_4.F	AAACTACAGTATGATGATTGCC	60
C1-4.R	CTGGTATATCAAATGGTGCTG	60
C34-37.F	ACGCATAAGATTCTCACATGC	60
SNaPshot primers 5´- 3´		
IR1_C19	TCTATGTTGGCAAACGATTT	50
IR1_C21	AAAAA AAAAA CAACATCTCTTTGCTTTCAC	50
IR1_C22	AAAAA AAAAA AAAAA GGTGGGCGGATCAAGAGTTC	50
IR1_C24	CACCTGTGAGGAAATAAAAA	50
IR1_C25	AAAAA AAAAA CATGTTTACCTTCCATTACA	50
IR1_C26	AAAAA AAAAA AAAAA AAATGTGTTCTGAACAGGAC	50
C1	AAGAAGCTTCAGAAAAGTTT	50
C2	AAAAA AAAAA TAACACTCTGTTGAATTTCC	50
C3	AAAAA AAAAA AAAAA AAAAA TTATTTTTCAACCTTTGTTT	50
C4	AAAAA AAAAA AAAAA AAAAA AAAAA AAAAA TGCCTTTCTGTGTGCCAGTC	50
C35	AAAAA AAAAA AATGCCAATCTAATAATGAT	50
Microsatellite primers 5´- 3´		
IR1_M2.F	FAMGATGTTGGATGTTCTGGCTG	58
IR1_M2.R	CAAGTAATTTGTGTGAGCAGG	58

8.3, Results

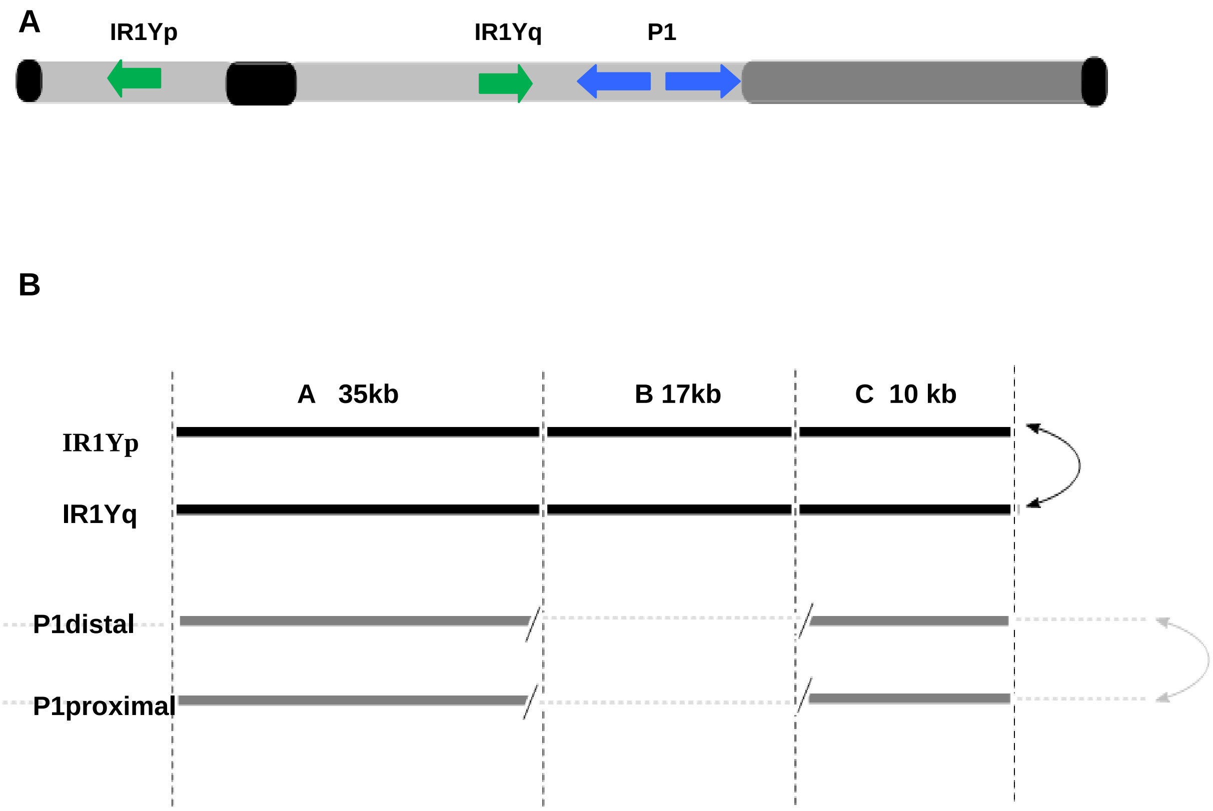
Despite the large physical distance separating the paralogs of IR1, 99.65% sequence similarity is observed between the 62-kb of sequence. As well as high sequence similarity between paralogs >99.5% sequence similarity is also observed between the regions towards the boundaries of IR1 and P1 (Figure 8.1A) leaving only a 17-kb region of unique Yp-Yq sequence similarity available for phylogenetic analysis. In this Chapter, IR1 will be sub-divided into three sequence blocks, termed A, B and C (Figure 8.1B).

Block A spans 35kb and is located towards the outer boundary of IR1. This region exhibits >99.97% sequence similarity with two regions of Yq which correspond to regions on the proximal and distal arms of P1.

Block B refers to the 17-kb region of unique Yp and Yq identity which will be studied phylogenetically for evidence of gene conversion.

Block C refers to a 10-kb region located towards the inner boundary of IR1. This region exhibits >90% sequence similarity with four Y chromosome paralogs, three of which are located on Yq and one located on Yp. Three of the paralogs span less than 1.5kb and display less than 95% sequence similarity with IR1. Due to the short sequence length and high divergence these paralogs were not included in further analysis. The remaining two paralogs are located on Yq and correspond to the proximal and distal arms of P1 and exhibit >99% sequence similarity with IR1.

In the first section of this Chapter, phylogenetic analysis of snPSVs and microsatellites



Figures not drawn to scale

Figure 8.1: Structure and location of P1 and IR1

a) IR1 consists of two paralogs, one located on the long arm of the Y chromosome (IRYq) and the second situated on the short arm (IR1Yp). IR1 also exhibits paralogy with palindrome 1 (P1) which is located on Yq approximately 5kb downstream of IR1Yq paralog.

b) IR1 can be subdivided into three regions termed A, B and C. Regions A and C which span 35 kb and 10 kb respectively both exhibit paralogy with P1 while region B which spans 17 kb exhibits only Yp-Yq paralogy.

located within block B will be carried out in order to determine whether gene conversion occurs between IR1 paralogs. The second section will carry out sequence analysis of blocks A and C which exhibit paralogy with P1 and cannot be studied phylogeneticaly for evidence of gene conversion.

8.3.1, Phylogenetic analysis of snPSVs located within block B

Based on the alignment of the reference sequence, seventy-nine PSVs were identified within block B of IR1. Of these, sixty-eight were snPSVs, ten were indels which range from 2-bp to 2.5kb, and one was a microsatellite showing length variation.

Six sub-regions of block B ranging from 60 - 600bp in length and covering a total of thirty-two snPSVs were chosen for re-sequencing. Regions were selected based on the length of uninterrupted sequence observed between two snPSVs identified from the reference sequence alignment. Each sub-region was sequenced in eight male individuals representing the major branches of the Y phylogeny and resulting data were also aligned with the corresponding regions of the JW, CV, and Yh database sequences.

From initial analysis no additional PSVs were identified from comparison of all available database sequences or through sequencing. Thirty of the thirty-two PSVs identified from the reference sequence alignment were observed at all major branches of the Y phylogeny, suggesting that these PSVs have arisen before the MRCA of the Y phylogeny, and undergone no observable subsequent change. Of the two PSVs which were not observed to be present in all chromosomes, the C40 C/T PSV was absent in the Yh sequence due to a CTT deletion which removes the PSV site on Yq. The CTT deletion was not observed in any chromosome other than the Yh chromosome, which belongs to haplogroup O1a and most likely represents a mutation specific to this sub-lineage or individual or a sequencing artifact. The C12 A/G PSV was only observed to be pseudoheterozygous in haplogroups P* and R1b1b2 while all remaining chromosomes were pseudohomozygous for the ancestral allele.

The large proportion of PSVs which appear to have arisen before the MRCA of the Y phylogeny could either suggest that gene conversion does not occur between IR1 paralogs or that it occurs at a very low rate. If conversion in this region is slow, eight chromosomes may not be sufficient to capture conversion events and a larger sample set may be required. To increase coverage of this region fourteen additional snPSVs from across block B were typed in a panel of 58 individuals representing the major haplogroups of the Y phylogeny but excluding all haplogroup N chromosomes. The C12 PSV which has previously only been observed in haplogroup P* and R1b1b2b chromosomes was also included in analysis. Phylogenetic analysis revealed the fourteen additional PSVs to also be pseudo-heterozygous across the Y phylogeny while the C12 PSV was only observed to be pseudoheterozygous in haplogroups P, Q and R and their subclades. All remaining haplogroups were pseudohomozygous for the ancestral allele as determined through comparison with the chimpanzee sequence.

C12 was further typed in 100 additional chromosomes including 50 chromosomes from haplogroups P, Q and R and their subclades. This typing confirmed the PSV to be confined to the P superclade of the Y phylogeny while all haplogroups outside of this branch are pseudohomozygous for the ancestral allele (figure 8.2). No evidence of conversion to the ancestral or derived alleles was observed within superhaplogroup P, suggesting that once a PSV arises within the phylogeny it persists in descendent chromosomes. Analysis of snPSVs identified within block B does not suggest that gene conversion has been occurring between IR1 paralogs in recent human evolution. Analysis of the C12 A/G PSV which arises in superhaplogroup P of the Y phylogeny indicates that once a PSV is introduced into the IR1 sequence it persists in all descendent chromosomes with no apparent conversion events observed. However, the possibility remains that rare gene conversion events occur which have not been captured within this sample set. It is possible that gene conversion may have been rapid before the paralogous sequences were disrupted some time before the MRCA of the Y phylogeny and if gene conversion does still occur in this region, the rate of conversion must be very low.

8.3.2, Seeking evidence of gene conversion through microsatellite typing

From the reference sequence alignment, only one microsatellite was identified within the 17kb of block B. The M2 CA(n) microsatellite showed only two repeat differences between its paralogous copies with 21 repeats on Yp and 23 repeats on Yq. Ideally a larger repeat difference between loci is needed to make conversion events easier to identify. As previously discussed in Chapter 5, microsatellites with large repeat difference between copies are more powerful for identifying conversion events than loci with three or fewer differences in repeat number, as conversion events can be more readily distinguished from the outcomes of single or two-step microsatellite mutation. Comparison of the available database sequences shows all to carry the CA(21,23)

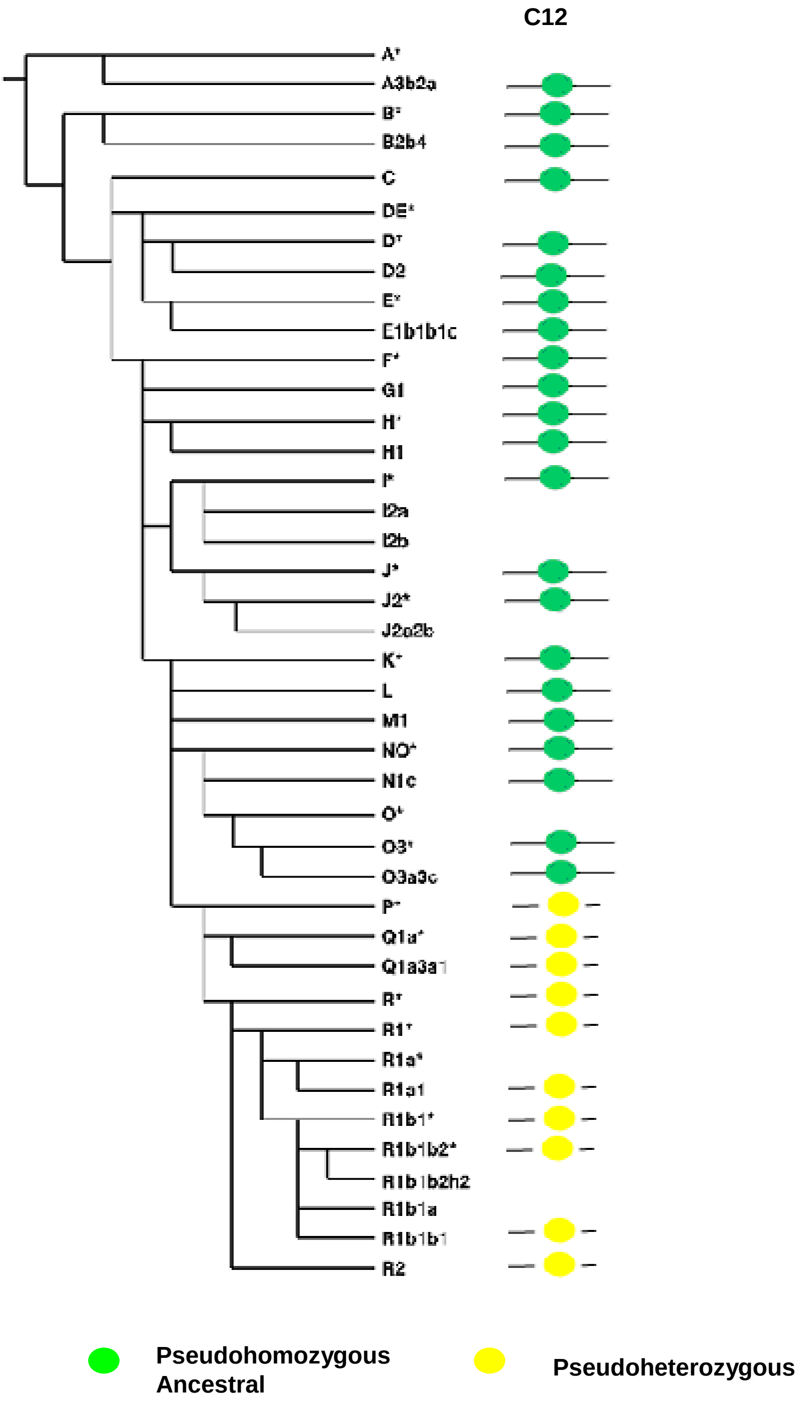


Figure 8.2: Phylogenetic analysis of the C12 PSV

One snPSV was shown to arise within the founder of Hg P with all descendant chromosomes being pseudoheterzygous. No evidence of gene conversion was observed.

haplotype as observed from the reference sequence alignment. To gain an understanding of diversity of M2 and identify haplogroups which display large repeat differences between microsatellite copies, M2 was typed in a panel of fifty-eight males carrying Y chromosomes from across the Y phylogeny (Figure 8.3A).

From the preliminary analysis of M2 four chromosomes from haplogroups A(xA3b2a), B2b4, O3a3c and R1a1 showed >4 differences in repeat number between copies and have the most potential to identify gene conversion events. However, as previously stated, a well represented and defined set of chromosomes are required in order to capture gene conversion events. As haplogroups A(xA3b2a), B2b4 and O3a3c are either not well represented or defined within the CEPH-HGDP panel only R1a1 was subjected to further typing.

8.3.2.1, Analysis of the M2 microsatellite

M2 was further typed in forty-five chromosomes from haplogroup R1a1 representing four different populations (supplementary table S8.1). Extended typing of M2 identified two chromosomes carrying the CA(18:23) haplotype observed from preliminary typing which leaves a 5-repeat difference between copies, while no chromosomes were identified which carry either the pseudohomozygous CA(18:18) or CA(23:23) haplotypes which would be suggestive of gene conversion (Figure 8.3B). One additional chromosome carrying the CA(20:25) haplotype was also identified giving a five-repeat differences between microsatellite copies; again. the pseudohomozygous CA(20:20) and CA(25:25) haplotypes which would be suggestive of gene conversion were not

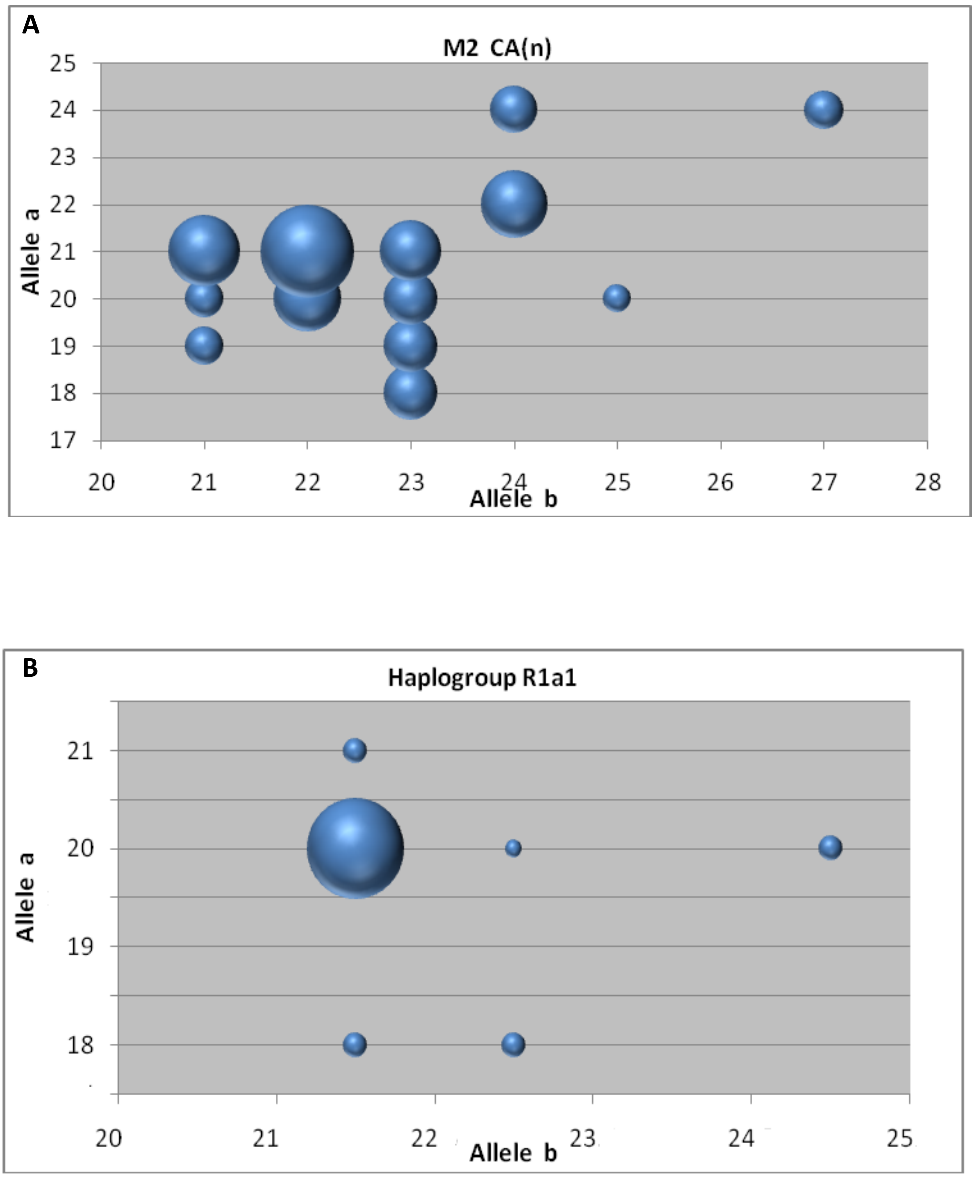


Figure 8.3: Phylogenetic analysis of the M2 microsatellite.

a) Microsatellite data from phylogenetic analysis of displayed as bubble plots where the X-axis represents the major allele and the Y-axis represents the minor allele. The area of the bubble is proportional to the number of chromosomes carrying that genotype. Haplogroup R1a1 was observed to carry the CA(20:25) haplotype and further analysis of R1a1 was carried out.

b) Two chromosomes were identified carrying the CA(18:23) haplotype giving a 5-repeat difference between copies. No chromosomes were observed to carry the CA(18:18) or CA(23:23) haplotypes, Two chromosomes carrying the CA(20:25) haplotype also giving a 5-repeat difference between copies. Again the CA(20:20) and CA(25:25) haplotypes were not observed. Intermediate haplotypes CA(18:22) CA(20:22) CA(18:23) and CA(2:22) haplotype were also observed within the sample set. This distribution of haplotypes shows 1-2-step mutational differences.

observed. In addition, the intermediate CA(20:22) and CA(20:23) haplotypes were identified within the sample set, with over 80% of the chromosomes typed carrying the CA(20:22) haplotype. Given that the intermediate haplotypes are observed in the allele distribution which reduces the difference between microsatellite copies to two repeats, this analysis cannot provide evidence of gene conversion occurring between duplicated microsatellites. From this analysis, M2 appears to be subject only to stepwise mutational processes. However; the possibility remains that historical gene conversion has occurred but has been masked by subsequent mutation.

From analysis of snPSVs and microsatellites within the 17kb region of unique Yp-Yq identity no evidence of gene conversion has been observed. The presence of two large indels and PSVs which appear to have become "fixed" since the MRCA of the Y phylogeny suggests that gene conversion does not occur between IR1 paralogs; however, due to paralogy with P1, over 45kb of sequence remains unanalysed. In order to determine whether gene conversion may be occurring between the regions of IR1 which also exhibit paralogy with P1, analysis of the reference sequence for these regions was carried out.

8.3.3, Analysis of regions which exhibit paralogy with P1

This section will utilise the available database sequences to establish whether gene conversion occurs within blocks A and C of IR1 which also exhibit paralogy with P1. This part of the study relies on the availability of the chimpanzee reference sequence to determine the ancestral state of human PSVs; however, relatively little is known about the structure of IR1 in chimpanzee. Ideally sequence comparisons with an independant primate species such as gorilla or macaque would also be performed to provide evidence of the deep-rooting ancestral state: however, sequence data are only available for female gorilla and macaque, and so no information on the Y chromosome is available. In this section a thorough analysis of the chimpanzee Y chromosome reference sequence will be carried out in order to determine the structure and reliability of the chimpanzee sequence for this region.

8.3.3.1, IR1 origins on the chimpanzee Y chromosome

Regions corresponding to the human IR1 paralogs were located in the March 2006 version of the chimpanzee reference sequence using the BLAT and convert functions in the UCSC genome browser. Both functions identified one region corresponding to the IR1Yp paralog on the short arm of the chimpanzee Y chromosome and a second region on the long arm corresponding to human IR1Yq. In contrast to the human sequence, no evidence of multiple regions of paralogy in the chimpanzee sequence was observed, Regions corresponding to the segments of P1 which exhibit paralogy with IR1 were also located in the chimpanzee reference sequence and interestingly, these sequences mapped to the same region previously identified as IR1. These findings could suggest that prior to speciation IR1 and P1 were the same structure and that a duplication event has occurred in the human lineage. It is also possible that IR1 and P1 both existed in the chimpanzee genome prior to speciation but have been lost from the chimpanzee Y chromosome during evolution. Also, as the reliability of the chimpanzee sequence has been questioned, it is possible that either a mis-assembly has occurred or that the chimpanzee reference sequence for this region is not complete. The identification of separate paralogs on Yp and Yq suggests that these regions correspond to IR1; however human P1 has been shown to contain multiple paralogs located within the palindrome arms, and while the regions of P1 which contain the *DAZ* genes are known to predate speciation (Rozen et al. 2003) it is possible that other paralogs within P1, such as P1.2 and IR5, have arisen from human-specific duplication event.

Analysis of the 2009 version of the chimpanzee reference sequence provides little additional information on the structure and location of IR1 in the chimpanzee genome. The newly published sequence is not available from the UCSC genome browser and a graphical view of the sequence was not available at the time of writing. From analysis of the complete chimpanzee sequence only one region corresponding to each IR1 paralog was identified suggesting that there is no duplication of this structure on the chimpanzee Y chromosome. Whether these sequences belong to IR1 or P1 in chimpanzee is not important, as both versions of the reference sequence are complete and do not appear to display multiple paralogy in the chimpanzee sequence, they provide a reliable indication of ancestral state for both IR1 and P1 in humans. However, if a human-specific duplication has occurred this may influence the significance of divergence calculations between these regions.

8.3.3.2, Human and chimpanzee sequence comparison of blocks A and C

Chimpanzee sequence was obtained for the regions which correspond to the paralogs of human IR1 and separate analysis of blocks A-C were carried out. The most interesting observation made from preliminary sequence comparisons was that in block B, which exhibits unique Yp-Yq sequence identity, equal numbers of PSVs arise due to mutation on IR1Yp and IR1Yq, with 38 (47.5%) snPSVs arising on IR1Yq and 42 snPSVs (52.5%) arising on IR1Yp. However, in blocks A and C, which share paralogy with P1 fewer PSVs arise through mutation on IR1Yq than through mutation on IR1Yp. In Block A only two PSVs (5%) arise due to mutation on IR1Yq in comparison to 35 (95%) due to mutation on IR1Yp While in block C only 1 PSV (1.5%) arises due to mutation on IR1Yq while, 69 (98.5%) arise due to mutation on IR1Yp. The significantly lower (P=0.0001 chi square) number of mutations arising on IR1Yq in blocks A and C compared to block B, which is known not to undergo gene conversion, is interesting as an equal number of mutations would be expected to arise on both Yp and Yq in the absence of gene conversion. This suggests that some mechanism other than mutation may be acting in regions A and C which prevents the accumulation of mutations on Yq and this could be explained by conservative gene conversion occurring between IR1Yq and P1. In order to address this possibility, sequence comparisons between IR1 and P1 in blocks A and C were carried out using the chimpanzee sequence to determine ancestral state.

8.3.4, Interspecies sequence divergence between blocks A and C

Analysis of the chimpanzee reference sequence suggests that either P1 or IR1 may have arisen as the result of a human-specific duplication and this must be taken into account when interpreting divergence calculations as a more recent duplication would create lower divergence between paralogs which could be mistaken for evidence of gene conversion. To gain some understanding of what may be occurring in these regions P1 sequences were obtained via the UCSC genome browser using landmark STS as described by Skaletsky et al. (2003), and divergence was calculated using DnaSP v5 (Rozas and Rozas 1999). P1 sequences were aligned with the IR1 paralogs and divergence between sequences was calculated. Divergence between the P1 paralogs was calculated as 0.05% while divergence between the IRYq paralog and P1 paralogs was lower at 0.04%. Interestingly, divergence between IRYp and the P1 paralogs was significantly higher at 0.49% (P=0.0001, using the 2-tailed Fisher exact test) which is similar to the 0.45% divergence observed between both paralogs of IR1 in regions A and C (Figure 8.4A). While it has previously been established that gene conversion occurs between the arms of P1 (Rozen et al. 2003) the low divergence observed between IR1Yq and P1 suggests that gene conversion may also occur between IR1 and P1. The higher divergence observed between IR1Yp and P1, which is similar to that observed between the IR1 paralogs which do not undergo frequent gene conversion, suggests that gene conversion between IR1Yp and P1 either does not occur or is very rare. Gene conversion occurring between P1 and IR1Yq could explain the lower number of mutations arising on IRYq and low divergence which is similar to divergence observed between P1, which is known to undergo gene conversion (Rozen et al. 2003). The full significance of divergence calculations carried out for this region is difficult to determine as the low divergence observed between paralogs could also result from a duplication event as well as from gene conversion, and duplication of IR1Yq to produce P1 (or vice versa) could explain the lower divergence observed between the three paralogs.

8.3.4.1, Evidence of gene conversion in blocks A and C

Sequence comparisons identified 72 nucleotides in blocks A and C where gene conversion appears to have occurred between all three Yq paralogs (P1proximal, P1distal and IR1Yq), while an additional three regions were identified where conversion appears to have occurred between only one paralog of P1 and IR1Yq. No apparent gene conversion events were observed between IR1Yp and any of the three Yq paralogs (Figure 8.4B). As analysis of sequence from one individual chromosome has been carried out it is only possible to identify conversion events to the derived allele as due to the homogenising effect of gene conversion, any conversion events to the ancestral allele are effectively “invisible”. As only conversion events to the derived allele can be identified, the observation of 72 apparent conversion events between all three Yq paralogs seems high, especially if gene conversion is conservative of the ancestral state as previously observed between palindromes. As it is possible that either P1 or IR1 has arisen through a human-specific duplication event, another possible explanation for these observations could be that duplication rather than gene conversion has resulted in the identical alleles observed between all three Yq paralogs. Even if this is the case there is evidence from sequence comparisons of three gene conversion events occurring between IR1Yq and P1 and it is possible that gene conversion occurs following duplication.

8.3.5, Evidence of conservative gene conversion

Analysis of snPSVs and sequence alignments suggests that gene conversion does not occur between the paralogous arms of IR1, but gene conversion does occur between the IR1Yq paralog and P1. Several observations made from sequence alignments of blocks A and C suggest that gene conversion between IR1Yq and P1 is conservative of the ancestral state. Firstly, comparison of the number of PSVs resulting from mutation on IR1Yp and IR1Yq reveals that in block B roughly equal numbers of PSVs arise from mutation on Yp and Yq, while in blocks A and C, which exhibit paralogy with P1, >95% of PSVs arise due to mutation on Yp.

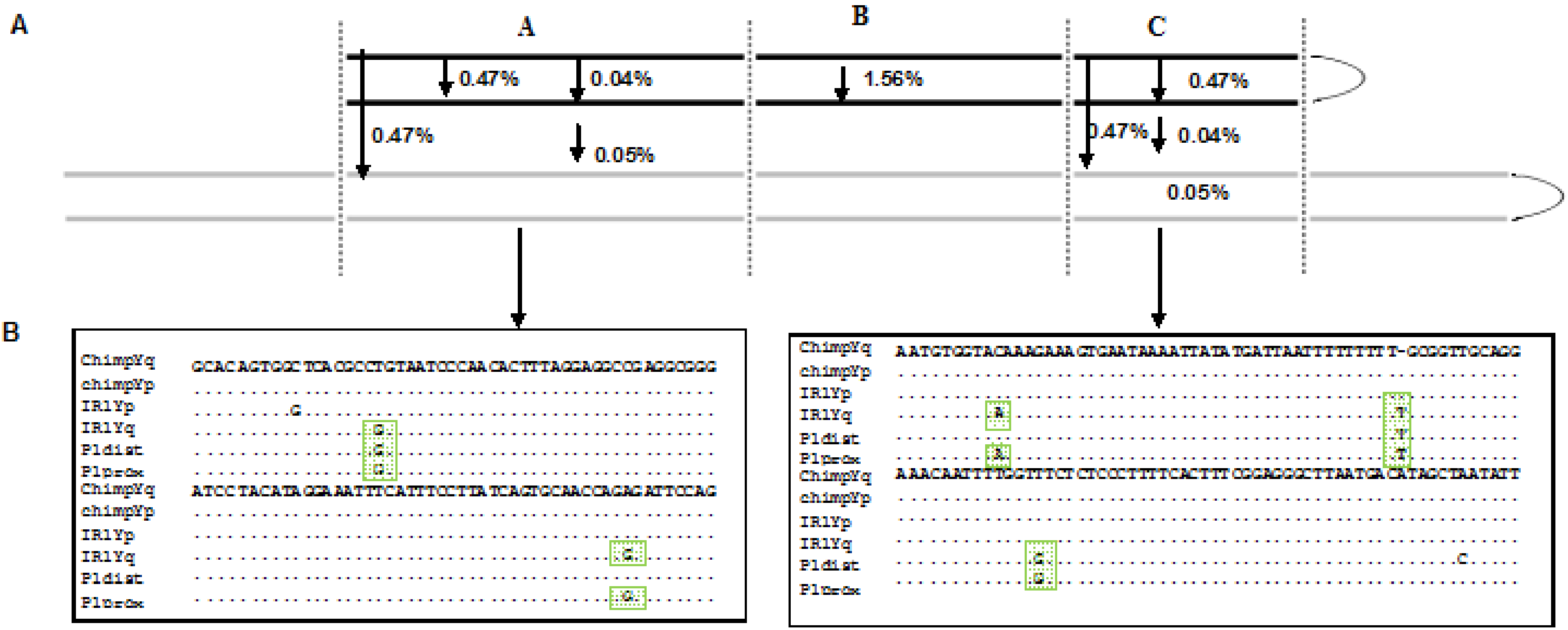

A
A
B
C
0.47%
0.04%
0.47%
0.05%
1.56%
0.47%
0.47%
0.04%
0.05%
B
ChimpYq
chimpYp
IR1Yp
IR1Yq
P1dist
P1prox
GCACAGTGGCTCACGCCTGTAATCCCAACACTTTAGGAGGCCGAGGGGG
G
ATCCTACATAGGAAATTTCATTTCCTTATCAGTGCAACCAGAGATTCCAG
AATGTGGTACAAAGAAACTGAATAAAATTATATGATTAATTTTTTTT-GGGGTTGCAGG
AAACAATTTTGGTTTCTCTCCCTTTTCACTTTCGGAGGGCTTAATGACATAGCTAATATT
C

Figure 8.4: Sequence analysis of regions A and C

a) Sequence divergence calculations for regions A and C revealed low divergence of 0.04% between IR1Yq and the P1 paralogs which is similar to the 0.05% divergence observed between the arms of P1. Divergence between IR1Yp and P1 was significantly higher at 0.47% (P=0.0001, 2-tailed Fisher exact test) and is similar to divergence of 0.45% observed between the paralogs of IR1 which have been shown not to undergo gene conversion.

b) From sequence comparisons no evidence of gene conversion between IR1Yp and both P1 paralogs was observed, while 3 regions were identified where gene conversion appears to have occurred between IR1Yq and one P1 paralog.

The significantly lower number of PSVs arising due to mutation on Yq (P=0.0001 chi square test) suggests that gene conversion between IR1Yq and P1 is conservative of the ancestral state. Conversion of a derived allele into the IR1Yq sequence would increase the number of PSVs which arise due to mutation on IRYq in blocks A and C, while conversions to the ancestral allele would be invisible and fewer PSVs would arise due to mutation on Yq.

Secondly, the significantly lower divergence between IR1 paralogs in the regions which have been shown to undergo gene conversion with P1 compared to the region of unique Yp-Yq identity, despite no evidence of gene conversion being observed between the IR1 paralogs, also suggests that conversion is conservative of the ancestral state (P=0.0001, 2-tailed Fishwer exact test). Conversion of PSVs which have arisen due to mutation on Yq back to the ancestral allele would lower divergence between the IR1 paralogs and only mutations occurring on IR1Yp would contribute to the overall sequence divergence between paralogs.

8.4, Discussion

Despite >99% sequence similarity being observed between IR1 paralogs no evidence of gene conversion in recent human evolution has been observed. Phylogenetic analysis of variants within the region of unique Yp-Yq identity reveals that once a variant is introduced into the IR1 sequence it remains in all observed descendent chromosomes. This and the presence of multiple PSVs which appear to be "fixed" since the MRCA of the Y phylogeny strongly suggest that conversion does not occur at an appreciable rate between IR1 paralogs.

Previously evidence of gene conversion has only been sought between paralogs forming part the same structure (palindrome or IR). Evidence of conservative gene conversion between IR1Yq and P1 provides evidence that gene conversion occurs between different classes of Y chromosome paralog and is not limited to paralogs which form part of the same structure (for example palindrome-palindrome). Analysis of the chimpanzee reference sequences suggests that either IR1 or P1 has arisen as the result of a duplication event in humans, and although the chimpanzee sequence provides enough information to infer ancestral state it offers no information as to whether low divergence is due to a human-specific duplication or due to gene conversion. Assuming that IR1 has not arisen as the result of a recent duplication event in humans, conservative gene conversion could also explain the low divergence observed between IR1 paralogs despite the lack of detectable gene conversion in recent human evolution. From this analysis it is apparent that gene conversion has occurred between the IR1Yq paralog and P1 following duplication. Evidence of NAHR occurring between IR1Yq and P1 also raises the possibility that other rearrangements mediated by NAHR, such as duplications, deletions and inversions could also potentially occur within this region.

While evidence of gene conversion between IR1 and P1 has come from analysis of the reference sequence, with recent advances in sequencing technology it may be possible to resequence on a large scale the regions of IR1 which display paralogy with IR1 to provide direct evidence of gene conversion between P1 and IR1 and also determine the rate of gene conversion between these regions. Publication of the 1000 Genomes Project may also provide further evidence of gene conversion between IR1Yq and P1.

Chapter 9: Seeking evidence of gene conversion between the chromosome and the X -transposed region on Yp

9.1, Introduction

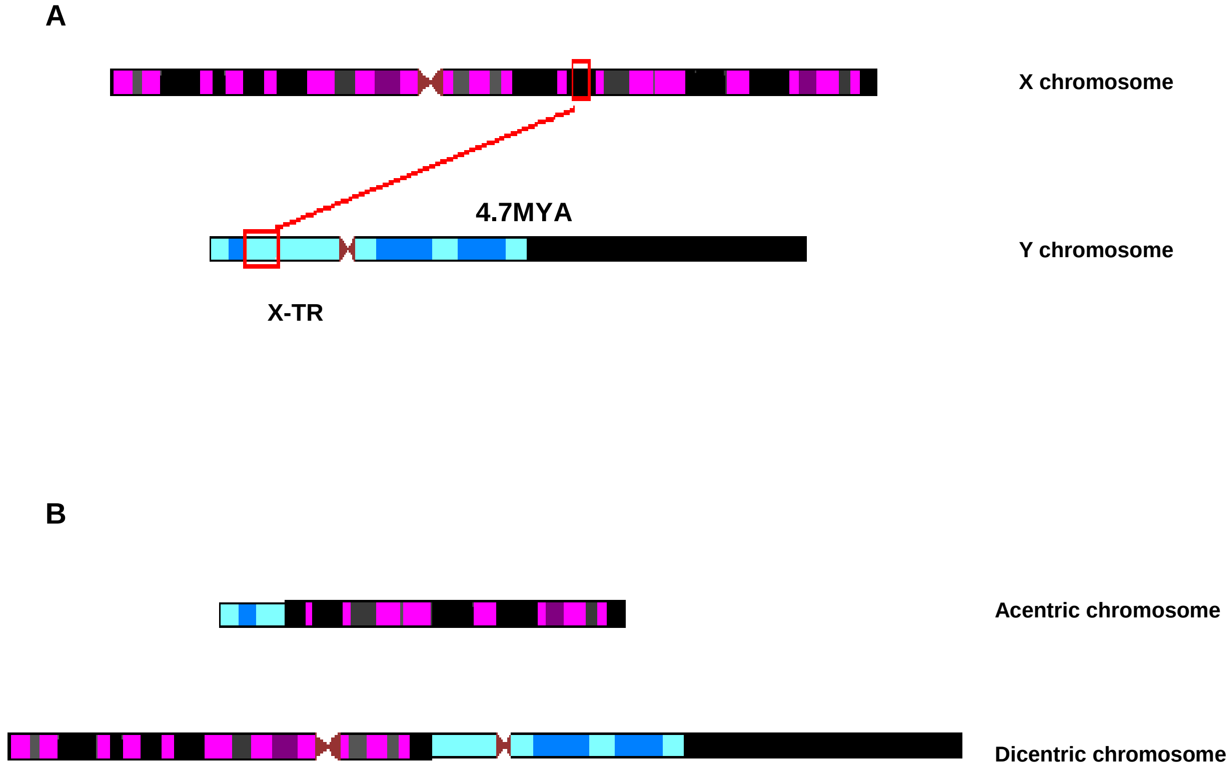
Approximatly 4.7MYA (Ross et al. 2005) a transposition occurred between the X and Y chromosomes which resulted in the transfer of a 3.8Mb block of sequence from the X to the Y chromosome (Figure 9.1A). Since this transposition, a series of inversions and deletions on the Y chromosome have shortened the transposed sequence block to 3.38Mb with a 200-kb segment which has become separated from the main sequence block. Three genes are located within the X transposed region: *TGIF2LX/Y*, *PCDH11X/Y* and a newly identified gene, each of these genes has a functional copy on the X chromosome. Sequence similarity between Xq21 and the XTR of the Y chromosome has been estimated to be approximately 98.78% excluding insertions and deletions and this high degree of sequence similarity suggests that there is the potential for gene conversion to occur.

9.1.1, Gene conversion between the X and Y chromosomes

Since this section of the study was completed, three papers have been published which have sought evidence of gene conversion occurring between the X and Y chromosomes. Rosser et al. (2009) and Cruciani et al. (2010) found evidence of gene conversion occurring between the X-Y homologous *PRKX/Y* genes at the site of a known XY-translocation hotspot, while Trombetta et al. (2009) provided evidence of gene conversion between the *VCX* and *VCY* genes. The rate of X-to-Y gene conversion between the *PRKX/Y* and *VCX/Y* genes is 1-2 orders of magnitude slower (Cruciani et al. 2010) than the Y-Y rate of gene conversion (2.2x10^{-4}) observed between palindrome arms (Rozen et al. 2003). The slower rate of gene conversion between the X-Y homolgous genes would be expected as the genes are located on different chromosomes and less likely to come into contact during meiosis.

9.1.2, Translocations between the X and Y chromosomes

Cytogenetic studies have shown that during meiosis pairing of the X and Y chromosomes is not restricted solely to the PARs and that the two chromosomes sometimes align along the entire chromosome length. Various X-Y translocations are ascertained in patients with sex differentiation disorders such as XX males, (Affara et al. 1986) which shows that recombination intermediates form between the X and Y chromosomes outside the PARs. While X-Y translocations are known to occur, no cytogenetic evidence of translocation between the XTR and Xq21 have been reported. Given the location of the sequences on Yp11 and Xq21 translocations would be expected to result in the production of either acentric or dicentric chromosomes (Figure 9.1B). While acentric chromosomes would be highly unstable and lost during meiosis, dicentric chromosomes are not always unstable and some such as those caused by the Robertsonian translocation usually remain stable (Page and Shaffer 1998). Dicentric chromosomes can be maintained via inactivation of one centromere and therefore remain viable (Howell, Roberts, and Beard 1976; Sarto et al. 1986). While translocations may in some cases result in non-viable chromosomes, gene conversion would not result in chromosomal rearrangements and would therefore be expected to persist in the general population.



Figures not drawn to scale

Figure 9.1: Structure of the XTR of the Y chromosome and hypothetical products of NAHR

a) A transposition event between the X and Y chromosomes 4.7 million years ago (MYA) resulted in the transposition of a 3.4mb block of sequence from Xq22 to Yp11,

b) No cytogenetic evidence of translocations between Xq22 and the XTR of the Y chromosome has been observed. Translocations would produce either a highly unstable acentric chromosome or a dicentric chromosome which may remain viable through inactivation of one centromere.

9.1.3, Chapter aims

This Chapter aims to seek evidence of gene conversion occurring between the XTR of the Y chromosome and Xq21. Phylogenetic analysis of snGSVs identified between the *TGIF2LX/Y* genes and the surrounding region of sequence similarity will be carried out in the first part of this Chapter. In the second part of this chapter analysis of X-Y gametologous microsatellites from across the XTR will be carried out.

9.2, Materials and methods

Analysis was carried out as described in chapter 2, with the following exceptions.

9.2.1, DNA samples

167 DNAs from the CEPH-HGDP panel were used for typing of the M1 and M3 microsatellites. M3 was further typed in 272 additional male and female CEPH-HGDP samples (supplementary table S9.1).

9.2.2, Chromosome specific-typing

9.2.2.1, PCR

PCR was carried out using 1-2µl of WGA DNA the buffer of et al. (Jeffreys, Neumann, and Wilson 1990), and 1U Kappa Taq. PCR conditions included initial denaturation at 95°C for 3 minutes followed by 94°C 30s, 60°C 30s, 70°C 60s, for 25 cycles. PCR products were treated with 4µl (4U) SAP and 1µl (1U) ExoI and incubated at 37°C for two hours to remove unincorporated dNTPs and primers.

9.2.2.2, Sequencing

Samples were sequenced using 1 pM primers and Big Dye terminator v 3.1 (Applied Biosystems) at the Protein and Nucleic Acid Laboratory (PNACL) at The University of Leicester. Reaction conditions were 94°C 30s, 96°C 10s, 50°C 5s 60°C 4 min for 30 cycles Unicorporated dye terminators were removed by adding 2µl of 2.2% (w/v) SDS and boiling at 96°C for 5 min and placed through an EDGE spin column. Sequence data were obtained using ABI 3730xl capillary electrophoresis apparatus (Applied Biosystems). Sequence analysis was carried out using Sequence Analysis v3.7 (Applied Biosystems) and sequences aligned using ClustalW.

9.2.2.3, Microsatellite typing

Secondary PCR was carried out using 1µl of purified PCR product as template, the Buffer of et al. (Jeffreys, Neumann, and Wilson 1990), 1U kappa Taq and 5µM primers in a 10µl reaction. PCR conditions were initial denaturation of 95°C for 5mins followed by 94°C 20s, 60°C 10s, 72°C 20s, for 15 cycles. Products were diluted and run on an ABI 3130xl Genetic Analyser against GeneScan™ 500 LIZ™ Size Standard. Analysis was carried out using genemaper v4 software (Applied Biosystems).

Table 9.1: Primers used in the analysis of the XTR

PCR Primers			
Primer name	**Forward primer 5´- 3´**	**Reverse primer 5´- 3´**	**Tm**
TGIf-GENE-PCR1	CTGGAGAGCTATAAACTGCG	ACGGACTCGGCTGGCAAG	60
TGIf-GENE-PCR2	CGGCCCGGCTGAGACC	CTGGTTGCTTCTCTCTTGAC	60
TGIf-GENE-PCR3	GGCCACAAAACGGGCAAAG	CCTTAGTGTCTTTCACCTGG	60
TGIF_1	TTCCCACAAGAGTGGTGTAG	AACATTTCATCTCAGACACAAG	60
TGIF_2	CCTTGGGTTAATTTGGAGGC	AGGGACCAGGAGGAGACC	60
TGIF_3	ACTTCAGTACCAATTAGTTGAC	CCCTGGTTGGGCTCTGTC	60
TGIF_4	CGTCACATCATGTATTAGAGAG	GGTCTGTTCTAGGATCTTTG	60
TGIF_5	GAGGAGGTCCTGCATCCTC	GGGAGGAAGGTGGACCTC	60
TGIF_6	GTTAACAGGTGCTTCTCAGG	CAGTGGAGGACCTTAAATGG	60
SNaPshot primers			
Primer name			
YH_GT	ACTATATTGAGAATCACAGA		
Microsatellite primers			
Primer name	**Forward primer 5´- 3´**	**Reverse primer 5´- 3´**	**Tm**
M1	6FAMGTTGCGGTCTGTGGGAGG	TGCAGACAGCCTATTCTG	60
M2	6FAMGGTGCTACACTGTCACAGG	AATTATACCATATTTATCTACCTC	60
M3	8AATGTGGTAGATATACACCATG	TAGCCATACTGACTGGTGTG	60
M1X	ATAAAGCAGGCAGGAAAACAC	ACAGTAGACTAGACTAAGCAG	60
M1Y	TCTAATCAGCTTCCAGTGAAC	ACAGTAGACTAGACTAAGCAC	60
M3X	AATCATTCTATTATAAAGACACAC	ACTAGTTTACATTCTCCCCAC	60
M3Y	GGAATATAAATCATTCTATTG	ATACTCAGCAATTAAATTGCTAG	60

9.3, Results

9.3.1, Interspecies sequence divergence between the XTR sequences

As the X to Y transposition occurred following human and chimpanzee speciation, relevant outgroup sequence is only available for the chimpanzee X chromosome. This is useful in the study of gene conversion as it can provide evidence of the ancestral state of human GSVs. Sequence analysis carried out in Chapter 4 revealed an average divergence of 1.21% across the XTR. Interspecies sequence divergence between human and chimpanzee X chromosome sequences was calculated as 0.98%, while divergence between the chimpanzee X chromosome and the human XTR was calculated as 1.53%. Interpreting the significance of divergence values between the XTR and Xq21 is difficult as the sex chromosomes have very different mutation rates with the Y chromosome mutating faster than the X chromosome (Lahn, Pearson, and Jegalian 2001). Since speciation, on average the X chromosome has diverged by 1%, the autosomes by 1.2%, while divergence between the Y chromosomes is much higher a 1.9% (Ebersberger et al. 2002). As gene conversion is expected to homogenize two sequences despite the very different mutation rates, frequent gene conversion would be expected to reduce the divergence between the human sequences; however, the degree of similarity observed between sequences will also depend on the rate of gene conversion.

9.3.2, The problem of determining phase

When seeking evidence of gene conversion between the X and Y chromosomes there is also the issue of phase. As previously discussed in chapter 5, when a variant is located within large segments of sequences which display a high degree of sequence similarity, co-amplification of both sequences occurs during PCR. On the Y chromosome this is not so much of an issue when identifying gene conversion events as exchange between paralogs can easily be observed. When typing snGSVs the problem of phase is more of an issue, especially for GSVs which are observed to be in the pseudoheterozygous state. Due to co-amplification, when a pseudoheterozygous GSV is identified there is no direct information as to which chromosome the derived allele lies on. When sequencing X-Y homologous genes it is possible to sequence each gene independently to provide evidence of phase, while in large stretches of DNA advantage can be taken of differences between the X and Y sequences and chromosome-specific PCR can be performed.

9.3.4, Seeking evidence of gene conversion between the *TGIF2LX-Y* genes

The *TGIF2LX and TGIF2LY* gene sequences were obtained from all database sequences and aligned with the chimpanzee *TGIF2LX* sequence. On the Y chromosome *TGIF2LY* spans 957bp and consists of 2 exons while the *TGIF2LX* gene is smaller at 943bp but also contains 2 exons. As the transposition event is known to be human-specific, the *TGIF2LX* gene has no gametolog on the chimpanzee Y chromosome and cannot be undergoing gene conversion, thus providing evidence of ancestral state.

From sequence comparisons a total of twelve snGSVs (G1-G12) were identified between the *TGIF2LX and TGIF2LY* gene copies (Table 9.1). Eleven of the GSVs were present in the CV, JW and reference sequences while one GSV (G2) was present only in the Yh sequence. To identify additional GSVs and overcome the ascertainment bias associated with typing snGSVs, a 1-kb segment containing the *TGIF2LX* and *TGIF2LY* genes were simultaneously sequenced in eight individuals representing the major haplogroups of the Y phylogeny and sequences were aligned along with the database sequences using clustalW (Higgins 2003; Larkin et al. 2007).

From sequencing of eight individuals, no additional GSVs were identified, and ten of the twelve GSVs identified from the database sequences were present in the pseudoheterozygous state in all eight chromosomes. Two GSVs, the G2 G/T GSV identified from the Yh sequence and the G9 A/G GSV identified from the reference sequence, were observed in the pseudoheterozygous and pseudohomozygous ancestral states while no pseudohomozygotes for the derived allele were observed for either GSV. As gene conversion is expected to be rare due to the large physical distance which separates the two genes, a larger sample set may be required for gene conversion events to be identified. To investigate these two GSVs further, the G2 and G9 GSVs along with the adjacent GSVs in the reference sequence were typed in a panel of 64 males representing the major haplogroups of the Y phylogeny using the SNaPshot minisequencing protocol (Figure 9.2).

9.3.4.1, Phylogenetic analysis of variant G9

Comparison of the G9 A/G GSV in humans and chimpanzee reveals the derived G-allele to have arisen on the X chromosome. Phylogenetic analysis showed the GSV to be present in the pseudoheterozygous A/G state in 34 (47%) of males and the ancestral pseudohomozygous A/A state in 38 (53%) of males typed, while no evidence of gene conversion of the *TGIF2LY* gene to the derived G-allele was observed. The adjacent G8

GSV	Y-allele	X-allele	Chimpanzee-allele
G1	T	G	G
G2	T	G	G
G3	T	G	T
G4	A	G	G
G5	G	A	A
G6	-	C	C
G7	A	G	A
G8	A	G	G
G9	A	G	A
G10	T	A	A
G11	T	C	T
G12	G	A	A

Table 9.1: snGSVs identified between the *TGIF2LX/Y* gene copies

Twele snGSVs were identifed between the *TGIF2LX/Y* gene copies and the ancestral state of each CSV was determined from comparisons with the chimpanzee sequence. Eight GSVS were shown to arise due to mutation on the Y chromosome while the remaining four arise due to variation on the X chromosomes.

and G10 PSVs which are located 5bp and 4bp either side of G9 in the reference sequence were both observed to be in pseudoheterozygous state in all males analyzed. Given that the derived allele is known to have arisen on the X chromosome this most likely represents an X chromosome SNP which is not converted into the Y chromosome sequence in the ancestors of the studied sequences.

9.3.4.2, Phylogenetic analysis of variant G2

Comparison of the G2 G/T GSV in humans and chimpanzee shows the derived T-allele to have arisen on the Y chromosome. From phylogenetic typing the GSV was observed to be in the pseudoheterozygous G/T state in 14 (19%) of males and the ancestral pseudohomozygous G/G state in 53 (81%) of males, while no pseudohomozygotes for the derived T-allele were identified. Phylogenetic analysis revealed all individuals from haplogroups N and O and their subclades to be in the pseudoheterozygous state suggesting that the founder of the NO branch of the Y phylogeny carried the T-allele. Two additional chromosomes, one, from haplogroup J2(xJ2a2b) and one from haplogroup R1b1b1, were also observed to be in the pseudoheterozygous state, while all remaining haplogroups were shown to be pseudohomozygous for the ancestral allele This distribution across the phylogeny is interesting and could have several possible explanations. The first explanation is that the derived T-allele has arisen within the ancestor of superhaplogroup F of the Y phylogeny and subsequent gene conversions to the ancestral allele has occurred in all descendent Y chromosomes other than those from haplogroups J2(xJ2a2b) N,O and R1b1b1. Although this is possible it seems unlikely, as gene conversion is expected to be slow. If it were the case, then assuming that the mutation has arisen at haplogroup F, a minimum of seventeen conversion events to the

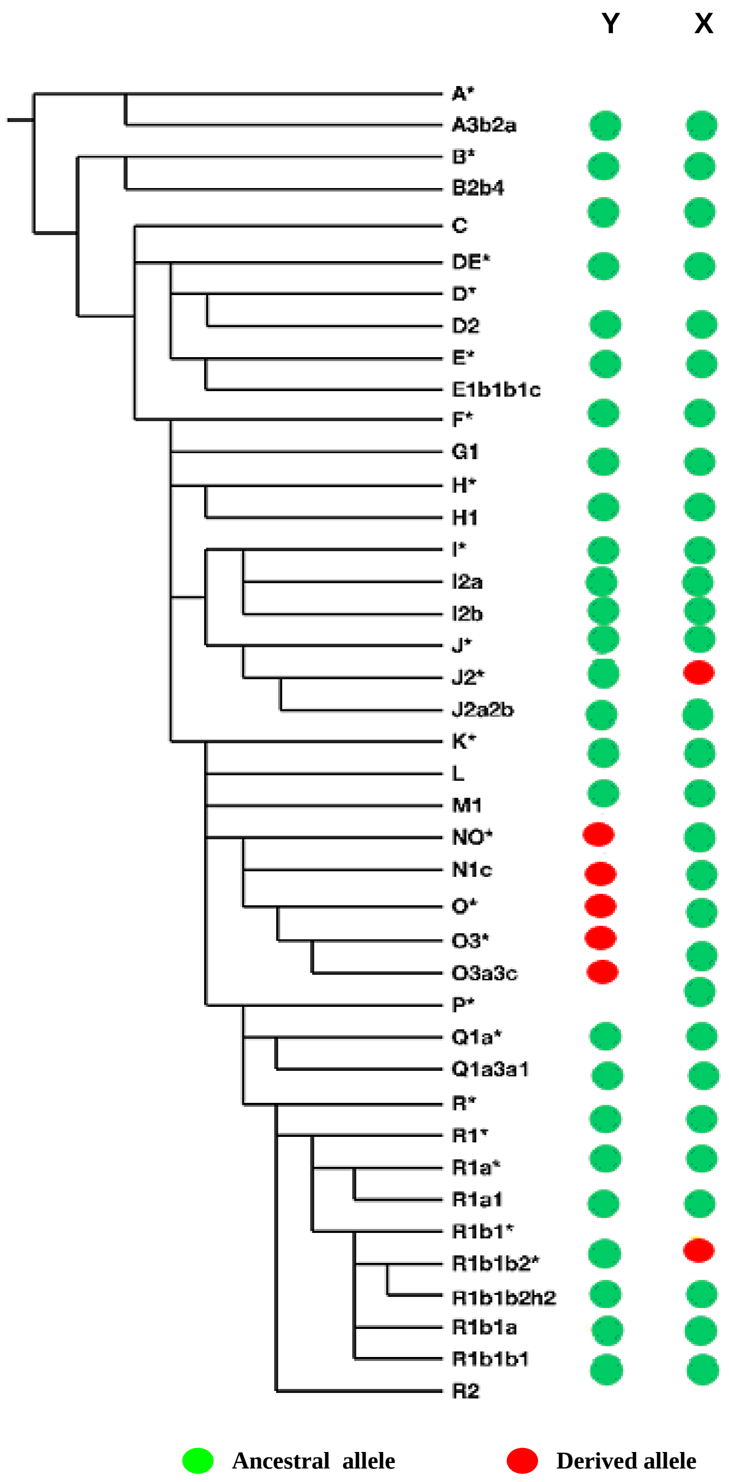


Figure 9.2: Phylogentic analysis of the G2 GSV
Sequencing reveals that the hg N and O chromosomes carry the derived T-allele on the Y chromosome as observed from the Yh sequence, while the haplogroup J2 and R1b1b2 chromosomes carry the derived T-allele on the X chromosome. This suggests that gene conversion has occurred between the *TGIF2LX* and *TGIF2LY* genes.

ancestral allele must have occurred for this distribution to be observed. If such frequent gene conversion were occurring then conversion events would be expected to be observed at other GSV sites especially at the site of the C1 GSV which lies 18bp upstream of C2.

Another possible explanation for this distribution is that three independent Y-chromosomal mutations have occurred at the same nucleotide within three different haplogroups of the Y phylogeny. This also seems unlikely as the GSV does not lie at the site of a highly mutable CpG dinucleotide and given that nucleotides other than those of CpG dinucleotides have a low average mutation rate of 2×10^{-8} per nucleotide per generation (Ballard et al. 2005) hypermutation seems highly unlikely.

The previous two explanations for this haplotype distribution have assumed that the derived T-allele is located on the Y chromosome as observed from the Yh reference sequence. However, as both genes were simultaneously sequenced the phase of the GSV on each chromosome is currently unknown and it is a possibility that the derived T-allele lies on the X chromosome rather than the Y. As all males carrying haplogroup N and O Y-chromosomes show the pseudoheterozygous state it is most parsimonious to assume that the T-allele is carried on the Y chromosome in these chromosomes. However, the possibility remains that the individuals from haplogroup J2(xJ2a2b) and R1b1b1 may carry the T-allele on the X chromosome, rather than the Y. This would be an interesting observation as it would suggest gene conversion has occurred between the X and Y chromosomes.

9.3.4.3, Determination of chromosome phase

In order to address the issue of phase for the G2 GSV, chromosome-specific sequencing was carried out. Primers were designed making use of GSVs which have previously been typed and are known to be heterozygous between chromosomes. To ensure chromosome specificity X- and Y-only somatic-cell hybrid DNAs were included as controls and the adjacent C1 and C3 GSVs which are known to be in the pseudoheterozygous state were also covered in the region sequenced.

Sequencing revealed that in all haplogroup N and O males the derived T-allele arises on the Y chromosome as observed in the Yh reference sequence. However, in the J2(xJ2a2b) and R1b1b1 individuals the derived T-allele lies on the X chromosome The adjacent C1 and C3 GSVs, which are known to be in the pseudoheterozygous state in these chromosomes, were shown to carry the allelic state observed from the reference sequence.

These data provide evidence of gene conversion occurring between the *TGIF2LX* and *TGIF2LY* genes; however, it has not been possible to determine the direction of gene conversion. Due to the diploid nature and lack of evolutionary phylogeny for the X chromosome it is not possible to determine when the derived T-allele arose on the X chromosome. On the Y chromosome, phylogenetic analysis shows the derived T-allele to have arisen within the founder of the NO branch of the Y phylogeny: however, it cannot be determined whether the derived T-allele originated on the Y chromosome or resulted from a conversion event which transferred the derived T-allele to the founder of the NO branch of the Y phylogeny.

9.3.5, Seeking evidence of gene conversion between the sequences surrounding *TGIF2LX-Y*

The *TGIF2LX* and *TGIF2LY* genes are situated within a 273-kb block of gametology which exhibits >98.79% similarity between the X and Y chromosomes. As evidence of exchange between the *TGIF2LX* and *TGIF2LY* genes has been demonstrated the sequence surrounding the genes was investigated to see if additional conversion events could be identified and the rate of gene conversion estimated.

Three regions ranging between 140bp and 540bp were resequenced in eight males carrying Y chromosomes representing the major branches of the Y phylogeny. Regions were selected based on the length of uninterrupted sequence between two GSVs identified from the reference sequence and covered a total of 51 snGSVs. Sequencing revealed all snGSVs identified from the reference sequence to also be present in the pseudoheterozygous state in all eight individuals and no additional GSVs were identified. Comparison with all available database sequences identified two additional GSVs, both in the Yh sequence. The first GSV located within the 240-bp sequence block consisted of a C/T GSV with the derived T-allele arising on the X chromosome. The second GSV located within the 540-bp sequence block consists of an A/G GSV with the derived G-allele arising on the Y chromosome. The Yh C/T GSV was observed in two individuals, carrying Y chromosomes from haplogroups E1b1b1c and F*. Chromosome-specific sequencing revealed that in both cases the X chromosome carries the derived T-allele and this GSV most likely represents an X chromosome SNP. The Yh A/G GSV was only observed in the Yh sequence and may represent a private mutation within this individual.

This study provides evidence that gene conversion occurs between the XTR and Xq21. Previous studies by Trombetta et al. (2009) did not find any evidence of gene conversion occurring between the *PCDH11X/Y* genes which are also located within the XTR which suggests that gene conversion between the XTR and Xq21 is not a frequent occurrence. Given that the XTR spans over 3Mb and that conversion events are expected to be rare, analysis of such small regions is very unlikely to identify conversion events. The presence of variants which appear to have become fixed during evolution means that extensive sequencing must be undertaken to overcome ascertainment bias and identify additional GSVs which have the potential to show gene conversion events.

9.3.6, Survey of X-Y gametologous microsatellites

Since some evidence of gene conversion has been observed in the XTR it was decided to survey the entire X-transposed region for microsatellites which may have the potential to identify further gene conversion events. Although identifying gene conversion events through this approach is more complicated than typing snGSVs it has been shown in Chapter 6 that gene conversion events can be identified through typing microsatellites which have >4 repeat difference between copies. Alignments of the reference sequence carried out in Chapter 4 revealed that some X-Y homologous microsatellites in the XTR have up to 25 repeat differences between copies and such large differences in repeat number may make it possible to identify gene conversion events: however, at the same time it may also be possible that such large differences between repeat number, which effectively create a large indel between sequences, may also disrupt HJ formation and prevent gene conversion from occurring.

9.3.6.1, Microsatellite typing

From sequence alignments carried out in Chapter 4, three microsatellites (M1-M3) with >13 repeat differences between X and Y copies were identified (Figure 9.3a). In all three microsatellites the smaller allele was located on the Y chromosome with the larger allele on the X chromosome. For the M2 microsatellite, the Y chromosome copy carries only seven repeat units and is not expected to be variable while the M1 and M3 microsatellites both have nine repeats and are expected to be more variable as mutation is known to occur at a faster rate with increased repeat number (Carvalho-Silva et al. 1999).

To gain an understanding of the variability of each microsatellite, fluorescently labelled primers were designed to amplify both X and Y microsatellites simultaneously and each microsatellite was typed in a panel of 105 males and 93 females covering 17 different populations. Sequencing of each microsatellite was carried out using chromosome-specific primers which make use of fixed sequence differences between the two chromosomes and the repeat number for each microsatellite was determined. The preliminary results show that for all three microsatellites the Y-chromosomal copy is much less variable than the X copy, with the Y-allele being monomorphic in the M2 and M3 microsatellites, and only the 9 and 10 alleles being observed for M1 (Figure 9.3b). These findings are similar to those published by Lopes et al. (2004) who analyzed seven independent X-Y homologous microsatellites in the region surrounding the *PCDH11X-Y* genes and found Y-linked microsatellites to be much less variable than X-linked microsatellites. From analysis of the data obtained by Lopes et al. (2004), no evidence of gene conversion was apparent between microsatellites surrounding a)

Microsatellite	X-copy	Y-copy	Repeat Difference
M1	TA(22)	TA(9)	13
M2	TA(23)	TA(7)	16
M3	TA(34)	TA(9)	25

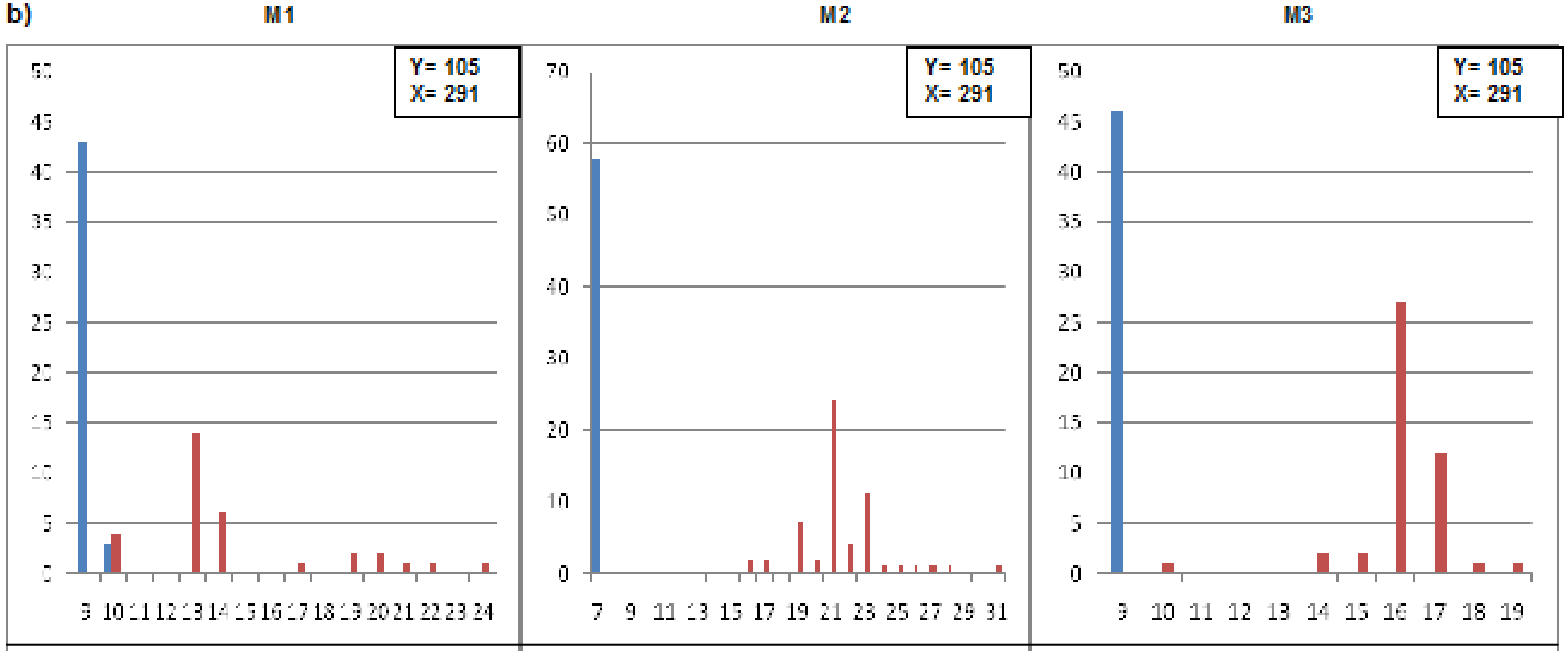


Figure 9.3: X- and Y-allelic diversity of three gametologous microsatellites.

a) Three X/Y homologous microsatellites were identified which displayed >13 repeat differences between X- and Y-linked copies.

b) For all microsatellites the Y-allele was predominantly monomorphic while the X-allele was more variable. For M2 the X-allele distribution ranged from 16-31 repeats, while M1 and M3 showed a gap of >3 repeats in the allele distribution with the smallest X-allele being the same size as the Y chromosome allele.

PCHD11X/Y and recently no evidence of gene conversion was observed between the *PCDH11X/Y* genes themselves (Trombetta et al. 2009).

While the Y-linked microsatellites are predominantly monomorphic, their X-linked gametologs are much more variable and most interestingly the M1 and M3 microsatellites both show a different pattern of X-allele distribution to the M2 microsatellite. While M2 has an X-distribution ranging from 16-31 repeats, the M1 microsatellite X-allele distribution ranges from 13 to 24 repeats in all but three individuals. These three individuals carry the 10-repeat allele, which precedes a gap of three repeats in the X-allele distribution. Similarly, for the M3 microsatellite the X chromosome allele distribution ranges from 14-19 repeats in all but one individual. This individual was observed to carry the 10-repeat allele preceding a four-repeat gap in the X-allele distribution. While gene conversion between the X and Y chromosomes could have resulted in the production of the 10-repeat alleles in the X chromosome allele distribution, it is also possible that these small alleles are outliers of the natural X-allele distribution, and the intermediate alleles have not been observed within the sample set. To investigate this further, both M1 and M3 were typed in an additional 204 individuals representing eight populations from the CEPH-HGDP panel (Cann et al. 2002). As the X-allele is of particular interest in this case an additional 100 females were also included in the sample set in order to increase coverage of the X chromosome.

9.3.6.2, Analysis of the M1 microsatellite (Figure 9.4a)

Further typing of the M1 microsatellite showed the Y–linked copy to have very low variability, with 103 chromosomes carrying the 9-repeat allele and one carrying the 10-

repeat allele. For the X-linked copy, extended typing identified ten chromosomes which carry the 12-repeat allele reducing the gap in the X-allele distribution to two repeats. As the intermediate alleles have been observed this distribution can now be more readily explained by mutation and the 10 X-repeat allele cannot be assumed to be the result of gene conversion. However, this does depend to some extent on the genealogical depth of the X-chromosomes sampled; due to lack of evolutionary phylogeny for the X-chromosome this remains unknown, which somewhat complicates the interpretation of these results.

9.3.6.3, Analysis of the M3 microsatellite (Figure 9.4b)

Further typing of the M3 microsatellite revealed the Y-linked copy to be monomorphic for the 9-repeat allele in all 104 Y chromosomes, while no additional X-linked alleles were identified. Two additional male individuals, one carrying the 9-repeat allele and one carrying the 10-repeat allele on the X chromosome were observed which still leaves a four- repeat gap in the X-allele distribution. This could suggest that the Y chromosome has converted the X to the 9-repeat allele, which has subsequently undergone a single-step mutation in one individual. Analysis of the genotype distributions across the sample set revealed the males carrying the TA(9:9) and TA(9:10) genotypes all to be from the African San population. Microsatellites from African populations are known to have significantly greater diversity ($P < 10^{-8}$) than non-African populations (Harpending et al. 1997) which further suggests that the observed 9- and 10-repeat X-alleles could be outliers in the natural X-chromosomal allele distribution. To investigate this further M3 was typed in an additional 130 individuals from four African populations along with 240 individuals from 9 non-African populations

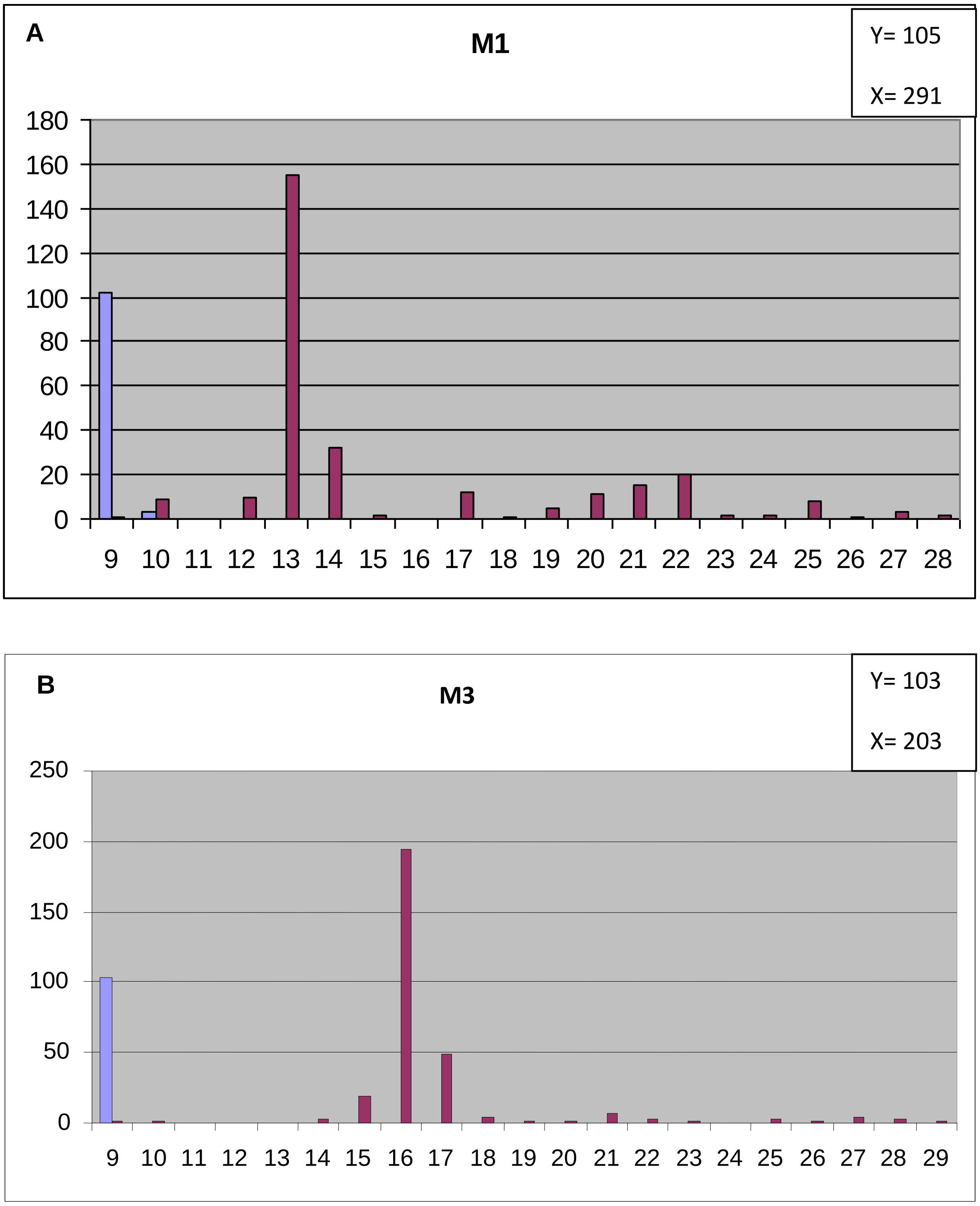


Figure 9.4: M1 and M3 microsatellite allele distributions

a) Extended typing of M1 identified 10 X chromosomes carrying the 12-allele which reduces the gap in the allele distribution to two. This distribution can now be explained by mutation.

b) Extended typing of M3 failed to identify any additional X-alleles in the distribution and a gap of four repeats difference remains. This gap cannot be explained by mutation alone and suggests that gene conversion may have created the smaller X-allele.

9.3.6.4, Population study of the M3 microsatellite

From the extended typing the X-allele distribution was observed to be much broader in African than non-African populations. The X-allele distribution for African populations ranged from 9 to 29 repeats with a mean allele size of 14, while in non-African populations allele size ranged from 14-30 repeats with a mean allele size of 16 (Figure 9.5). The 9- and 10-repeat X-alleles were only observed in central and southern African populations while no such alleles were observed outside of Africa or in the north African populations. The TA(9:9) and TA(9:10) genotypes were observed in four of six males available from the African San population. While it is unlikely that the TA(9:9) genotypes observed in this study have resulted from separate gene conversion events, it is possible that one conversion event has occurred and the observed 9-repeat alleles have descended from the converted ancestor while the 10-repeat allele has occured from a single-step mutation from allele 9. While it seems possible that gene conversion could have produced the (TA9:9) genotypes the possibility remains that this genotype is an outlier in natural distribution of alleles¸ and the intermediate genotypes have not been observed within this sample set.

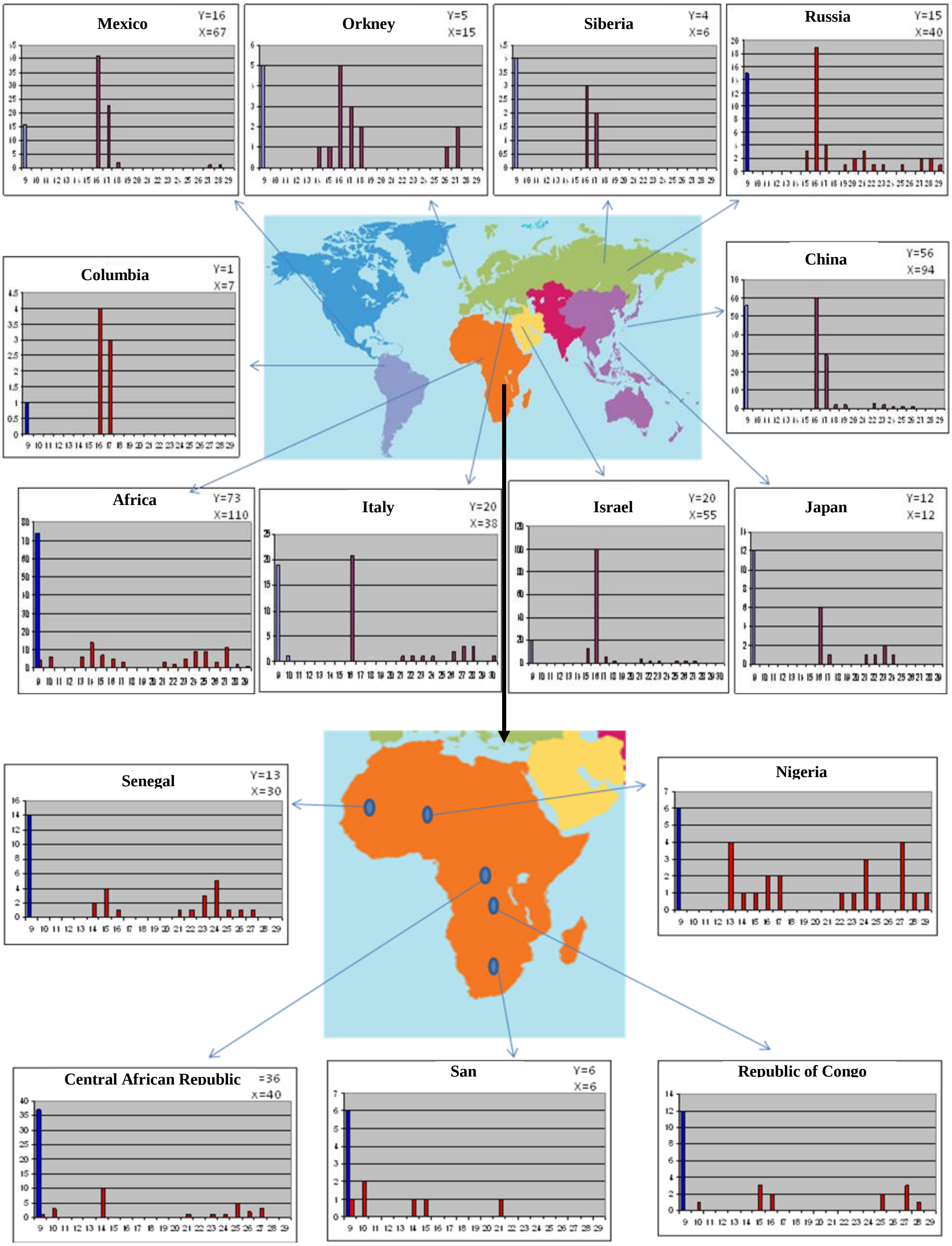


Figure 9.5: A population study of the M3 microsatellite
The outlying 10 X-allele was only observed in individuals from the Central African and African San populations, while no 10 X-alleles were observed in non-African populations.

9.4, Discussion

This study has provided evidence of gene conversion occurring between the *TGIF2LZX/Y* genes at the site of a single snGSV, while no evidence of gene conversion was observed from analysis of the regions surrounding *TGIF2LX/Y*. From analysis of X-Y duplicated microsatellites it has not been possible to determine that gene conversion occurres between Xq21 and the XTR. Analysis of the M3 microsatellite identified a possible historical Y-to-X gene conversion event in central and south African populations. However, as African populations are known to generally be more diverse than non-African populations this haplotype could be part of the natural X-allele distribution. Many studies of microsatellite mutation have reported an upward mutation bias (Brinkmann et al. 1998a; Karafet et al. 1998; Carvalho-Silva et al. 1999) suggesting that microsatellites will be larger in African populations and it has been suggested that that mean microsatellite length is higher in African populations than non-African populations (Amos, Flint, and Xu 2008). While it is generally thought that differences in microsatellite length arises from replication slippage, Amos (2009) suggests that microsatellites will reach a certain length and then contract causing expanded populations to carry relatively shorter microsatellite alleles and this could explain the small 9- and 10- alleles which are only observed in the African populations.

From analysis of X-Y duplicated microsatellites carried out in this study, the X- linked copy was observed to be much more variable than the Y-linked copy which is a similar finding to those of Scozzari et al. (1997), Karafet et al. (1999) and Lopes et al. (2004) which suggests that this may be true of X-Y duplicated microsatellites in general. This observation is interesting as at the time of transposition the X and Y chromosomes would carry identical numbers of repeats for each microsatellite. As microsatellites mutate though replication slippage and as the Y chromosome undergoes more replications via spermatogenesis, Y-linked microsatellites might be expected to undergo more mutation than X-linked microsatellites. In fact the opposite is observed with the Y linked microsatellite showing very little variation since the MRCA of the Y phylogeny. It has been suggested that the lower diversity of Y-linked copies may reflect a small X-allele being transferred to the Y chromosome at the time of transposition, which has either led to the stabilization of the microsatellite or loss of polymorphism on the Y chromosome (Carvalho-Silva et al. 1999).

Although evidence of gene conversion has been observed between the XTR and Xq21 it is not likely to be at a rate which will significantly alter the estimation of transposition time. Previous studies on *PCDH11X/Y* (Trombetta et al. 2009) and the surrounding regions have not provided evidence of gene conversion. The presence of GSVs which have become fixed since the MRCA of the Y phylogeny as well as the lack of described translocations between Yp11 and Xq21 could suggest that NAHR between these regions is not a frequent occurrence. As only one conversion event has been identified the rate of gene conversion could not be determined and a more detailed analysis of these regions would need to be carried out in order to determine if gene conversion is likely to significantly alter estimations of transposition time.

Chapter 10: Determining a direction of gene conversion.

10.1, Introduction

Many studies of gene conversion in diverse species have provided evidence that gene conversion is directional, and this appears to have different effects in each species. Several studies of gene conversion in pathogens have revealed that gene conversion favours the derived allele (Brayton et al. 2007) a mechanism which is thought to alter protein coat expression to escape the immune response (Brayton et al. 2007). In mammalian genomes gene conversion occurring between duplicated genes has been shown to favour the incorporation of G and C alleles into a sequence and this is believed to increase the stability of AT rich regions making them less prone to mutation (Galtier et al. 2001).

On the Y chromosome differences in the direction of gene conversion have been observed between different classes of Y chromosome paralog. Bosch et al. (2004) have provided evidence of significantly more proximal-to-distal conversion events between directly orientated HERV sequences and Trombetta et al. (2009) have recently suggested that the *VCY* genes act as a sequence acceptor from *VCX* during gene conversion. Rozen et al. (2003) have suggested that gene conversion between palindrome arms is conservative of the ancestral sequence. In this study evidence to support conservative gene conversion has been obtained through calculations of interspecies sequence divergence and analysis of snPSVs.

10.1,1, Chapter aims

While the data obtained in this study appears to provide evidence of conservative gene conversion, as other biases in the direction of gene conversion have been observed this chapter will also discuss these possible directions as an alternative explanation to conservative gene conversion.

10.2, Results

10.2.1, Analysis of conservative gene conversion between Yq paralogs?

In this study analysis of snPSVs identified between the arms of P6 suggested that gene conversion between palindrome arms may be conservative of the ancestral sequence. This conclusion is supported by divergence calculations which show that interspecies sequence divergence between palindrome arms is significantly lower (P=0.0001, two-tailed Fisher exact test) than that between the spacers, which are non-duplicated and cannot undergo gene conversion. While typing a subset of snPSVs appears to support these observations it is possible that other biases in the direction of gene conversion may also explain these observations. This section will bring together observations made in other Chapters to determine whether gene conversion may be conservative of the ancestral state.

Interspecies divergence calculations carried out for two additional palindromes also suggest that gene conversion may be conservative. Divergence calculations carried out for P8 suggests that gene conversion occurs rapidly between palindrome arms, while the significantly lower (P=0.0001, 2-tailed Fisher exact test) interspecies divergence observed between palindrome arms (1.7%) in comparison to the spacer region (3.9%) suggests that gene conversion may be conservative of the ancestral sequence. Phylogenetic analysis of two P8 snPSVs has provided limited evidence of conservative gene conversion: however, as gene conversion between the arms of human P8 appears to be very rapid, identifying conversion events has been difficult. Similarly bioinformatic analysis of the P7 reference sequence reveals the interspecies divergence between palindrome arms to be significantly lower (at 0.21%) than the 3.27% divergence observed between the spacers (P=0.0001, 2-tailed Fisher exact test).

This, along with very low divergence of 0.01% observed between human palindrome arms, suggests that conservative gene conversion also occurs rapidly between the arms of P7: however, as very low divergence and relatively few PSVs were identified between the palindrome arms, phylogenetic analysis of PSVs identified within his palindrome was not carried out.

Additional evidence of conservative gene conversion has also come from analysis of IR1. While no direct evidence of gene conversion was observed directly between the IR1 paralogs, sequence analysis has provided evidence of gene conversion occurring between the IR1Yq paralog and paralogs of P1, but not between IR1Yp and P1. Several observations made in this part of the study suggest that gene conversion occurring between IR1Yq and P1 is conservative of the ancestral state.

Firstly, in the region of IR1 which exhibits only unique Yp-Yq sequence identity, roughly equal numbers of PSVs arise due to mutation on Yp and Yq, while in the regions which have been shown to undergo gene conversion with P1, >97% of PSVs arise due to mutation on Yp. In the absence of gene conversion equal numbers of mutations would be expected to arise on Yp and Yq across the entire paralog as observed in the region of unique Yp-Yq identity. The reduced number of PSVs arising on Yq in the regions which have been shown to undergo gene conversion with P1 can be explained by conservative gene conversion. Conversion of a PSV which has arisen due to mutation on IR1Yq, back to the ancestral allele would effectively “erase” a PSV site allowing only mutations arising on IR1Yp to be observed. In contrast, conversion of a derived allele arising on IR1Yq into the P1 sequence would create more PSVs in the IR1 sequence due to mutations on Yq. Similarly, conversion of a derived allele from P1 into the IR1Yq sequence would also cause more PSVs to arise due to mutation on Yq (Figure 10.1). Secondly, the relatively low divergence between IR1 paralogs since the MRCA of the Y phylogeny despite the absence of gene conversion, suggests that conversion between IR1Yq and P1 has been conservative. Conversion of "mutations" arising on Yq to the ancestral allele would lower divergence between the IR1 paralogs as only mutations occurring on IR1Yp will contribute to the overall divergence between IR1 paralogs.

Data produced in this study support the findings of Rozen et al. (2003) that gene conversion is conservative of the ancestral state; however, as other biases in the direction of gene conversion are also known to occur, these will also be examined to see if they can offer an alternative explanation for the observations made in this study.

10.2.2, Analysis of proximal-to-distal gene conversion

Bosch et al. (2004) have previously shown through a phylogenetic analysis that gene conversion between two HERV sequences which flank the *AZFa* region on Yq favours proximal-to-distal conversion, with 22 proximal-to-distal conversion events being identified, compared to only one distal-to-proximal conversion event. It may be possible that similar directional biases also apply to other types of Y chromosome paralog.

Analysis of all snPSVs typed in this study reveals that, of the eight PSVs which have been shown to undergo frequent gene conversion, six (C6, C7, C11, C17, C24 and C40)

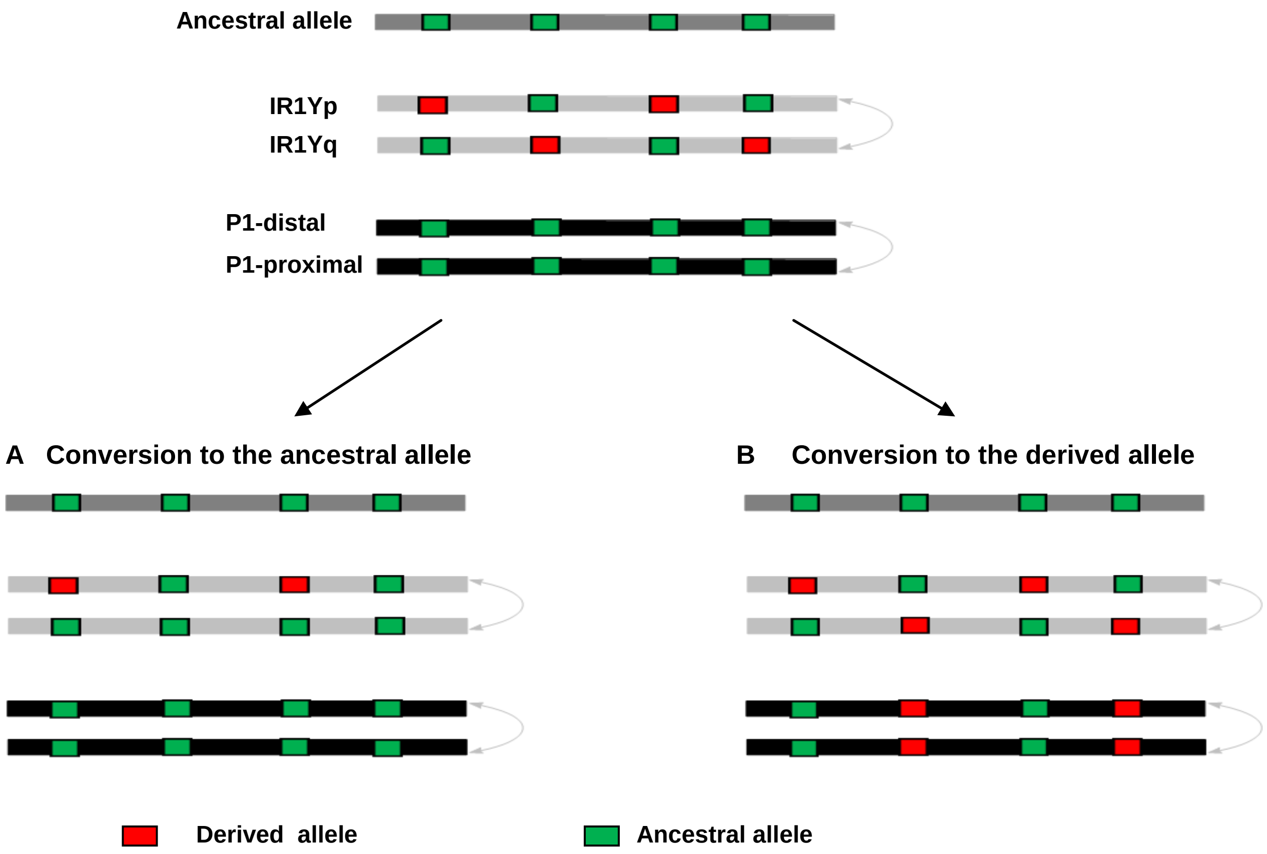


Figure 10.1: Evidence of conservative gene conversion between IR1Yq and P1
The lower number of PSVs arising due to mutation on IR1Yq suggestes that gene conversion is conservative of the ancestral state. Conversion of a mutation arising on IR1Yq back to the ancestral allele would reduce the number of mutations arising between IR1 paralogs as only mutations on IR1Yp would be observed (A) while conversion of a mutation arising on IR1Yq to the derived allele would increase the number of PSVs observed between IR1 paralogs (B).

carry the derived allele on the distal arm of the palindrome while only two (C10 and C12) carry the derived allele on the proximal arm. Of the six PSVs that have arisen due to mutation on the distal palindrome arm, more proximal-to-distal conversion events than distal-to-proximal events were identified in each case (Table 10.1). For this subset of PSVs significantly more proximal-to-distal conversion events were observed (P=0.0007; Chi square test) with a total of 32 proximal-to-distal conversion events being identified compared to only ten distal-to-proximal conversion events. Of the two PSVs which have arisen due to mutation on the proximal arm of the palindrome more distal-to-proximal conversion events were observed for each PSV with a total of 13 distal-to-proximal conversion compared to five proximal-to-distal conversion events. This difference was not statistically significant (P=0.0593; chi square test), however, as only two PSVs analyzed carried the derived allele on the proximal arm of the palindrome this could be due to an ascertainment bias. As it has not been possible to individually phase each PSV analyzed in this study the phase has been determined from the reference sequence and it has been assumed that no rearrangements have occurred. However it is possible that rearrangements such as inversions may have occurred within some chromosomes which has switched alleles between palindrome arms. This analysis does not appear to suggest that there is an arm-to-arm bias in the direction of gene conversion whereby one arm acts preferentially as a sequence donor. These data appear to support conservative gene conversion with the unmutated arm acting more frequently as the sequence donor during gene conversion. To seek further evidence of an arm bias, analysis of the P8 palindrome was also carried out.

Interspecies sequence comparison reveals that both P8 PSVs have arisen due to mutation on the proximal arm of the palindrome. As gene conversion appears to occur

PSV	Proximal arm allele	Distal arm allele	Ancestral state	Mutated arm	Proximal-to-distal conversion events	Distal-to-proximal conversion events
C6	G	A	G	Distal	**8**	**1**
C7	C	G	C	Distal	**7**	**4**
C10	A	C	C	Proximal	2	6
C11	C	T	C	Distal	**3**	**0**
C12	C	T	T	Proximal	3	7
C17	G	A	G	Distal	**7**	**4**
C24	A	T	A	Distal	**3**	**1**
C40	C	G	C	Distal	**4**	**0**

Table 10.1: Analysis of proximal-to-distal conversion events between P6 snPSVs
6 of the 8 PSVs which have been shown to undergo gene conversion have arisen due to mutation on the distal arm of the palindrome, while only two have arisen due to mutation on the proximal arm. Of the 6 PSVs which have arisen due to mutation on the distal arm significantly more proximal-to-distal gene conversion events were observed (P=0.0007, chi square test). The two PSVs which have arisen due to mutation on the proximal arm showed more distal-to-proximal conversion events.

rapidly between the arms of P8, under a proximal-to-distal direction of gene conversion, the derived allele for each PSV would be expected to be rapidly converted into the palindrome sequence. This analysis does not suggest that there is a proximal-to-distal bias in the direction of gene conversion between the arms of P8. It has not been possible to determine whether there is an arm-to-arm bias in the direction of gene conversion between IR1Yq and P1. Under a proximal-to-distal model of gene conversion it would be expected that IR1Yq would act as a sequence "donor" to P1 as it is located more proximal to the centromere of the chromosome. As evidence of gene conversion has been determined from sequence alignments it is difficult to recognise a proximal-to-distal bias of gene conversion. While conversions to the ancestral allele would be "invisible", it would not be possible to determine which arm has acted as sequence donor during conversions to the derived allele.

10.2.3, Analysis of BGCgc between Yq paralogs.

There is growing literature to suggest that in mammalian genomes gene conversion between duplicated genes favours an increase in GC content, with regions frequently undergoing gene conversion having a higher GC content than regions in which gene conversion is absent (Galtier et al. 2001). This section will describe an analysis of the GC content of Yq paralogs which have previously been shown to undergo gene conversion, in order to determine whether Biased Gene Conversion to G or C allele (BGCgc) could be an alternative explanation to conservative gene conversion. As palindrome arms undergo gene conversion, while the spacers do not, BGCgc would be expected to increase the GC content of palindrome arms in relation to the spacer. Several factors must be taken into consideration when comparing the GC content between different regions. Firstly, GC content is known to vary between genic and non-genic sequences, and since the arms of palindromes but not the spacers contain genes this could lead to a bias. There is also the issue of relative GC-content in Alus and LINEs which can also lead to a similar bias. For this reason all sequences analyzed in this section were repeat-masked and repetitive regions and genic sequences will not be included in the analysis of GC content. Analysis of the GC content of IR1 will also be carried out with the GC content of IR1Yp which does not undergo gene conversion being expected to be lower than that of IR1Yq, which has been shown to undergo gene conversion with P1.

The GC content of P6 palindrome arms was calculated as 32.86% while the GC content of the spacer was lower at 32.25%. While the lower GC content of the spacer in relation to the palindrome arms would be expected under BGCgc this difference is not significant (P=0.11, 2-tailed Fisher exact test) and does not suggest that BGCgc occurs between the arms of P6. Of the 10 snPSVs typed in this study, 9 have an ancestral allele which is either a G or a C so BGCgc and conservative gene conversion cannot easily be distinguished from each other. One PSV, C12 T/C which arises at the root of haplogroup C of the Y phylogeny, has an ancestral T-allele, with the "mutant" C-allele arising on the proximal arm of the palindrome. Phylogenetic typing of this PSV identified 13 conversion events, 12 of which returned the PSV back to the ancestral T-allele while only one converted to the derived C-allele (Figure 10.2). This analysis does not provide evidence of BGCgc between the arms of P6. However, as only one PSV has an ancestral state which is not a G or a C it is difficult to determine the full significance of this observation. As gene conversion between the arms of P8 appears to be very rapid, under BGCgc the GC content of the palindrome arms would be expected to be

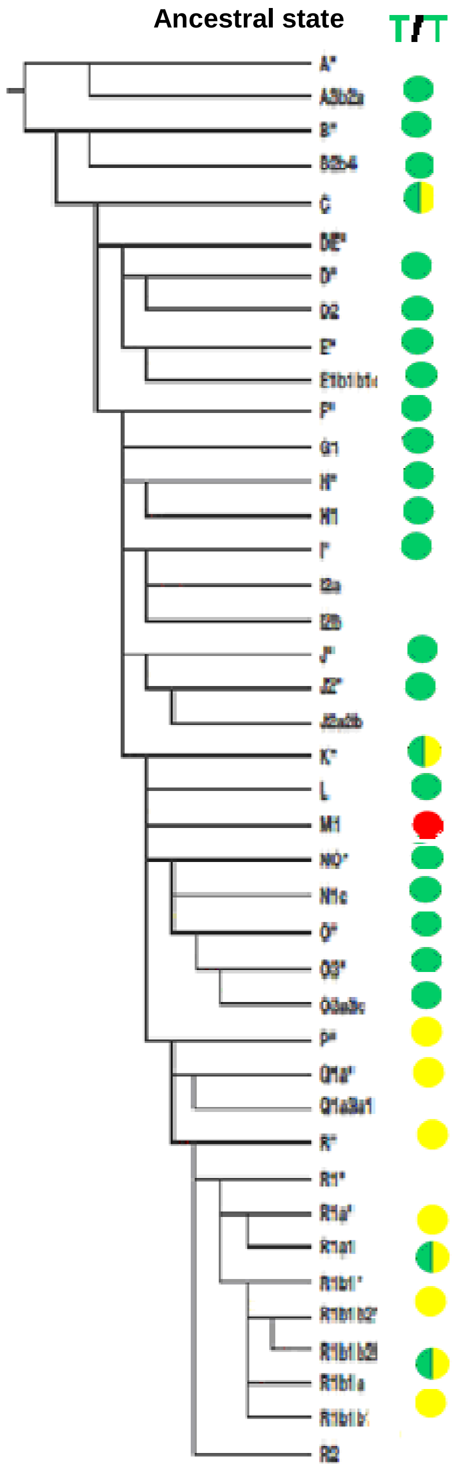


Figure 10.2: Evidence of BGCgc from analysis of the C12 PSV
Of the 10 snPSVs previously analyzed only the C12 PSVs was shown to arise due to a derived G or C allele. Only one conversion of the derived C allele into the palindrome sequence was observed while 7 conversions to the ancestral T allele were observed.

much higher than that of the spacer. As P8 contains the *VCY* genes the gene sequences will be removed and compared separately to the rest of the palindrome sequence. The CG content for the palindrome arms was calculated as 37.7% compared to 37.2% for the spacer, despite the GC content of the palindrome arms being slightly higher than that of the spacer this difference is not statistically significant (P=0.76, 2-tailed Fisher exact test) and does not provide evidence of BGCgc between the arms of P8.

Analysis of the *VCY* reference sequences was carried out separately to P8 and the ancestral state was inferred from comparisons with the chimpanzee and *VCX* sequences. From analysis of human and chimpanzee alignments 16 possible gene conversion events were identified were identified between the VCY genes. Of these 9 introduce a C or G allele into the sequence, while five events were shown to remove a G or C allele. Two possible conversion events were identified which occur at a CpG dinucleotide and as gene conversion cannot be distinguished from hypermutation these two sites were excluded from further analysis. From this analysis gene conversion does not appear to significantly increase the GC content of the VCY genes (P=0.16, chi-square test).

In the study of BGCgc between IR1 paralogs, the GC content of IR1Yp (which does not undergo gene conversion) would be expected to be lower than that of IRYq, which undergoes gene conversion with P1. The GC content of IR1Yq was calculated as 36.19%, while the GC content of IR1Yp was slightly lower at 35.76%. Although GC content of IR1Yp is lower than that of IR1Yq, as would be expected under BGCgc, this difference is not statistically significant (P=0.29, 2-tailed Fisher exact test). Analysis of mutations occurring on IR1Yp in the region of unique Yp-Yq paralogy reveals that mutation alone removes more GC-alleles than it produces, with 45 GC-alleles being removed from the sequence while only 36 are introduced. This suggests that the lower GC-content of IR1Yp compared to IR1Yq is most likely to be the result of mutation, and not due to BGCgc between IR1Yq and P1 paralogs.

From this analysis no significant evidence of BGCgc was observed for any of the regions analysed in this study and data obtained in this study suggests that gene conversion occurring between Y chromosome paralogs is conservative of the ancestral state.

10.2.4, Analysis of directional gene conversion between X-Y homologous genes

Several studies have suggested that gene conversion between X-Y gametologous genes is also directional. Establishing direction of gene conversion between X and Y genes is more complicated than it is for Y chromosome paralogs, as the X chromosome is three times more prevalent in the population than the Y chromosome, and hence Y-to-X conversion events may either be over represented on the X chromosome or be lost during meiosis.

Most recently Trombetta et al. (2009) have claimed that the *VCY* genes act as a sequence acceptor from *VCX* during gene conversion and while calculations of interspecies sequences divergence appears to support this, only one potential *VCX*-to-*VCY* conversion event was identified in this study; however, it was not possible to definitively determine the direction of gene conversion. (Rosser et al. 2009) have also recently provided evidence of gene conversion occurring between the *PRKX/Y* genes located within the X-degenerate regions of the Y chromosome. In this study more conversions from the Y to the X chromosome were identified. However, as this could be attributed to the larger population size of the X chromosome. Data obtained in this study has also provided evidence of gene conversion between the *TGIF2lX/Y* genes located within the XTR of the Y chromosome. As only one conversion event was identified a direction of gene conversion could not be established. While it has been established that gene conversion occurs between X-Y gametologous genes it has not been possible to determine the direction of gene conversion.

10.3, Discussion

Several studies have revealed that gene conversion appears to be directional and that the direction of gene conversion differs between species. On the human Y chromosome a direction of gene conversion has only been observed between two HERV sequences and directional gene conversion between other paralogs has only been hypothesized.

Analysis carried out in this chapter suggests that gene conversion between the paralogs of P6, P8, and IR1 is likely to be conservative of the ancestral sequence. No evidence of a proximal-to-distal bias in the direction of gene conversion has been observed between any paralogs analyzed in this study. Additionally no significant evidence to suggest BGCgc has been obtained for any of the regions analyzed in this study. Although the data obtained in this study appear to suggest that gene conversion between Y chromosome paralogs is most likely to be conservative of the ancestral sequence, this may not necessarily be the case for all Y chromosome paralogs. While it seems plausible that gene conversion between genic regions of the Y chromosome is likely to be conservative of the ancestral sequence in order to protect spermatogenic genes from mutation, the direction of gene conversion may be different for non-genic regions where incorporation of a derived allele into the sequence will not have effects on fertility. In the case of gene conversion between Y chromosome genes, if conversion to the derived allele causes male infertility the conversion event will not be observed leading to over representation of conversion events to the ancestral allele.

Chapter 11: Final discussion

Over the past seven years several studies have provided evidence of gene conversion occurring between different classes of Y chromosome paralog (Bosch and Jobling 2003; Rozen et al. 2003; Rosser, Balaresque, and Jobling 2009; Cruciani et al. 2010) some of which have suggested biases in the direction of gene conversion (Bosch and Jobling 2003; Rozen et al. 2003). This study provides evidence that gene conversion occurs between multiple Yq paralogs and is not limited to a particular class of paralog or to genic regions of the Y chromosome. The most striking finding of this study is that gene conversion occurring between Yq paralogs appears to be conservative of the ancestral state, as shown by the significantly lower divergence (P=0.001, 2-tailed Fisher exact test) between palindrome arms compared to the non-duplicated spacer regions. The main evidence to support conservative gene conversion has come from the phylogenetic analysis of snPSVs identified between the arms of P6, with significantly more conversions to the ancestral allele being observed (P=0.0001, Chi square test).

While these data appear to support the interspecies divergence calculations and provide significant evidence of conservative gene conversion, only a small subset of PSVs which were identified from a single reference sequence have been typed, which has created several biases in this study. As only one reliable reference sequence was available when this study was commenced all PSVs analyzed have been identified from a single reference sequence. This creates an ascertainment bias as PSVs which have undergone gene conversion in the reference sequence will not have been identified, leading to under-representation of gene conversion events. Also as only 10 of the 28 PSVS identified from the reference sequence were typed this may not give an accurate reflection of what is happening along the entire length of the palindrome arms - the frequency of gene conversion events may potentially differ between different regions of the palindrome. Resequencing of palindrome arms in diverse Y chromosome haplogroups would overcome this bias; however, due to the large size and repetitive nature of the palindromes, resequencing of P6 was not attempted in this study. Recent advances in DNA sequencing technology have made it possible to obtain data on a much larger scale than is possible with the traditional Sanger sequencing method used here. Next-generation sequencing technologies such as the 454 FLX (Roche), Solexa GA (Illumina), SOLiD (Applied Biosystems) and Polonator G.007 (Dover systems) are now widely used (Gupta 2008). Next-generation sequencing technologies produce read lengths of between 30-400bp, and while they offer three to four orders of magnitude more sequence than Sanger sequencing and are capable of generating >1 Gb of sequence data in a single run, the short read lengths mean that large regions are covered by multiple overlapping reads which do not enable the phase to be determined in diploid or pseudodiploid regions. Thus, these advances in sequencing technology make it possible to sequence the entire palindrome arms in diverse Y chromosomes, and to identify additional PSVs and gene conversion events to be identified; however, they do not overcome the issue of phase. Publication of the findings of the 1000 Genomes Project is also expected to overcome some of the biases encountered in this study. This project aims to sequence the genomes of over 1000 anonymous individuals from diverse populations using next-generation sequencing technologies. Provided the sequence coverage is sufficient, publication of these data is expected to identify additional PSVs from diverse Y chromosome haplogroups which can be analyzed to seek evidence of gene conversion; however, the phase will be unknown, and the haploid nature of the Y may mean that there are problems with sequence coverage. Given the high inherent error rates of the technologies employed, this may mean that there are problems with data reliability.

The introduction of third-generation sequencing technologies which are based on single-molecule sequencing promises the potential to obtain much longer read lengths. In 2008 Helicos tSMS (Helicos Biosciences) became the first third-generation sequencing platform to be commercially available; however, this platform currently only offers 30bp read lengths. More single-molecule sequencing technologies such as PacBio (Pacific Biosystems), Nanopore (University of California), and ZS genetics TM (ZS genetics) are currently under development but are not yet commercially available. These sequencing technologies aim to produce >100,000bp of sequence, which offers the potential to sequence entire palindrome arms and will also allow the phase of PSVs to be determined.

One of the main aims of this study has been to determine whether there is a bias in the direction of gene conversion occurring between Y chromosome paralogs. While data obtained in this study suggest that gene conversion is conservative of the ancestral state there are other possible biases in the direction of gene conversion which may potentially occur. In this study we carried out analysis as to whether there is an arm-to-arm bias in the direction of gene conversion with one arm preferentially acting as sequence donor during gene conversion. While this has not suggested that there is an arm-to-arm bias it has been assumed that the phase of each PSV is the same as observed in the reference sequence alignment. It is possible that rearrangements (in particular, inversions) may have occurred which have lead to the switching of alleles between palindrome arms and an arm-to-arm bias may not be observed as the phase in each chromosome analyzed is unknown.

The issue of phase is more of a problem when typing duplicated microsatellites, as each microsatellite is expected to mutate independently it is not known which repeat number is associated with each palindrome arm. While the phase of each PSV has not been determined in this study there are several techniques which could allow this to be done in principle, including the physical separation of homologous sequences through allele-specific long-PCR, DEASH (Jeffreys and May 2003) or pulsed-field gel electrophoresis, as well as direct phasing through the introduction of third-generation sequencing technologies.

The use of bioinformatic software such as GENCOV (Sawyer 1989) is making it possible to determine if gene conversion is occurring *in silico*, and with increasing numbers of reference sequences being published it may be possible to determine if gene conversion is occurring without the need for experimental analysis. The GENCOV software has not been used in this study as there has been shown to be a higher rate of false positives when only two sequences are used in the analysis and when conversion tract lengths are <200bp (McGrath, Casola, and Hahn 2009). As the current analysis has been based on the alignment of the reference sequence, and resequencing of the palindrome arm was not carried out, the use of GENCOV is likely to introduce false positives into this study. The use of computer simulations is becoming increasing popular in modelling the effect that gene conversion would be expected to have on the human genome. Marais et al. (2010) have recently modelled the effect of gene conversion on the Y chromosome over a given number of generations. From these simulations it was shown that gene conversion opposes degeneration and has an advantageous effect on the Y chromosome. A similar approach could be used when looking for evidence of directional gene conversion between Y chromosome paralogs. Simulations could be performed where biases in the direction of gene conversion are introduced and the expected influence on sequence divergence observed. These could then be compared to the actual sequence divergence to determine whether there is a bias in the direction of gene conversion.

This study has provided an overview of gene conversion occurring between Y chromosome paralogs and has provided evidence of gene conversion occurring between various classes of Y chromosome paralog. More work will need to be carried out to establish the rate and directionality of gene conversion, and comparative analyses with other species should help us to understand the selective forces at work in patterning variation on sex chromosomes.

Supplementary information

HGDP number	Population	Geographic origin	2003 Haplogroup	2008 Haplogroup
HGDP01406	Bantu N.E.	Kenya	A(xA3b2a)	A(xA3b2a)
HGDP00987*	San	Namibia	A(xA3b2a)	A(xA3b2a)
HGDP00991	San	Namibia	B(xB2b4)	B(xB2b4)
HGDP00453	Biaka Pygmy	Central African Republic	B(xB2b4)	B(xB2b4)
HGDP00992	San	Namibia	B2b4	B2b4
HGDP00452*	Biaka Pygmies	Central African Republic	B2b4	B2b4
HGDP01204	Oroqen	China	C	C
HGDP00545	Papuan	New Guinea	C	C
HGDP00747	Japanese	Japan	D(xD2)	D(xD2)
HGDP01183	Yizu	China	D(xD2)	D(xD2)
HGDP00752	Japanese	Japan	D2	D2
HGDP01253	Mozabite	Algeria (Mozabite)	E(xE3b3)	E(xE1b1b1c)
HGDP00538	French	France	E(xE3b3)	E(xE1b1b1c)
HGDP00923*	Yoruba	Nigeria	E(xE3b3)	E(xE1b1b1c)
HGDP01412	Bantu N.E.	Kenya	E(xE3b3)	E(xE1b1b1c)
HGDP00628	Bedouin	Israel (Negev)	E3b3	E1b1b1c
HGDP00528	French	France	F*	F*
HGDP01317	Lahu	China	F*	F*
HGDP00893*	Russian	Russia	G	G
HGDP00017	Brahui	Pakistan	G	G
HGDP00254	Pathan	Pakistan	H(xH1)	H(xH1)
HGDP00041	Brahui	Pakistan	H1	H1
HGDP00887	Russian	Russia	I	I
HGDP00808	Orcadian	Orkney Islands	I	I
HGDP00015	Brahui	Pakistan	J(xJ2)	J(xJ2)
HGDP01076	Sardinian	Italy	J(xJ2)	J(xJ2)
HGDP00007	Brahui	Pakistan	J2(xJ2f2)	J2(xJ2a2b)
HGDP00525*	French	France	J2(xJ2f2)	J2(xJ2a2b)
HGDP01245	Xibo	China	K(xL,M,N,O,P)	K(xL,M,N,O,P)
HGDP00540	Papuan	New Guinea	K(xL,M,N,O,P)	K(xL,M,N,O,P)
HGDP01298	Uygur	China	L	L
HGDP00003	Brahui	Pakistan	L	L
HGDP00548	Papuan	New Guinea	M	M1

HGDP00490	NAN Melanesian	Bougainville	M	M1
HGDP00715	Cambodian	Cambodia	N(xN3)	N(xN1c)
HGDP01101	Tujia	China	N(xN3)	N(xN1c)
HGDP01249	Xibo	China	N3	N1c
HGDP00879	Russian	Russia	N3	N1c
HGDP00711	Cambodian	Cambodia	O(xO3)	O (xO3)
HGDP00748	Japanese	Japan	O(xO3)	O (xO3)
HGDP00019	Brahui	Pakistan	O3(xO3e)	O3(O3a3c)
HGDP00819*	Han	China	O3(xO3e)	O3(O3a3c)
HGDP00749	Japanese	Japan	O3e	O3a3c
HGDP00412	Burusho	Pakistan	O3e	O3a3c
HGDP01343	Naxi	China	P*	P*
HGDP00100*	Hazara	Pakistan	P*	P*
HGDP00144	Makrani	Pakistan	Q(xQ3a)	Q1a(xQ1a3a1)
HGDP00834	Surui	Brazil	Q(xQ3a)	Q1a(xQ1a3a1)
HGDP00364	Burusho	Pakistan	R*	R*
HGDP00023	Brahui	Pakistan	R*	R*
HGDP00011	Brahui	Pakistan	R1a1	R1a1
HGDP00807	Orcadian	Orkney Islands	R1a1	R1a1
HGDP01261	Mozabite	Algeria (Mozabite)	R1b*	R1b1
HGDP01067	Sardinian	Italy	R1b*	R1b1
HGDP00102	Hazara	Pakistan	R1b2	R1b1b1
HGDP00595*	Druze	Israel (Carmel)	R1b3(xR1b3e)	R1b1b2
HGDP00511	French	France	R1b3(xR1b3e)	R1b1b2
HGDP00141	Makrani	Pakistan	R2	R2
HGDP00005	Brahui	Pakistan	R2	R2

Table S2.1: CEPH-HGDP DNAs used for phylogenetic analysis

For phylogenetic analysis 64 DNA samples representing 31 different Y chromosome haplogroups and 17 populations were selected from the CEPH–HGDP panel (Cann et al. 2002). For re-sequencing a subset of 8 chromosomes (marked with asterisks), each representing a different haplogroup and population, were selected.

Information on the 2003 haplogroup (Jobling and Tyler-Smith 2003) for each DNA sample, determined from binary marker and microsatellite analysis, were kindly provided prior to publication by Peter de Knijff (subsequently published by Shi et al. (2009). The 2008 haplogroup for each sample was determined from comparison of the allelic states of binary markers, to new haplogroup resolution described by Karafet et al. 2008.

	a	b	c	d
Oven temperature (°C)	60	60	60	60
First readout time (ms)	300	N/A	N/A	N/A
Second readout time (ms)	300	N/A	N/A	N/A
Pre run voltage (kVolts)	15	15	15	15
Pre run time (Sec)	180	180	180	180
Injection voltage (kVolts)	1.5	1.2	1.6	1.2
Injection time (Sec)	25	18	15	23
Voltage number of steps (nk)	40	30	30	20
Voltage step interval (Sec)	15	15	15	15
Voltage Tolerance (kVolts)	0.6	N/A	N/A	N/A
Data delay time (Sec)	405	120	250	60
Run voltage (kVolts)	8.5	8.5	13.4	15
Run time (Sec)	5640	1499	2800	3500
Ramp delay (sec)	600	N/A	N/A	N/A
Current stability (uA)	30	5	5	5

Supplementary table S2.2: ABI parameters
Data obtained in this study was collected from an ABI 3730xl or 3130xl prism genetic analyser where stated. The parameters used were varied depending on the assay used and the machine on which products were run. N/A applies to settings which are used on the ABI 370xl which are not applicable to the 3130xl.

a) Parameters for the ABI 3730xl used for sequencing reactions performed with big dye v3.1

b) Parameters for the ABI 3130xl used for sequencing reactions performed with big dye v1.1

c) Parameters for the ABI 3130xl used for SNaPshot analysis

d) Parameters for the ABI 3130xl used for microsatellite analysis

STS markers compared in chapter 3
sY1247 – 2,609,521-2,809,806
sY14 – 2,615,096-2,815,586
sY109 – 5,384,172-5,584,646
sY52 - 5,651,980-5,852,463
sY1200 - 11,146,901-11,347,212
sY746 - 12,842,302-13,042,517
sY1066 – 13,643,509-13,843,639
sY1310 – 16,404,553-16,604,990
sY1311 - 16,417,192-16,617,630
sY1227 – 18,472,421-18,672,738

Genes compared in chapter 3
AMELY- 6,796,078-6,800,661
PCDH11- 4,928,267-5,028,748
VCY - 14,607,046-14,607,786

1Kb segments compared in chapter 3
3,000,000-3,001,000
5,000,000-5,001,000
7,000,000-7,001,000
9,000,000-9,001,000
11,000,000-11,001,000
13,000 000-13,001,000
15,000,000-15,001,000
17,000,000-17,001,000
19,000,000-19,001,000
21,000,000-21,001,000

Supplementary information 3.1
Regions of the Y chromosom analysed in chapter three

Sequence co-ordinates from UCSC genome browser March 2006

Palindrome 6
Human arm 1- chrY:16780827-16890820
Human arm2 - chrY:16937051-17047071

Chimp arm 1 - chrY:21,657,338-21,771,309
Chimp arm 2 - chrY:21,817,481-21,931,415

Palindrome 7
Human arm 1 - chrY:16496133-16504854
Human arm 2 - chrY:16,517,494-16,526,218

Chimp arm1 – chrY:21,120,012-21,128,203
Chimp arm2 – chrY:21,140,849-21,149,040

Palindrome 8
Human arm 1 - chrY:14602927-14640931
Human arm 2 - chrY:14644347-14681749

Chimp arm 1 - chrY:19,178,282-19,206,927
Chimp arm 2 - chrY:19,256,805-19,274,165

TGIF2LY chrY:3,507,126-3,508,082
TG1F2LX chrX:89,063,741-89,064,466

VCX - chrX:7,770,303-7,772,184
VCYA - chrY:14,607,046-14,607,786
VCYB - chrY:14,677,492-14,678,232

IR1
Yp - chrY:7506530-7599980
Yq - chrY:23214221-23309811

Chimp Yp chrY:317,528-390,990
Chimp Yq chrY:2,803,557-2,899,233

Supplementary information 3.2
Sequence co-ordinates used to obtaine reference sequence data from the UCC genome browser

Palindrome 6			
Variant name	**Type**	**variant**	**Distance from outer boundary**
C1	SnPSV	C/T	290bp
C2	SnPSV	C/t	350bp
C3	INDEL	GAT/- - -	3.76-kb
C4	Poly T variation	T/ -	6.7-KB
C5	SnPSV	C/T	8.8-KB
C6	SnPSV	A/G	9.8-KB
C7	SnPSV	G/C	9.9-KB
C8	SnPSV	A/T	11.1-KB
C9	SnPSV	G/ -	12.5-KB
C10	SnPSV	A/C	13.4-KB
C11	SnPSV	T/C	13.7-KB
C12	SnPSV	T/C	14.8-KB
C13	SnPSV	T/C	15.3-KB
C14	SnPSV	T/C	17.3-KB
C15	SnPSV	A/G	18.6-KB
C16	SnPSV	T/C	18.7-KB
C17	SnPSV	A/G	18.8.-KB
C18	Poly A variation	AAAA/ - - - -	19.5-KB
C19	SnPSV	A/C	19.6-KB
C20	SnPSV	A/C	19.7-KB
C21	SnPSV	C/T	20.3-KB
C22	SnPSV	A/G	21.3-KB
C23	Microsatellite	CA(17/18)	25.4-KB
C24	SnPSV	T/A	26.2-KB
C25	Microsatellite	A/T(17/19)	26.3-KB
C26	POLY T variation	T/ -	26.6-KB
C27	SnPSV	G/ -	32.9-KB
C28	SnPSV	G/-	35.9-KB
C29	SnPSV	C/T	46.2-KB
C30	Poly A variation	A/-	57.5-KB
C31	SnPSV	A/G	59-KB
C32	SnPSV	C/T	62.4-KB
C33	Microsatellite	TAA(3)TAAAA(3/5)	65-KB
C34	Microsatellite	GT12/13	66-KB
C35	Microsatellite	CA(4)TA(1)CA(11/12)	69.2-KB
C36	SnPSV	A/T	69.3-KB
C37	Poly T variation	T/-	72.3-KB
C38	Poly A variation	AA/ - -	72.4-KB
C39	Microsatellite	GT(20/22)	75.2-KB
C40	SnPSV	C/G	81.1-KB
C41	SnPSV	A/G	82.9-KB
C42	SnPSV	A/G	82.9-KB

C43	SnPSV	A/G	82.9-KB
C44	Complex Microsetellite	GAAA BASED	83-KB
C45	Complex Microsetellite	GAAA BASED	83.1-KB
C46	Poly A variation	A/ -	90.2-KB
C47	Microsatellite	GT(20/22)	94.3-KB
C48	Microsatellite	GAAA(10/11)	94.9-KB
C49	SnPSV	C/T	95.1-KB
C50	SnPSV	C/ -	103.1-KB
Palindrome 7			
C1	SNPSV	T/G	615bp
C2	Microsatellite	ATG(3/4)	2.2-kb
C3	Microsatellite	AT(14/15)	4.7-KB
C4	SNPSV	C/A	8.4-kb
C5	SNPSV	A/C	8.7-kb
Palindrome 8			
C1	SNPSV	C/-	7kb
C2	SNPSV	C/G	26-KB

Supplementary table 4.1
PSVs identified from alignment of the reference sequence for palindromes 6-8

Sample	Population	Haplogroup
H1080	Bhutan	N*
H1121	Bhutan	N*
H1127	Bhutan	N*
H1142	Bhutan	N*
H1154	Bhutan	N*
H1800	Bhutan	N*
H1974	Bhutan	N*
H1985	Bhutan	N*
H1586	Bhutan	N*
H1604	Bhutan	N*
H1630	Bhutan	N*
H1360	Bhutan	N*
H1373	Bhutan	N*
H1392	Bhutan	N*
H1406	Bhutan	N*
H1421	Bhutan	N*
H1482	Bhutan	N*
H1599	Bhutan	N*
H1920	Bhutan	N*
H1368	Bhutan	N*

Supplementary table S6.1.
Haplogroup N* chromosomes used in the analysis of the C39 microsatellite

Brahui	Pakistan	R1a1
Brahui	Pakistan	R1a1
Brahui	Pakistan	R1a1
Brahui	Pakistan	R1a1
Balochi	Pakistan	R1a1
Balochi	Pakistan	R1a1
Balochi	Pakistan	R1a1
Balochi	Pakistan	R1a1
Balochi	Pakistan	R1a1
Balochi	Pakistan	R1a1
Makrani	Pakistan	R1a1
Makrani	Pakistan	R1a1
Sindhi	Pakistan	R1a1
Sindhi	Pakistan	R1a1
Sindhi	Pakistan	R1a1
Sindhi	Pakistan	R1a1
Sindhi	Pakistan	R1a1
Sindhi	Pakistan	R1a1
Sindhi	Pakistan	R1a1
Sindhi	Pakistan	R1a1
Pathan	Pakistan	R1a1
Pathan	Pakistan	R1a1
Pathan	Pakistan	R1a1
Pathan	Pakistan	R1a1
Pathan	Pakistan	R1a1
Pathan	Pakistan	R1a1
Pathan	Pakistan	R1a1
Pathan	Pakistan	R1a1
Kalash	Pakistan	R1a1
Kalash	Pakistan	R1a1
Kalash	Pakistan	R1a1
Burusho	Pakistan	R1a1
Burusho	Pakistan	R1a1
Bedouin	Israel (Negev)	R1a1
Bedouin	Israel (Negev)	R1a1
Russian	Russia	R1a1
Russian	Russia	R1a1
Russian	Russia	R1a1
Russian	Russia	R1a1
Uygur	China	R1a1
Uygur	China	R1a1
Tu	China	R1a1
Adygei	Russia Caucasus	R1a1
Orcadian	Orkney Islands	R1a1

Supplementary table S8.1.
Haplogroup R1a1 chromosomes used in the analysis of the M2 microsatellite

HGDP00473	Biaka Pygmies	Central African Republic
HGDP00475	Biaka Pygmies	Central African Republic
HGDP00477	Biaka Pygmies	Central African Republic
HGDP00448	Biaka Pygmies	Central African Republic
HGDP00461	Biaka Pygmies	Central African Republic
HGDP00464	Biaka Pygmies	Central African Republic
HGDP00465	Biaka Pygmies	Central African Republic
HGDP00466	Biaka Pygmies	Central African Republic
HGDP00469	Biaka Pygmies	Central African Republic
HGDP00470	Biaka Pygmies	Central African Republic
HGDP00459	Biaka Pygmies	Central African Republic
HGDP00460	Biaka Pygmies	Central African Republic
HGDP00479	Biaka Pygmies	Central African Republic
HGDP00479	Biaka Pygmies	Central African Republic
HGDP00452	Biaka Pygmies	Central African Republic
HGDP00453	Biaka Pygmies	Central African Republic
HGDP00454	Biaka Pygmies	Central African Republic
HGDP00455	Biaka Pygmies	Central African Republic
HGDP00457	Biaka Pygmies	Central African Republic
HGDP00458	Biaka Pygmies	Central African Republic
HGDP01090	Biaka Pygmies	Central African Republic
HGDP01092	Biaka Pygmies	Central African Republic
HGDP00981	Biaka Pygmies	Central African Republic
HGDP00985	Biaka Pygmies	Central African Republic
HGDP00986	Biaka Pygmies	Central African Republic
HGDP01084	Biaka Pygmies	Central African Republic
HGDP01085	Biaka Pygmies	Central African Republic
HGDP01086	Biaka Pygmies	Central African Republic
HGDP01087	Biaka Pygmies	Central African Republic
HGDP01088	Biaka Pygmies	Central African Republic
HGDP01089	Biaka Pygmies	Central African Republic
HGDP01094	Biaka Pygmies	Central African Republic
HGDP00908	Mandenka	Senegal
HGDP00904	Mandenka	Senegal
HGDP00905	Mandenka	Senegal
HGDP00906	Mandenka	Senegal
HGDP00919	Mandenka	Senegal
HGDP01199	Mandenka	Senegal
HGDP01200	Mandenka	Senegal
HGDP00910	Mandenka	Senegal
HGDP00911	Mandenka	Senegal
HGDP00912	Mandenka	Senegal
HGDP00913	Mandenka	Senegal
HGDP00914	Mandenka	Senegal
HGDP00917	Mandenka	Senegal
HGDP01202	Mandenka	Senegal
HGDP01283	Mandenka	Senegal
HGDP01285	Mandenka	Senegal
HGDP01286	Mandenka	Senegal
HGDP00474	Mbuti Pygmies	Democratic Republic of Congo
HGDP00449	Mbuti Pygmies	Democratic Republic of Congo
HGDP00450	Mbuti Pygmies	Democratic Republic of Congo

HGDP00462	Mbuti Pygmies	Democratic Republic of Congo
HGDP00463	Mbuti Pygmies	Democratic Republic of Congo
HGDP00468	Mbuti Pygmies	Democratic Republic of Congo
HGDP00907	Mbuti Pygmies	Democratic Republic of Congo
HGDP00478	Mbuti Pygmies	Democratic Republic of Congo
HGDP00982	Mbuti Pygmies	Democratic Republic of Congo
HGDP00983	Mbuti Pygmies	Democratic Republic of Congo
HGDP00984	Mbuti Pygmies	Democratic Republic of Congo
HGDP01081	Mbuti Pygmies	Democratic Republic of Congo
HGDP00924	Yoruba	Nigeria
HGDP00925	Yoruba	Nigeria
HGDP00920	Yoruba	Nigeria
HGDP00921	Yoruba	Nigeria
HGDP00922	Yoruba	Nigeria
HGDP00924	Yoruba	Nigeria
HGDP00925	Yoruba	Nigeria
HGDP00926	Yoruba	Nigeria
HGDP00927	Yoruba	Nigeria
HGDP00929	Yoruba	Nigeria
HGDP00931	Yoruba	Nigeria
HGDP00932	Yoruba	Nigeria
HGDP00933	Yoruba	Nigeria
HGDP00934	Yoruba	Nigeria
HGDP01384	Adygei	Russia Caucasus
HGDP01386	Adygei	Russia Caucasus
HGDP01382	Adygei	Russia Caucasus
HGDP01370	French Basque	France
HGDP01362	French Basque	France
HGDP01363	French Basque	France
HGDP01365	French Basque	France
HGDP01367	French Basque	France
HGDP01372	French Basque	France
HGDP01373	French Basque	France
HGDP01374	French Basque	France
HGDP01375	French Basque	France
HGDP01376	French Basque	France
HGDP01377	French Basque	France
HGDP00802	Orcadian	Orkney Islands
HGDP00803	Orcadian	Orkney Islands
HGDP00804	Orcadian	Orkney Islands
HGDP00795	Orcadian	Orkney Islands
HGDP00797	Orcadian	Orkney Islands
HGDP00798	Orcadian	Orkney Islands
HGDP00799	Orcadian	Orkney Islands
HGDP00800	Orcadian	Orkney Islands
HGDP00805	Orcadian	Orkney Islands
HGDP00807	Orcadian	Orkney Islands
HGDP00666	Sardinian	Italy
HGDP00667	Sardinian	Italy
HGDP00669	Sardinian	Italy
HGDP00670	Sardinian	Italy
HGDP00671	Sardinian	Italy

HGDP01065	Sardinian	Italy
HGDP01075	Sardinian	Italy
HGDP01066	Sardinian	Italy
HGDP01067	Sardinian	Italy
HGDP01068	Sardinian	Italy
HGDP01069	Sardinian	Italy
HGDP01071	Sardinian	Italy
HGDP01072	Sardinian	Italy
HGDP01077	Sardinian	Italy
HGDP01079	Sardinian	Italy
HGDP01071	Sardinian	Italy
HGDP01166	Tuscan	Italy
HGDP01167	Tuscan	Italy
HGDP01161	Tuscan	Italy
HGDP01162	Tuscan	Italy
HGDP01169	Tuscan	Italy
HGDP01155	North Italian	Italy (Bergamo)
HGDP01156	North Italian	Italy (Bergamo)
HGDP01147	North Italian	Italy (Bergamo)
HGDP01149	North Italian	Italy (Bergamo)
HGDP01153	North Italian	Italy (Bergamo)
HGDP01171	North Italian	Italy (Bergamo)
HGDP01173	North Italian	Italy (Bergamo)
HGDP01310	Dai	China
HGDP01308	Dai	China
HGDP01217	Daur	China
HGDP01213	Daur	China
HGDP01214	Daur	China
HGDP01215	Daur	China
HGDP01216	Daur	China
HGDP01241	Hezhen	China
HGDP01242	Hezhen	China
HGDP01234	Hezhen	China
HGDP01235	Hezhen	China
HGDP01236	Hezhen	China
HGDP01237	Hezhen	China
HGDP01238	Hezhen	China
HGDP01239	Hezhen	China
HGDP01240	Hezhen	China
HGDP01196	Miaozu	China
HGDP01197	Miaozu	China
HGDP01230	Mongola	China
HGDP01231	Mongola	China
HGDP01223	Mongola	China
HGDP01224	Mongola	China
HGDP01225	Mongola	China
HGDP01226	Mongola	China
HGDP01206	Oroqen	China
HGDP01207	Oroqen	China
HGDP01242	Hezhen	China
HGDP01234	Hezhen	China
HGDP01235	Hezhen	China

HGDP01236	Hezhen	China
HGDP01237	Hezhen	China
HGDP01238	Hezhen	China
HGDP01239	Hezhen	China
HGDP01240	Hezhen	China
HGDP01196	Miaozu	China
HGDP01197	Miaozu	China
HGDP01230	Mongola	China
HGDP01231	Mongola	China
HGDP01223	Mongola	China
HGDP01224	Mongola	China
HGDP01225	Mongola	China
HGDP01226	Mongola	China
HGDP01206	Oroqen	China
HGDP01207	Oroqen	China
HGDP01203	Oroqen	China
HGDP01098	Tujia	China
HGDP01099	Tujia	China
HGDP01101	Tujia	China
HGDP01102	Tujia	China
HGDP01103	Tujia	China
HGDP01104	Tujia	China
HGDP01302	Uygur	China
HGDP01303	Uygur	China
HGDP01305	Uygur	China
HGDP01306	Uygur	China
HGDP01246	Xibo	China
HGDP01247	Xibo	China
HGDP01248	Xibo	China
HGDP01250	Xibo	China
HGDP01180	Yizu	China
HGDP01181	Yizu	China
HGDP01183	Yizu	China
HGDP01184	Yizu	China
HGDP01185	Yizu	China
HGDP01186	Yizu	China
HGDP01188	Yizu	China
HGDP01189	Yizu	China
HGDP01312	Dai	China
HGDP01313	Dai	China
HGDP01315	Dai	China
HGDP01316	Dai	China
HGDP01317	Dai	China
HGDP01318	Dai	China
HGDP01323	Dai	China
HGDP01324	Dai	China
HGDP01325	Dai	China
HGDP01326	Dai	China
HGDP01322	Lahu	China
HGDP01344	Naxi	China
HGDP01345	Naxi	China
HGDP01337	Naxi	China

HGDP01339	Naxi	China
HGDP01340	Naxi	China
HGDP01341	Naxi	China
HGDP01346	Naxi	China
HGDP01332	She	China
HGDP01333	She	China
HGDP01328	She	China
HGDP01335	She	China
HGDP01336	She	China
HGDP01356	Tu	China
HGDP01347	Tu	China
HGDP01348	Tu	China
HGDP01349	Tu	China
HGDP01351	Tu	China
HGDP01352	Tu	China
HGDP01353	Tu	China
HGDP00952	Yakut	Siberia
HGDP00967	Yakut	Siberia
HGDP00968	Yakut	Siberia
HGDP00964	Yakut	Siberia
HGDP00969	Yakut	Siberia
HGDP00757	Japanese	Japan
HGDP00747	Japanese	Japan
HGDP00748	Japanese	Japan
HGDP00749	Japanese	Japan
HGDP00750	Japanese	Japan
HGDP00752	Japanese	Japan
HGDP00753	Japanese	Japan
HGDP00758	Japanese	Japan
HGDP00768	Japanese	Japan
HGDP00759	Japanese	Japan
HGDP00762	Japanese	Japan
HGDP00763	Japanese	Japan

Supplementary table S9.1.
Chromosomes used in the analysis of the M3 microsatellite

References

Affara, N. A., M. A. Ferguson-Smith, J. Tolmie, K. Kwok, M. Mitchell, D. Jamieson, A. Cooke, and L. Florentin. 1986. Variable transfer of Y-specific sequences in XX males. Nucl. Acids Res. **14**:5375-5387.

Amos, W., J. Flint, and X. Xu. 2008. Heterozygosity increases microsatellite mutation rate, linking it to demographic history. BMC Genet. **14;9:72.**

Bachrati, C. Z., R. H. Borts, and I. D. Hickson. 2006. Mobile D-loops are a preferred substrate for the Bloom's syndrome helicase. Nucleic Acids Res **34**:2269–2279

Bagnall, R., K. Ayres, P. Green, and F. Giannelli. 2005. Gene conversion and evolution of Xq28 duplicons involved in recurring inversions causing severe hemophilia A. Genome Research **15**.

Bahlo, M., and R. Griffiths. 2000. The largest study of Y-SNP variation free from ascertainment bias, which is based on DHPLC and proposes a common ancestor, Inference from gene trees in a subdivided population. . . Popul. Biol Theor **57**.

Bailey, J. A., A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler. 2001. Segmental duplications: Organization and impact within the current Human Genome Project assembly. Genome Research **11**:1005-1017.

Balaresque, P., A. Sibert, E. Heyer, and B. Crouau-Roy. 2007. Unbiased interpretation of haplotypes at duplicated microsatellites. Ann Hum Genet **71**:209-219.

Ballard, D. J., C. Phillips, G. Wright, C. R. Thacker, C. Robson, A. P. Revoir, and D. S. Court. 2005. A study of mutation rates and the characterisation of intermediate, null and duplicated alleles for 13 Y chromosome STRs. Forensic Sci. Int. **155**:65-70.

Bandelt, H.-J., P. Forster, and A. Röhl. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. **16**:37-48.

Bandelt, H.-J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. Genetics **141**:743-753.

Batzer, M., D. Ray, J. Xing, P. Callinan, J. Myers, D. Hedges, R. Garber, D. Witherspoon, and L. Jorde. 2003. Alu elements and hominid phylogenetics. PNAS **100**.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucl. Acids Res. **27**:573-580.

Bernstein, R., A. Wadee, J. Rosendorff, A. Wessels, and T. Jenkins. 1986. Inverted Y chromosome polymorphism in the Gujerati Muslim Indian population of South Africa. Hum. Genet. **74**:223-229.

Bhowmick, B. K., Y. Satta, and N. Takahata. 2007. The origin and evolution of human ampliconic gene families and ampliconic structure. Genome Res **17**:441-450.

Blanchard. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res **19**:1527-1541.

Blanco, P., M. Shlumukova, C. A. Sargent, M. A. Jobling, N. Affara, and M. E. Hurles. 2000. Divergent outcomes of intra-chromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. J. Med. Genet. **37**:752-758.

Bosch, E., M. E. Hurles, A. Navarro, and M. A. Jobling. 2004. Dynamics of a human interparalog gene conversion hotspot. Genome Res. **14**:835-844.

Bosch, E., and M. A. Jobling. 2003. Duplications of the *AZFa* region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. Hum. Mol. Genet. **12**:341-347.

Brayton, K., G. Palmer, J. Futse, C. Leverich, D. Knowles, and F. Rurangirwa. 2007. Selection for Simple Major Surface Protein 2 Variants during Anaplasma marginale Transmission to Immunologically Naïve Animals. Infection and Immunity **75**:1502-1506.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Hühne, and B. Rolf. 1998a. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. **62**:1408-1415.

Brinkmann, B., M. Klintschar, F. Neuhuber, and B. Rolf. 1998b. Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. Am J Hum Genet. **62**:1408-1415.

Brudno, M., C. Do, G. Cooper, M. Kim, E. Davydov, E. Green, A. Sidow, and S. Batzoglou. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Research **13**:721-731.

Brunet, M., F. Guy, D. Pilbeam, H. T. Mackaye, A. Likius, D. Ahounta, A. Beauvilain, C. Blondel, H. Bocherens, and J. R. Boisserie. 2002. A new hominid from the Upper Miocene of Chad, Central Africa. Nature **418**.

Burk, R., and K. Smith. 1985. Characterisation and evolution of a single copy sequence from the human Y chromosome. Mol. Cell. Biol. **5**:576-581.

Cann, H. M., C. de Toma, L. Cazes, M. F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Y. Chu, C. Carcassi, L. Contu, R. F. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Y. Huang, J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. P. Qian, Q. F. Shu, J. J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset, and L. L. Cavalli-Sforza. 2002. A human genome diversity cell line panel. Science **296**:261-262.

Carvalho-Silva, D. R., F. R. Santos, M. H. Hutz, F. M. Salzano, and S. D. J. Pena. 1999. Divergent human Y-chromosome microsatellite evolution rates. J. Mol. Evol. **49**:204-214.

Casanova, M., P. Leroy, C. Boucekkine, J. Weissenbach, C. Bishop, M. Fellous, M. Purrello, G. Fiori, and M. Siniscalco. 1985. A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science **230**:1403-1406.

Causio, F., D. Canale, L. M. Schonauer, R. Fischetto, T. Leonetti, and N. Archidiacono. 2000. Breakpoint of a Y chromosome pericentric inversion in the DAZ gene area. A case report. J Reprod Med **45**:591-594.

Chen, F. C., and W. H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68**:444-456.

Chen, J. M., D. N. Cooper, N. Chuzhanova, C. Ferec, and G. P. Patrinos. 2007. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet **8**:762-775.

Cruciani, F., B. Trombetta, V. Macaulay, and R. Scozzari. 2010. About the X-to-Y gene conversion rate. Am. J. Hum. Genet. **in press**.

Dahlberg, J., J. Hu, B. Cote, and E. Lund. 1983. Isolation and characterization of genomic mouse DNA clones containing sequences homologous to tRNAs and 5S rRNA. Nucleic Acids Res **25**:4809-4821.

de Knijff, P., T. Kraaijenbrink, E. J. Parkin, D. R. Carvalho-Silva, G. L. van Driem, G. Barbujani, C. Tyler-Smith, and M. A. Jobling. 2009. Genetic and linguistic borders in the Himalayan Region. Becoming Eloquent: :181-201.

Dean, F. B., S. Hosono, L. H. Fang, X. H. Wu, A. F. Faruqi, P. Bray-Ward, Z. Y. Sun, Q. L. Zong, Y. F. Du, J. Du, M. Driscoll, W. M. Song, S. F. Kingsmore, M. Egholm, and R. S. Lasken. 2002. Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl. Acad. Sci. USA **99**:5261-5266.

Devriendt, K., S. Frintsa, J. Frynsa, L. Lagaec, M. Syrroua, and P. Marynenb. 2001. Xp22.3; Yq11.2 chromosome translocation and its clinical manifestations. Annales de Génétique **44**:71-76.

Ebersberger, I., D. Metzler, C. Schwarz, and S. Pääbo. 2002. Genomewide Comparison of DNA Sequences between Humans and Chimpanzees . AJHG **70**:1490-1497.

Ehrlich, M., and R. Wang. 1981. 7.5-Methylcytosine in eukaryotic DNA. Science **212**.

Emrie, P., C. Jones, T. Hofmann, and J. Fisher. 1988. The coding sequence for the human 18,000-dalton hydrophobic pulmonary surfactant protein is located on chromosome 2 and identifies a restriction fragment length polymorphism. . Somat Cell Mol Genet **14**:105-110.

Ferguson-Smith, M., A. O'Reilly, N. Affara, E. Simpson, P. Chandler, and E. Goulmy. 1992. A molecular deletion map of the Y chromosome long arm defining X and autosomal homologous regions and the localisation of the HYA locus to the proximal region of the Yq euchromatin. Hum Mol Genet **1**:379-385.

Galtier, N., and L. Duret. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet **23**:273-277.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. GC-Content Evolution in Mammalian Genomes:The Biased Gene Conversion Hypothesis. Genetics **159**:907-911.

Graves, J. A., E. Koina, and N. Sankovic. 2006. How the gene content of human sex chromosomes evolved. Curr. Opin. Genet. Dev. **16**:219-224.

Gupta, P. 2008. Single-molecule DNA sequencing for future genomics research. Trends in Biotechnology **26**.

Hall, T. 2005. BioEdit v7.0.5

Hamer, D. H., and L. Li. 1995. Recombination and allelic association in the Xq/Yq homology region. Hum. Mol. Genet. **4**.

Hammer, M. F. 1995. A recent common ancestry for human Y chromosomes. Nature **378**:376-378.

Hammer, M. F., and S. L. Zegura. 2002. The human Y chromosome haplogroup tree: nomenclature and phylogeny of its major divisions. Annu. Rev. Anthropol. **31**:303-321.

Harpending, H., L. Jorde, A. Rogers, M. Bamshad, S. Watkins, P. Krakowiak, S. Sung, and J. Kere§. 1997. Microsatellite diversity and the demographic history of modern humans. PNAS **94**.

Hattori, T., T. Ogata, K. Muroya, G. Sasaki, G. Nishimura, and H. Kitoh. 2002. Development of Turner Skeletal Features SHOX Nullizygosity and Haploinsufficiency in a Japanese Family: Implication for the Development of Turner Skeletal Features. J. Clin. Endocrinol. Meta **87**:1390-1394.

Haussler, D., W. Kent, C. Sugnet , T. Furey, K. Roskin , T. Pringle, and A. Zahler. 2002. The human genome browser at UCSC. Genome Res **12**:996-1006.

Higgins, T. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res **31**.

Howell, R., S. Roberts, and R. Beard. 1976. Dicentric X isochromosomes in man. J Med Genet. **13**:496-500.

Hughes, J., H. Skaletsky, T. Pyntikova, T. Graves, S. van Daalen, P. Minx, R. Fulton, S. McGrath, D. Locke, C. Friedman, B. Trask, E. Mardis, W. Warren, S. Repping, Rozen,S. Wilson, R. and Page, D. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature **Vol 463|28**

Hughes, J. F., H. Skaletsky, T. Pyntikova, P. J. Minx, T. Graves, S. Rozen, R. K. Wilson, and D. C. Page. 2005. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. Nature **437**:100-103.

Hughes, J. F., H. Skaletsky, S. Rozen, R. K. Wilson, and D. C. Page. 2006. Has the chimpanzee Y chromosome been sequenced? Nat Genet **38**:853-854; author reply 854-855.

Hurles, M. 2004. Gene duplication: the genomic trade in spare parts. PLoS Biol **2**:E206.

Hurles, M. E., and M. A. Jobling. 2001. Haploid chromosomes in molecular ecology: lessons from the human Y. Mol. Ecol. **10**:1599-1613.

Hurles, M. E., and M. A. Jobling. 2003. A singular chromosome. Nat Genet **34**:246-247.

Huson, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14**:68-73.

Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23**:254-267.

Ira, G., A. Malkova, G. Liberi, M. Foiani, and J. E. Haber. 2003. Srs2 and Sgs1–Top3 suppress crossovers during double-strand break repair in yeast. Cell **115**:401-411.

Istrail, S., G. Sutton, L. Florea, A. Halpern, C. Mobarry, R. Lippert, B. Walenz, H. Shatkay, I. Dew, J. Miller, M. Flanigan, N. Edwards, R. Bolano, D. Fasulo, H. BV., i. S. Hannenhall, R. Turner, S. Yooseph, F. Lu, D. Nusskern, B. Shue, X. Zheng, F. Zhong, A. Delcher, D. Huson, S. Kravitz, L. Mouchard, K. Reinert, K. Remington, A. Clark, M. Waterman, E. Eichler, M. Adams, r. M. Hunkapille, E. Myers, and J. Venter. 2004. Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci **101**:1916-1921.

Jacobs, P. A., A. Sharp, K. Kusz, J. Jaruzelska, M. Szarras-Czapnik, and J. Wolski4. 2004. Familial X/Y translocations associated with variable sexual phenotype. J Med Genet. **41**:440-444.

Jeffreys, A. J., and C. A. May. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nature Genet. **36**:152-156.

Jeffreys, A. J., and C. A. May. 2003. DNA enrichment by allele-specific hybridization (DEASH): a novel method for haplotyping and for detecting low-frequency base substitutional variants and recombinant DNA molecules. Genome Res **13**:2316-2324.

Jeffreys, A. J., R. Neumann, and V. Wilson. 1990. Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell **60**:473-485.

Jobling, M. A. 2008. Copy number variation on the human Y chromosome. Cytogenet. Genome Res. **123**:253-262.

Jobling, M. A., A. Pandya, and C. Tyler-Smith. 1997. The Y chromosome in forensic analysis and paternity testing. Int. J. Legal Med. **110**:118-124.

Jobling, M. A., and C. Tyler-Smith. 2003. The human Y chromosome: an evolutionary marker comes of age. Nat. Rev. Genet. **4**:598-612.

Kamp, C., P. Hirschmann, H. Voss, K. Huellen, and P. H. Vogt. 2000. Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. Hum. Mol. Genet. **9**:2563-2572.

Karafet, T., P. de Knijff, E. Wood, J. Ragland, A. Clark, and M. F. Hammer. 1998. Different patterns of variation at the X- and Y-chromosome-linked microsatellite loci DXYS156X and DXYS156Y in human populations. Hum. Biol. **70**:979-992.

Karafet, T. M., F. L. Mendez, M. Meilerman, P. A. Underhill, S. L. Zegura, and M. F. Hammer. 2008. New binary polymorphisms reshape and increase resolution of the human Y-chromosomal haplogroup tree. Genome Res. **18**:830-838.

Kayser, M., R. Kittler, A. Erler, M. Hedman, A. C. Lee, A. Mohyuddin, S. Q. Mehdi, Z. Rosser, M. Stoneking, M. A. Jobling, A. Sajantila, and C. Tyler-Smith. 2004. A comprehensive survey of human Y-chromosomal microsatellites. Am. J. Hum. Genet. **74**:1183-1197.

Kehrer-Sawatzki, H., and D. Cooper. 2007. Understanding the recent evolution of the human genome: insights from human–chimpanzee genome comparisons. Human Mutation **28**:99-130.

Lahn, B., N. Pearson, and K. Jegalian. 2001. The human Y chromosome, in the light of evolution. Nature Reviews Genetics **2**:207-216.

Lahn, B. T., and D. C. Page. 1999. Four evolutionary strata on the human X chromosome. Science **286**:964-967.

Lander, E. S.L. M. LintonB. BirrenC. NusbaumM. C. ZodyJ. BaldwinK. DevonK. DewarM. DoyleW. FitzHughR. FunkeD. GageK. HarrisA. HeafordJ. HowlandL. KannJ. LehoczkyR. LeVineP. McEwanK. McKernanJ. MeldrimJ. P. MesirovC. MirandaW. MorrisJ. NaylorC. RaymondM. RosettiR. SantosA. SheridanC. SougnezN. Stange-ThomannN. StojanovicA. SubramanianD. WymanJ. RogersJ. SulstonR. AinscoughS. BeckD. BentleyJ. BurtonC. CleeN. CarterA. CoulsonR. DeadmanP. DeloukasA. DunhamI. DunhamR. DurbinL. FrenchD. GrafhamS. GregoryT. HubbardS. HumphrayA. HuntM. JonesC. LloydA. McMurrayL. MatthewsS. MercerS. MilneJ. C. MullikinA. MungallR. PlumbM. RossR. ShownkeenS. SimsR. H. WaterstonR. K. WilsonL. W. HillierJ. D. McPhersonM. A. MarraE. R. MardisL. A. FultonA. T. ChinwallaK. H. PepinW. R. GishS. L. ChissoeM. C. WendlK. D. DelehauntyT. L. MinerA. DelehauntyJ. B. KramerL. L. CookR. S. FultonD. L. JohnsonP. J. MinxS. W. CliftonT. HawkinsE. BranscombP. PredkiP. RichardsonS. WenningT. SlezakN. DoggettJ. F. ChengA. OlsenS. LucasC. ElkinE. UberbacherM. FrazierR. A. GibbsD. M. MuznyS. E. SchererJ. B. BouckE. J. SodergrenK. C. WorleyC. M. RivesJ. H. GorrellM. L. MetzkerS. L. NaylorR. S. KucherlapatiD. L. NelsonG. M. WeinstockY. SakakiA. FujiyamaM. HattoriT. YadaA. ToyodaT. ItohC. KawagoeH. WatanabeY. TotokiT. TaylorJ. WeissenbachR. HeiligW. SaurinF. ArtiguenaveP. BrottierT. BrulsE. PelletierC. RobertP. WinckerA. RosenthalM. PlatzerG. NyakaturaS. TaudienA. RumpH. M. YangJ. YuJ. WangG. Y. HuangJ. GuL. HoodL. RowenA. MadanS. Z. QinR. W. DavisN. A. FederspielA. P. AbolaM. J. ProctorR. M. MyersJ. SchmutzM. DicksonJ. GrimwoodD. R. CoxM. V. OlsonR. KaulN. ShimizuK. KawasakiS. MinoshimaG. A. EvansM. AthanasiouR. SchultzB. A. RoeF. ChenH. Q. PanJ. RamserH. LehrachR. ReinhardtW. R. McCombieM. de la BastideN. DedhiaH. BlockerK.

HornischerG. NordsiekR. AgarwalaL. AravindJ. A. BaileyA. BatemanS. BatzoglouE. BirneyP. BorkD. G. BrownC. B. BurgeL. CeruttiH. C. ChenD. ChurchM. ClampR. R. CopleyT. DoerksS. R. EddyE. E. EichlerT. S. FureyJ. GalaganJ. G. R. GilbertC. HarmonY. HayashizakiD. HausslerH. HermjakobK. HokampW. H. JangL. S. JohnsonT. A. JonesS. KasifA. KaspryzkS. KennedyW. J. KentP. KittsE. V. KooninI. KorfD. KulpD. LancetT. M. LoweA. McLysaghtT. MikkelsenJ. V. MoranN. MulderV. J. PollaraC. P. PontingG. SchulerJ. R. SchultzG. SlaterA. F. A. SmitE. StupkaJ. SzustakowkiD. Thierry-MiegJ. Thierry-MiegL. WagnerJ. WallisR. WheelerA. WilliamsY. I. WolfK. H. WolfeS. P. YangR. F. YehF. CollinsM. S. GuyerJ. PetersonA. FelsenfeldK. A. WetterstrandA. Patrinos, and M. J. Morgan. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. Clustal W and Clustal X version 2.0. Bioinformatics **23**:2947-2948.

Levy, S., G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. MacDonald, A. W. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. A. Kravitz, D. A. Busam, K. Y. Beeson, T. C. McIntosh, K. A. Remington, J. F. Abril, J. Gill, J. Borman, Y. H. Rogers, M. E. Frazier, S. W. Scherer, R. L. Strausberg, and J. C. Venter. 2007. The diploid genome sequence of an individual human. PLoS Biol. **5**:e254.

Liu, Y., and S. West. 2004. Happy Hollidays: 40th anniversary of the Holliday junction. Nat Rev Mol Cell Biol. **5**:937-944.

Marcus, M., R. Tantravahi, V. Dev, Miller DA, and M. OJ. 1976. Human-mouse cell hybrid with human multiple Y chromosomes. Nature **262**:63-65.

McGrath, C., C. Casola, and M. Hahn. 2009. Minimal Effect of Ectopic Gene Conversion Among Recent Duplicates in Four Mammalian Genomes. Genetics **182**:615-622.

Morris, B. J., and A. H. Mangs. 2007. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. current genomics:129–136. .

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science **310**:321-324.

Myers, S., C. Freeman, A. Auton, P. Donnelly, and G. McVean. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. **40**:1124-1129.

Nachmana, M., and S. Crowella. 2000. Estimate of the Mutation Rate per Nucleotide in Humans Genetics, **156**.

Navarro-Costa, P., C. Plancha, and J. Gonçalves. 2010. Genetic Dissection of the AZF Regions of the Human Y Chromosome: Thriller or Filler for Male (In)fertility? Journal of Biomedicine and Biotechnology **2010**.

Nielsen, K., J. Kasper, M. Choi, T. Bedford, K. Kristiansen, D. Wirth, S. Volkman, J. Szostak, T. Orr-Weaver, R. Rothstein, and F. Stahl. 1983. The double-strand-break repair model for recombination. Cell **33**.

Olson, M. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet. **64**:18-23.

Page, D. 1986. Sex reversal: deletion mapping the male-determining function of the human Y chromosome. . Cold Spring Harb Symp Quant Biol **51**.

Page, S., and L. Shaffer. 1998. Chromosome stability is maintained by short intercentromeric distance in functionally dicentric human Robertsonian translocations. Chromosome Research **6**:115-122.

Palmer, G., and K. Brayton. 2007. Gene conversion is a convergent strategy for pathogen antigenic variation. TRENDS in Parasitology **23**.

Parkin, E. J., T. Kraayenbrink, G. L. van Driem, K. Tshering, P. de Knijff, and M. A. Jobling. 2006. 26-locus Y-STR typing in a Bhutanese population sample. Forensic Sci. Int. **161**:1-7.

Povey, S., Jeremiah SJ, Barker RF, Hopkinson DA, Robson EB, Cook PJL, Solomon E, Bobrow M, Carritt B, and B. KE. 1980. Assignment of the human locus determining phosphoglycolate phophatase (PGP) to chromosome 16. . Ann. Hum. Genet. **43**:241-248.

Qamar, R., Q. Ayub1, A. Mohyuddin, A. Helgason, K. Mazhar, A. Mansoor, Zerjal.T., C. Tyler-Smith, and Q. Mehdi. 2002. Y-Chromosomal DNA Variation in Pakistan. The American Journal of Human Genetics **70**:1107-1124.

Quintana-Murci, L., and M. Fellous. 2001. The human Y chromosome: the biological role of a“functional wasteland”. Journal of Biomedicine and Biotechnology **1**.

Rappold, G. A. 1993. The pseudoautosomal regions of the human sex chromosomes. Hum. Genet **92:315–324**.

Repping, S., H. Skaletsky, L. Brown, S. K. van Daalen, C. M. Korver, T. Pyntikova, T. Kuroda-Kawaguchi, J. W. de Vries, R. D. Oates, S. Silber, F. van der Veen, D. C. Page, and S. Rozen. 2003. Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. Nat. Genet. **35**:247-251.

Repping, S., H. Skaletsky, J. Lange, S. Silber, F. Van Der Veen, R. D. Oates, D. C. Page, and S. Rozen. 2002. Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. Am. J. Hum. Genet. **71**:906-922.

Repping, S., S. K. van Daalen, L. G. Brown, C. M. Korver, J. Lange, J. D. Marszalek, T. Pyntikova, F. van der Veen, H. Skaletsky, D. C. Page, and S. Rozen. 2006. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat. Genet. **38**:463-467.

Ross, M., D. Grafham, A. Coffey, S. Scherer, K. McLay, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, Frankish A, Lovell FL, Howe KL, Ashurst JL, Fulton RS, Sudbrak R, Wen G, Jones MC, Hurles ME, Andrews TD, Scott CE, Searle S, Ramser J, Whittaker A, Deadman R, Carter NP, Hunt SE, Chen R, Cree A, Gunaratne P, Havlak P, Hodgson A, Metzker ML, Richards S, Scott G, Steffen D, Sodergren E, Wheeler DA, Worley KC, Ainscough R, Ambrose KD, Ansari-Lari MA, Aradhya S, Ashwell RI, Babbage AK, Bagguley CL, Ballabio A, Banerjee R, Barker GE, Barlow KF, Barrett IP, Bates KN, Beare DM, Beasley H, Beasley O, Beck A, Bethel G, Blechschmidt K, Brady N, Bray-Allen S, Bridgeman AM, Brown AJ, Brown MJ, Bonnin D, Bruford EA, Buhay C, Burch P, Burford D, Burgess J, Burrill W, Burton J, Bye JM, Carder C, Carrel L, Chako J, Chapman JC, Chavez D, Chen E, Chen G, Chen Y, Chen Z, Chinault C, Ciccodicola A, Clark SY, Clarke G, Clee CM, Clegg S, Clerc-Blankenburg K, Clifford K, Cobley V, Cole CG, Conquer JS, Corby N, Connor RE, David R, Davies J, Davis C, Davis J, Delgado O, and Deshazo D. 2005. The DNA sequence of the human X chromosome. Nature **434**.

Ross, M. T., D. R. Bentley, and C. Tyler-Smith. 2006. The sequences of the human sex chromosomes. Curr. Opin. Genet. Dev. **16**:213-218.

Rosser, Z. H., P. Balaresque, and M. A. Jobling. 2009. Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. Am. J. Hum. Genet. **85**:130-134.

Rozas, J., and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15**:174-175.

Rozen, S., H. Skaletsky, J. D. Marszalek, P. J. Minx, H. S. Cordum, R. H. Waterston, R. K. Wilson, and D. C. Page. 2003. Abundant gene conversion between arms of massive palindromes in human and ape Y chromosomes. Nature **423**:873-876.

Samonte, R. V., and E. E. Eichler. 2002. Segmental duplications and the evolution of the primate genome. Nature Reviews Genetics **3**:65-72.

Sarto, G., E. Therman, C. Trunca, and E. Kuhn, . 1986. Dicentric chromosomes and the inactivation of the centromere. . Hum Genet. **72**:191-195.

Sawyer, S. 1989. Statistical Tests for Detecting Gene Conversion 1. Mol Biol Evol **6**:526-538.

Schlecht, J., M. E. Kaplan, K. Barnard, T. Karafet, M. F. Hammer, and N. C. Merchant. 2008. Machine-learning approaches for classifying haplogroup from Y chromosome STR data. PLoS Comput. Biol. **4**:e1000093.

Schwartz, A., D. Chan, L. Brown, R. Alagappan, D. Pettay, C. Disteche, B. McGillivray, A. de la Chapelle, and D. Page. 1998. Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. Hum Mol Genet. **7**:1-11.

Schwartz, S., W. Kent, A. Smit, Z. Zhang, R. Baertsch, R. Hardison, D. Haussler, and W. Miller. 2003. Human-mouse alignments with BLASTZ. Genome Res. **13**.

Sekido, R., and R. Lovell-Badge. 2008. Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer. Nature **453**:930-934.

Sen, S. K. 2006. Human genomic deletions mediated by recombination between *Alu* elements. *Am. J. Hum Genet*. **79**.

Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, R. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S.-F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.-P. Yang, R. H. Waterston, R. K. Wilson, S. Rozen, and D. C. Page. 2003. The male-specific region of the human Y chromosome: a mosaic of discrete sequence classes. Nature **423**:825-837.

Stone, A. C., R. C. Griffiths, S. L. Zegura, and M. F. Hammer. 2002. High levels of Y-chromosome nucleotide diversity in the genus *Pan*. Proc. Natl. Acad. Sci. USA **99**:43-48.

Sun, C., H. Skaletsky, S. Rozen, J. Gromoll, E. Nieschlag, R. Oates, and D. C. Page. 2000. Deletion of *azoospermia factor a* (*AZFa*) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum. Mol. Genet. **9**:2291-2296.

Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner, and M. W. Feldman. 2000. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. Proc. Natl. Acad. Sci. USA **97**:7360-7365.

Tremblay, A., M. Jasin, and P. Chartrand. 2000. A doublestrand break in a chromosomal LINE element can be repaired by gene conversion with various endogenous LINE elements in mouse cells. Mol. Cell. Biol **20**.

Trombetta, B., F. Cruciani, P. A. Underhill, D. Sellitto, and R. Scozzari. 2009. Footprints of X-to-Y gene conversion in recent human evolution. Mol. Biol. Evol.

Tyler-Smith, C. 2008. An evolutionary perspective on Y-chromosomal variation and male infertility. Int. J. Androl. **31**:376-382.

Tyler-Smith, C., K. Howe, and F. R. Santos. 2006. The rise and fall of the ape Y chromosome? Nature Genetics **38**.

Underhill, P., L. Jin, A. Lin, S. Mehdi, T. Jenkins, D. Vollrath, R. Davis, L. Cavalli-Sforza, and P. Oefner. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res. **7**:996-1005.

Van Esch, H., K. Hollanders, L. Badisco, C. Melotte, P. Van Hummelen, R. Rober, V. Koen Devriendt, J. Fryns1, P. Marynen, and G. Froyen. 2005. Deletion of VCX-A due to NAHR plays a major role in the occurrence of mental retardation in patients with X-linked ichthyosis. Human Molecular Genetics, **14**.

Venter, J. C.M. D. AdamsE. W. MyersP. W. LiR. J. MuralG. G. SuttonH. O. SmithM. YandellC. A. EvansR. A. HoltJ. D. GocayneP. AmanatidesR. M. BallewD. H. HusonJ. R. WortmanQ. ZhangC. D. KodiraX. Q. H. ZhengL. ChenM. SkupskiG. SubramanianP. D. ThomasJ. H. ZhangG. L. G. MiklosC. NelsonS. BroderA. G. ClarkC. NadeauV. A. McKusickN. ZinderA. J. LevineR. J. RobertsM. SimonC. SlaymanM. HunkapillerR. BolanosA. DelcherI. DewD. FasuloM. FlaniganL. FloreaA. HalpernS. HannenhalliS. KravitzS. LevyC. MobarryK. ReinertK. RemingtonJ. Abu-ThreidehE. BeasleyK. BiddickV. BonazziR. BrandonM. CargillI. ChandramouliswaranR. CharlabK. ChaturvediZ. M. DengV. Di FrancescoP. DunnK. EilbeckC. EvangelistaA. E. GabrielianW. GanW. M. GeF. C. GongZ. P. GuP. GuanT. J. HeimanM. E. HigginsR. R. JiZ. X. KeK. A. KetchumZ. W. LaiY. D. LeiZ. Y. LiJ. Y. LiY. LiangX. Y. LinF. LuG. V. MerkulovN. MilshinaH. M. MooreA. K. NaikV. A. NarayanB. NeelamD. NusskernD. B. RuschS. SalzbergW. ShaoB. X. ShueJ. T. SunZ. Y. WangA. H. WangX. WangJ. WangM. H. WeiR. WidesC. L. XiaoC. H. YanA. YaoJ. YeM. ZhanW. Q. ZhangH. Y. ZhangQ. ZhaoL. S. ZhengF. ZhongW. Y. ZhongS. P. C. ZhuS. Y. ZhaoD. GilbertS. BaumhueterG. SpierC. CarterA. CravchikT. WoodageF. AliH. J. AnA. AweD. BaldwinH. BadenM. BarnsteadI. BarrowK. BeesonD. BusamA. CarverA. CenterM. L. ChengL. CurryS. DanaherL. DavenportR. DesiletsS. DietzK. DodsonL. DoupS. FerrieraN. GargA. GluecksmannB. HartJ. HaynesC. HaynesC. HeinerS. HladunD. HostinJ. HouckT. HowlandC. IbegwamJ. JohnsonF. KalushL. KlineS. KoduruA. LoveF. MannD. MayS. McCawleyT. McIntoshI. McMullenM. MoyL. MoyB. MurphyK. NelsonC. PfannkochE. PrattsV. PuriH. QureshiM. ReardonR. RodriguezY. H. RogersD. RombladB. RuhfelR. ScottC. SitterM. SmallwoodE. StewartR. StrongE. SuhR. ThomasN. N. TintS. TseC. VechG. WangJ. WetterS. WilliamsM. WilliamsS. WindsorE. Winn-DeenK. WolfeJ. ZaveriK. ZaveriJ. F. AbrilR. GuigoM. J. CampbellK. V. SjolanderB. KarlakA. KejariwalH. Y. MiB. LazarevaT. HattonA. NarechaniaK. DiemerA. MuruganujanN. GuoS. SatoV. BafnaS. IstrailR. LippertR. SchwartzB. WalenzS. YoosephD. AllenA. BasuJ. BaxendaleL. BlickM. CaminhaJ. Carnes-StineP. CaulkY. H. ChiangM. CoyneC. DahlkeA. D. MaysM. DombroskiM. DonnellyD. ElyS. EsparhamC. FoslerH. GireS. GlanowskiK. GlasserA. GlodekM. GorokhovK. GrahamB. GropmanM. HarrisJ. HeilS. HendersonJ. HooverD. JenningsC. JordanJ. JordanJ. KashaL. KaganC. KraftA. LevitskyM. LewisX. J. LiuJ. LopezD. MaW. MajorosJ.

McDanielS. MurphyM. NewmanT. NguyenN. NguyenM. NodellS. PanJ. PeckM. PetersonW. RoweR. SandersJ. ScottM. SimpsonT. SmithA. SpragueT. StockwellR. TurnerE. VenterM. WangM. Y. WenD. WuM. WuA. XiaA. Zandieh, and X. H. Zhu. 2001. The sequence of the human genome. Science **291**:1304-1351.

Verma, R., J. Rodriguez, and H. Dosik. 1982. The clinical significance of pericentric inversion of the human Y chromosome: a rare "third" type of heteromorphism. . J Hered. **73(3)**.

Vincent, B. J. 2003. Following the LINEs: an analysis of primate genomic variation at human-specific LINE1 insertion sites. Mol. Biol. Evol. **20**.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature **452**:872-876.

Willard, H. 2003. Tales of the Y chromosome. Nature **19;423(6942)**.

Wu, L., and I. D. Hickson. 2003. The Bloom’s syndrome helicase suppresses crossing over during homologous recombination. . Nature **426**:870-874.

Xue, Y., Q. Wang, Q. Long, B. Ling, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, D. MacArthur, M. Quail, N. Carter, H. M. Yang, and C. Tyler-Smith1. 2009. Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. Current Biology **19**.

Yen, H., S. Tsai, S. Wenger, M. Steele, T. Mohandas, and L. Shapiro. 1991. X/Y translocations resulting from recombination between homologous sequences on Xp and Yq. Proc. Natl. Acad. Sci. USA **88**:8944-8948.

Zhang, J., J. Wang, W. Wang, R. Li, Y. Li, G. Tian, L. Goodman, W. Fan, and L. Jun. 2008. The diploid genome sequence of an Asian individual. Nature. **456**:60-65.

Zhi, D. 2007. Sequence correlation between neighboring Alu instances suggests post-retrotransposition sequence exchange due to Alu gene conversion. Gene **390**.