Gene conversions and selection in the gene families of primates

By: Nicholas Petronella

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of the University of Ottawa in partial fulfillment of the requirements for the M.Sc. degree in the Ottawa-Carlton Institute of Biology

Thèse soumise à la Faculté des études supérieurs et postdoctorales de l'Université d'Ottawa en vue de l'obention de la maîtrise en biologie de l'Institat de biologie d'Ottawa-Carlton

Examiners

The following members of the Ottawa-Carlton Institute of Biology

Dr. Guy Drouin
Supervisor

Dr. Stéphane Aris-Brosou
Examining Member (University of Ottawa)

Dr. Marc Ekker
Examining Member (University of Ottawa)

Dr. Ashkan Golshani
Examining Member (Carleton)

Acknowledgements

First and foremost I would like to thank my supervisor Guy Drouin for giving me the opportunity to carry out my Master's degree at the University of Ottawa. His help and guidance was always available and extremely appreciated. In my two years at the University of Ottawa I have learned a great deal and most of it is thanks to him.

I would also like to thank my committee members Stéphane Aris-Brosou and Michel Dumontier who provided me with much needed constructive feedback throughout the course of my research.

In addition, I would like to thank my examiners Marc Ekker and Ashkan Golshani for their comments and feedback on my thesis project.

Lastly, I would like to thank my parents, Jack and Genevieve Petronella for all their love and support through all my years schooling. They are responsible for the person I am today and the completion of this research project would not have been possible without them.

Thanks again to all,

Nicholas Petronella

Abstract

We used the GENECONV program, the Hsu et al. (2010) method and phylogenetic analyses to analyze the gene conversions which occurred in the growth hormone, folate receptor and trypsin gene families of six primate species. Significant positive correlations were found between sequence similarity and conversion length in all but the trypsin gene family. Converted regions, when compared to non-converted ones, also displayed a significantly higher GC-content in the growth hormone and folate receptor gene families. Finally, all detected gene conversions were found to be less frequent in conserved gene regions and towards functionally important genes. This suggests that purifying selection is eliminating all gene conversions having a negative functional impact.

Résumé

Nous avons utilisé le programme GENECONV, la méthode de Hsu et al. (2010) et des arbres phylogénétiques pour analyser les conversions géniques qui se sont produites dans les familles multigéniques codants pour les hormones de croissance, les récepteurs de folate et les trypsines chez six espèces de primates. Des corrélations positives significatives ont été trouvées entres la similarité des séquences et la longueur des conversions dans toutes les familles, sauf celle codant pour la trypsine. Chez les familles codants pour les hormones de croissance et les récepteurs de folate, les régions converties ont aussi un contenu en bases G et C significativement plus élevées que les régions non-converties. Finalement, nos résultats démontrent que les conversions géniques sont moins fréquentes dans les régions conservées des gènes et vers les gènes fonctionnellement importants. Ceci suggère que la sélection purificatrice élimine les conversions géniques ayant un impact fonctionnel négatif.

Table of Contents

Chapter 1.

Introduction

Gene conversion is a non-reciprocal type of homologous recombination in which there is a unidirectional exchange of sequence information between two homologous sequences. Gene conversions are the main form of homologous recombination in eukaryotes that are initiated by double strand breaks. Following a double strand break, gene conversion mediates the fashion in which the sequence exhibiting this break is repaired by transferring genetic information from an intact homologous sequence to the region that contains the double strand break (Szostak et al., 1983, Stahl, 1996, Keeney, 2001). Gene conversion events are capable of occurring between sequences on homologous, the same or different chromosomes. Conversions that occur between alleles at the same locus are known as allelic gene conversions whereas ectopic gene conversions refer to events that occur between two sequences found either on the same or different chromosomes (Petes and Hill, 1998). The first evidence for gene conversion came from a study on the meiotic recombination in yeast done by Lindegren in 1953 where allelic ratios after meiotic recombination were not as expected (Lindgreen, 1953). Since this study, gene conversions have been observed and studied in a multitude of organisms from bacteria to humans (Shen et al., 1986, Slightom et al., 1985, Wang et al., 1999, Chen et al, 2007).

As previously stated, gene conversions are known to occur after double strand breaks, which mostly take place during both meiosis and mitosis. During yeast meiosis, these breaks are created via the SPO11 meiotic recombination protein enzyme. In contrast, during mitosis, double strand breaks can manifest through radiation, stalled replication forks, or specialized endonucleases (enzyme capable of cleaving phosphodiester bonds) (Chen et al., 2007). Gene conversions are also associated with crossing over, another form of homologous recombination. It has been found in previous studies that crossover and gene

conversion events are well associated (Paques and Haber 1999). Moreover, conversions that occur during meiosis have a higher association with crossover events, compared to those that occur during mitosis, due to the formation and resolving of Holiday Junctions (HJs). This relationship between these two forms of recombination brought about the formation of two gene conversion models (Chen et al., 2007, Haber et al., 2004, Szostak et al., 1983, Paques and Haber, 1999). These two models are the double strand break repair (DSB repair) and the synthesis-dependent strand annealing (SDSA) models. The creation of the SDSA model was to account for the noticeable absence of crossing over during mitotic gene conversions while the DSB repair model is associated with meiotic gene conversion events.

The DSB repair and the SDSA models both begin in a similar fashion. Once a double stand break occurs, the ends are removed by 5' → 3' endonucleases resulting in two 3' single stranded DNA (ssDNA) tails (shown in Figure 1.1). The tails then search the genome for homologous sequences. Once a homolog is found, one of the aforementioned 3' tails invades the homologous sequence, forming a displacement loop (D-loop), which is subsequently extended through DNA synthesis (shown in grey in Figure 1.1). This D-loop then extends, allowing the other 3' single stranded DNA tail to pair. What follows is where the two models differ. In the SDSA model, the non-template strand is displaced and anneals to the other 3' ssDNA (shown in Figure 1.1). In the DSB repair model however, the remaining loose ends are joined by DNA synthesis resulting in the formation of two Holiday Junction (HJ) structures (shown in Figure 1.1). In the case of the SDSA model, once the newly synthesized strand is annealed, DNA synthesis follows and ligates all loose ends, resulting exclusively in a gene conversion event with no crossover products. In DSB repair, it is the resolving of the HJs that will produce either a crossover or non-crossover (gene conversion) product. HJs can be cleaved by a resolvase cutting either the two noncrossed

strands or the two crossed strands (shown in Figure 1.1). If both HJs are cut in the same fashion, only non-crossover products are formed. Conversely, if the HJs are not resolved in the same way, crossing over will occur. In addition, most gene conversion is a consequence of the mismatch repair that must occur in order to resolve mispaired nucleotides between the donor and acceptor DNA sequences (Chen et al., 2007, Haber et al., 2004, Szostak et al., 1983, Paques and Haber, 1999).

1.1. Characteristics of gene conversion

Previous studies have identified certain characteristics of gene conversion events. The most notable of these characteristics is that gene conversion generally requires high sequence similarity between interacting sequences. Studies involving a variety of organisms from yeast to mammals demonstrated that the frequency and size of conversion events depends on the degree of sequence similarity of involved genes (Shen and Huang, 1986, Wand et al., 1999, Ezawa et al., 2006, Drouin, 2002, Liskay et al., 1987, Waldman and Liskay, 1988, Elliott et al., 1998). For example, in a review by Chen at al., (2007) on the subject of gene conversions being mechanisms of human disease, all gene conversions that were found to be correlated with disease involved sequences exhibiting a sequence similarity between 88 and 95%. Additionally, Benovoy and Drouin (2009) observed that large conversions in the human genome involve sequences that are on average at least 89% similar. Moreover, this aforementioned requirement for gene conversions to occur explains why gene families are ideal candidates for conversion events. Gene families consist of all genes that belong to a certain group of repeated sequences, usually formed by duplication events, therefore consisting of highly homologous sequences (Brooker, 2005).

In addition to sequence similarity, the frequency and length of conversion events has been found to be inversely proportional to the physical distance between interacting

sequences (Schildkraut et al., 2005, Benovoy and Drouin, 2009). For example, members of a gene family that are located farther away will usually not convert as often as those who are in close proximity to each other.

1.2. Effects of gene conversion

Gene conversions are also known to affect the GC-content of converted sequences. The biased gene conversion hypothesis states that due to biases in DNA repair mechanisms, higher gene conversion frequencies will lead to higher GC-content (Meunier and Duret 2004, Galtier et al., 2001, Kudla et al., 2004). For example, it was found that the duplicated regions of the yeast and *Arabidopsis* genomes exhibited an increase in GC-content due to gene conversion between genes being more than 88% similar (Benovoy et al., 2005).

Gene conversion must not be neglected when studying ancestral relationships between sequences. The increase in similarity between genes must be taken into consideration when investigating the orthologous relationships of genes between species. For example, phylogenetic relationships demonstrated in a tree may be modified due to conversion events. The increase in similarity resulting from conversion events can lead to a loss of the orthologous relationships between sequences due to genes now being more similar to each than to their ancestral orthologs (Drouin et al. 1999). As a result, if a conversion is large enough, the conversion event can be observed in a phylogenetic tree.

Moreover, conversion events also have the capability of causing concerted evolution, a process in which an individual member of a gene family does not evolve independently relative to its other family members, further affecting the ancestral relationships between sequences (Carter et al., 2009). Consequently, it is theoretically possible to detect gene conversion by reference to neighbouring sequences that did not evolve in concert (Graur and Li, 2000). One example of such a case is in the concerted

evolution of $^{G}\gamma$- and $^{A}\gamma$-globin. These two genes were found to be created by a duplication event approximately 35 million years ago (Slightom et al., 1985). Due to the divergence of the African apes (humans, chimpanzee and gorilla) we would expect the $^{G}\gamma$ orthologous genes from apes to be more similar to each other rather than to $^{A}\gamma$ paralogs. This, however, was not found to be the case. Panel (a) in Figure 1.2 represents a phylogenetic tree for exon 3 of the $^{G}\gamma$- and $^{A}\gamma$-globin genes, which was found to be unconverted, preserving the expected orthologous relationships. However, frequent conversion events occurring in exons 1 and 2 of the $^{G}\gamma$- and $^{A}\gamma$-globin genes resulted in a loss of the expected orthologous relationships, causing the $^{G}\gamma$- and $^{A}\gamma$-globin genes to be more similar to each other rather than to their respective ancestral sequences (shown in Figure 1.2 panel (b) ; Slightom et al., 1985).

Due to gene conversion events having the capacity to overwrite important genetic regions within sequences, some disadvantageous conversions may be selected against. In a previous study on green and red visual pigment genes it was observed that introns 4 and 2 of these genes were identical, most likely due to a recent gene conversion event occurring during the evolution of these genes in the human lineage (Shyue et al., 1994). It was further hypothesized however, that if gene conversion events would further extend into the neighbouring exon regions, these conversions would have detrimental effects and would be selected against since they would lower the ability to distinguish between the colours red and green (Shyue et al., 1994).

The primary fashion in which gene conversions can cause undesirable effects is when mutations that have accumulated in pseudogenes are donated to highly similar functional genes, overwriting important regions (Bateman et al., 2007, Vanita et al., 2001, Heinen et al., 2006, Friaes et al., 2006, Bagnall et al., 2005, Chen et al., 2007). For example, it was found that a sequence change can occur when CRYBB2 (a gene producing lens related

proteins) is converted by CRYBP1, a highly similar pseudogene, resulting in cataracts (Bateman et al., 2007). Similarly, when the CYP21A2 gene (a gene coding for a steroid related enzyme) is converted by CYP21A1P, a highly similar pseudogene, a point mutation commonly found in CYP21A1P can be transferred into CYP21A2 causing adrenal hyperplasia (Friaes et al., 2006). Moreover, Chen et al., (2007) published a review surveying all shown connections between gene conversions and genetic diseases that have previously been discovered. In this review, it is stated that 44 gene conversion events have been identified, so far, to be connected with genetic disease.

In contrast, previous research has also shown that some gene conversion events may be selectively advantageous. In a previous study on bacterial genomes, it was observed that gene conversion is integral in the generation of antigenic variation allowing some bacterial pathogens to avoid the host immune system (Santoyo and Romero, 2005). Furthermore, in past studies involving chicken, duck and zebra finch, it was observed that these species are able to generate antibody diversity via gene conversion with a pool of pseudogenes, allowing them to produce an effective defence mechanism against pathogens (McCormack et al., 1993, Das et al., 2010, Lundqvist et al., 2006).

1.3. Methods for detecting gene conversions

The surge in freely available computational biology tools and software has brought about a variety of recombination detection programs, some of which specialize in gene conversion detection. Two such gene conversion detection programs were used in this research project.

Firstly, there is GENECONV, which searches for possible conversion events and outputs a probability by analyzing the monomorphic and polymorphic sites of the inputted sequences (Sawyer, 1989). In order for GENECONV to identify these aforementioned sites, at least three sequences must be inputted. Monomorphic sites (shown in black in Figure

1.3) are those that are fixed, possibly for biological reasons. These sites are identified and subsequently removed by GENECONV from the analysis. Polymorphic sites (shown in red in Figure 1.3) are areas in the sequences that can potentially vary. These sites are used in the analysis and are randomly permuted a minimum of 10,000 times. GENECONV then calculates a p-value based on how many times stretches of identical nucleotides (shown in blue in Figure 1.3) as long as those observed in the actual sequences are reproduced in the simulated sequences. This is computed through an SSCF (sum of squares of condensed fragments). The SSCF is calculated for each polymorphism permutation by first identifying fragments that are similar. Lengths of these similar fragments are then squared and summed giving a number that represents the SSCF for that particular polymorphism permutation (see Figure 1.3). GENECONV then outputs a p-value for each potential conversion event found, representing the probability of a conversion event being responsible for the observed fragment similarities. The p-value is the number of permuted data sets divided by 10,000 having an SSCF-value greater than or equal to that of the observed data. Only the global fragments with simulated P-values of less than 0.05 are considered significant since these are corrected for multiple comparisons via the Bonferroni correction. This form of multiple comparison compensation lowers the statistical significance interval for "n" number of tests to 0.05/n, thus avoiding spurious positives. Pairwise fragments, even with simulated p-values of less than 0.05, should rarely be considered since they are not corrected for multiple comparisons.

A second computational method that allows for the detection of gene conversion is the Hsu et al., 2010 pipeline. In past gene conversion research, the investigation of conversion events was vastly restricted to those occurring between pairs of pseudogenes or protein coding genes. However, in reality, conversions have the potential to occur between any pair of highly similar genetic regions (Chen et al., 2007). In addition, previous studies

were only capable of analyzing a few thousand pairs of genes, whereas the Hsu et al., 2010 pipeline allows for the potential to analyze more than one million paralogous pairs of regions.

In order to achieve such a sizable analysis, the Hsu et al., 2010 pipeline based itself and improved upon a previous computational program developed by Boni et al., 2007. This aforementioned program proved fruitful in detecting gene conversions and other recombination events, however, it operated on an algorithm that called for considerable amounts of computer memory. The Hsu et al., 2010 pipeline was successful in reducing the amount of computer memory required to carry out an analysis by implementing and modifying the data structures implemented to store data used and created by the algorithm.

The input required for the Hsu et al., 2010 pipeline consists of a Newick-formatted species tree of the sequences being analyzed, a sequence file for each species and separate files for each species that indicate the locations and ranges of all exons within the inputted sequence files (see Figure 1.4). It must be noted, however, that the currently released version of the Hsu et al., 2010 pipeline can only be used for detecting gene conversions that occur between sequences on the same chromosome.

The first step in the Hsu et al., 2010 pipeline is the generation of all self alignments (aligning each sequence to itself) and pair-wise inter-species alignments using the LASTZ alignment software (see Figure 1.4). This alignment data is then inputted into the CAGE software, which is used to identify all pair-wise orthologs in the alignment data and whose output consists of a subset of inter-species alignments. The last step of the pipeline is the conversion detector, which examines each pair of paralogous regions, together with their orthologs and performs several statistical tests to infer conversion events.

The general manner by which the Hsu et al., 2010 pipeline conversion detector operates is that it infers the ancestral relationships between inputted sequences, making

comparisons amongst them and examining if portions of sequences within species are more similar to each other than to their respective ancestral counterpart. The first step in the analyses is to identify all paralogous intervals within the sequence files of each species, M1 and M2. Secondly, an orthologous sequence, C1, is then identified for each paralogous interval from a species that is at an appropriate evolutionary distance (somewhat after the duplication and before the conversion). This ortholog is used to factor out the differences in the rates of evolutionary change along M1. For example, some regions of M1 could have evolved neutrally but at a rate that is influenced by neighbouring nucleotides. Contrastingly, others may be protein coding or regulatory regions under selection. These two paralogs (M1 and M2) and their corresponding ortholog (C1) are referred to as a "triplet". These triplets are then examined for instances where parts of the paralogous sequences are more similar to each other than to their corresponding ortholog. Instances of high M1-M2 similarity are suggested to have resulted from gene conversion. In addition to triplet testing, the Hsu Method extends to quadruplet testing where two orthologs are identified for a paralogous pair (i.e. C1 and C2 are orthologs for M1 and M2). Quadruplet testing often has increased specificity and sensitivity with respect to the former for conversion events. Nevertheless, both tests are used in conjunction when calculating the probabilities of potential conversion events to ensure accuracy.

Like GENECONV, the Hsu et al., 2010 pipeline applies several statistical tests with each iteration and in order to correct for multiple-comparisons, the Bonferroni correction is applied (Holm, 1979).

The final output of the Hsu et al., 2010 pipeline consists of the following: the identification of the species in which the conversion was found, the positions (corresponding to the sequence file) of the paralogs used in the testing, the strand (forward or reverse) on which the conversion occurred, the identity of the paralogs, the length of the

conversion event and its matching p-value, the direction of the conversion event (which gene was the donor or acceptor) and the positions where the conversion can be found (for both the donor and the acceptor).

The accuracy and efficiency for both GENECONV and the Hsu et al., 2010 pipeline have been tested in previous studies. GENECONV was found to be a useful tool to detect gene conversions except when the sequences analyzed are all very similar or when the sequences compared are more than 20% divergent (Drouin et al. 1999, Posada et al., 2001, Posada et al., 2002). If all sequences are very similar, this method exhibits type II error (false negative results) and will fail to detect gene conversions due to the lack of polymorphic sites to permute. Conversely, this method gives type I error (false positive results) when the sequences being compared are more than 20% divergent (Posada et al., 2001).

The Hsu et al., 2010 pipeline has not been as extensively studied as GENECONV due to its recent release, however, studies have compared its performance to other gene conversion detection programs and techniques. In a simulation study, the Hsu et al., 2010 pipeline exhibited the highest sensitivity of all other techniques by identifying the most true conversion events (Song et al., 2011). Nevertheless, the primary reasons why this pipeline was chosen over GENECONV in our later research projects is due to its ability to take into account the ancestral relationships of sequences when detecting gene conversions and its performance not being dependant on the quality of the inputted multiple sequence alignment, which allows for the detection of conversions in the flanking regions of genes. Additionally, the Hsu et al., 2010 pipeline is capable of determining the direction of the conversion event, identifying which gene was the acceptor and donor of a conversion. Currently, the only way to determine the direction of a conversion with GENECONV is to use

the phylogenetic position of converted genes with relevant sequences and base the origin of converted sequences on the patterns of nucleotide variation outside the converted regions.

1. 4. Goals

Presented here are three studies each examining the various effects of gene conversions in three different gene families of multiple primate genomes. Firstly, chapter 2 deals with gene conversions in the growth hormone gene family of 13 primate species and the stronger homogenizing effects of gene conversion in the Hominidae lineage. Secondly, chapter 3 looks at the folate receptor gene family of six primate genomes and how certain essential folate receptor genes remain unconverted due to the necessary conservation of important folate receptor sequences. Lastly, chapter 4 examines the trypsin gene family of six primate species where strong purifying selection against gene conversion occurring in previously identified critical amino acid regions was observed.

Figure Legends

Figure 1.1. Diagram representing the double strand break repair (DSB repair) and synthesis-dependant strand-annealing (SDSA) models of gene conversion. Both models share similar initiation steps. Portions of sequences highlighted in grey represent ligation via DNA synthesis.

Figure 1.2. Phylogenetic trees for the $^{G}\gamma$- and $^{A}\gamma$-globin genes from human, chimpanzee, and gorilla. Panel (a) represents a tree of exon 3 of the $^{G}\gamma$- and $^{A}\gamma$-globin genes which was found to be unconverted, preserving the expected orthologous relationships. Shown in panel (b) is a phylogenetic tree of exons 1 and 2 of the $^{G}\gamma$- and $^{A}\gamma$-globin genes, which have frequently converted, resulting in a loss of the orthologous relationships. Figure modeled according to the results found by Slightom et al., 1985.

Figure 1.3. Theoretical example of how the GENECONV method for gene conversion detection works. Sites with a '*' in the example of an original inputted sequence are the variable sites. Nucleotides that are in red represent polymorphic sites (those that are free to vary). Nucleotides in black represent conserved sites. Nucleotides in blue represent regions that are identical, possibly due to gene conversion.

Note: Only 1 permutation of 10 000 is shown.

Figure 1.4. Required input, programs/software and steps of the Hsu et al., 2010 pipeline. Boxes represent input/output and ovals represent programs/software.

Figure 1.1

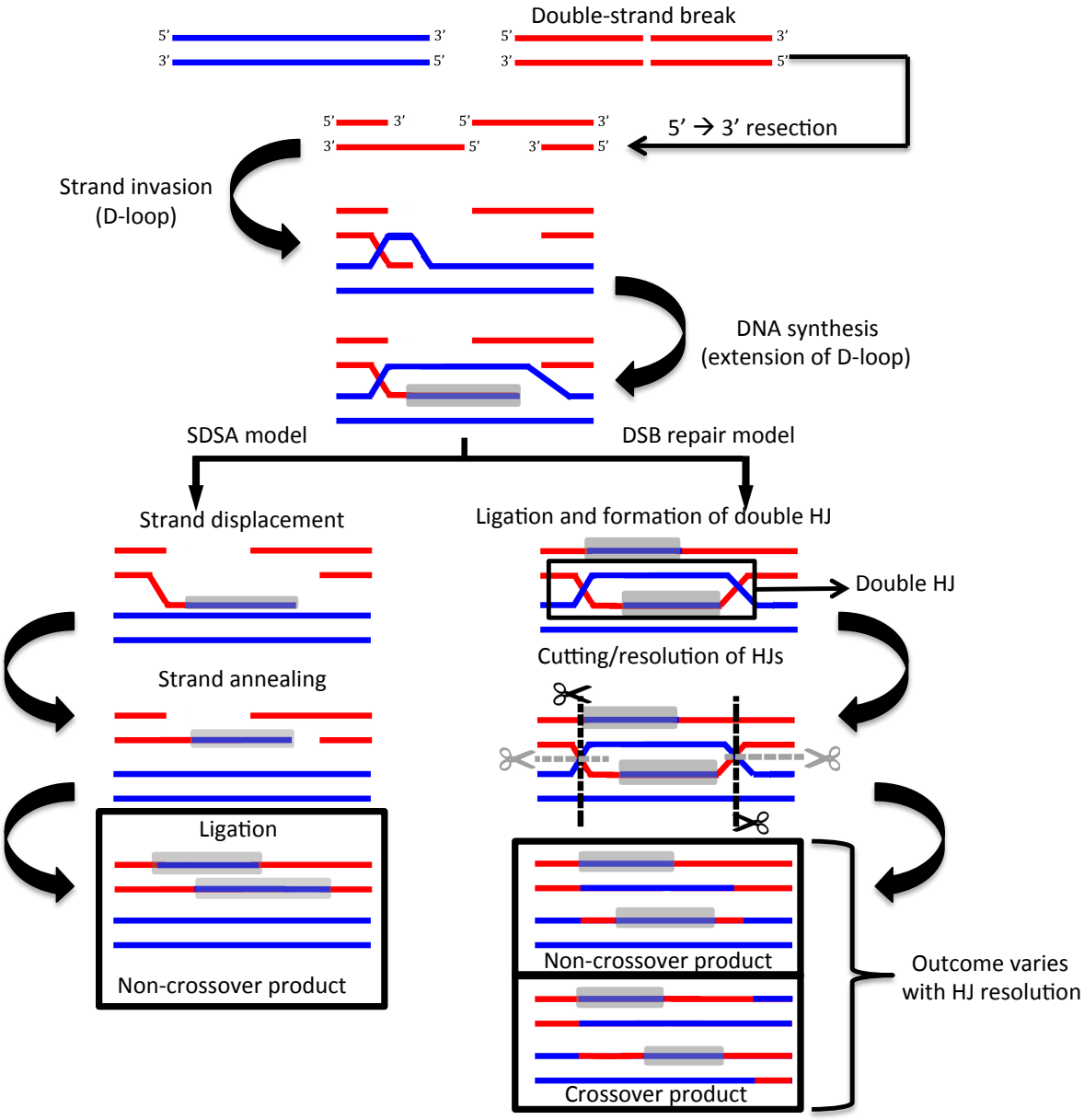

Double-strand break
5′
3′
3′
5′
5′ → 3′ resection
Strand invasion
(D-loop)
DNA synthesis
(extension of D-loop)
SDSA model
DSB repair model
Strand displacement
Strand annealing
Ligation
Non-crossover product
Ligation and formation of double HJ
Double HJ
Cutting/resolution of HJs
Non-crossover product
Crossover product
Outcome varies
with HJ resolution

Figure 1.2

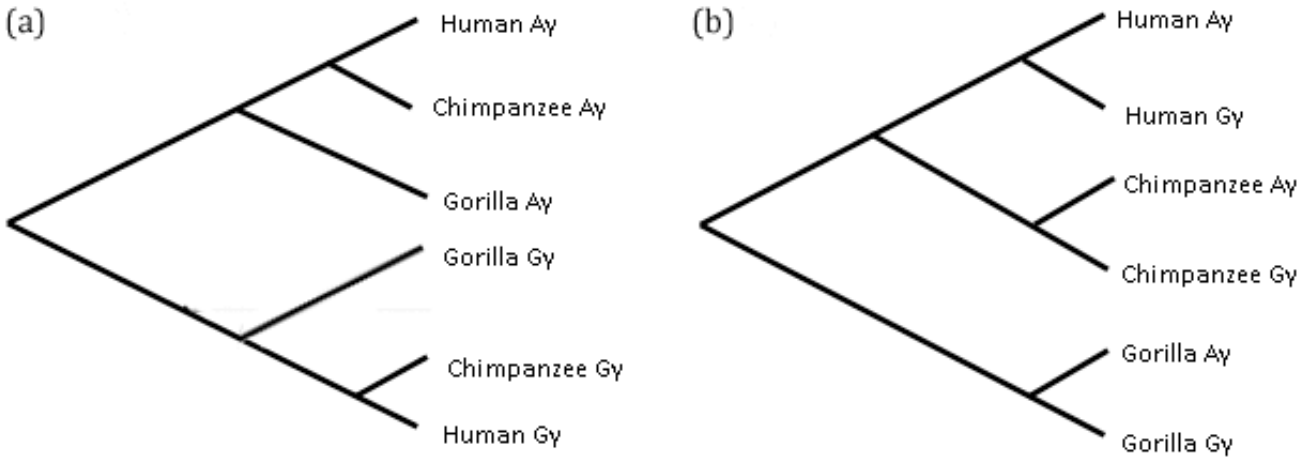

(a)
Human Aγ
Chimpanzee Aγ
Gorilla Aγ
Gorilla Gγ
Chimpanzee Gγ
Human Gγ
(b)
Human Aγ
Human Gγ
Chimpanzee Aγ
Chimpanzee Gγ
Gorilla Aγ
Gorilla Gγ

Figure 1.3

GENECONV:

Example of original inputted sequence:

```
ATG CCA CCC CCG CCT CCA CCC CCG TAA
ATG CCC CCC CCG CCT CCA CCG CCC TAA
ATG CCC CCG CCC CCA CCG CCG CCA TAA
      *   *   *   *   *   *   *
```

Condensed fragments:

```
A CCC CCG CCT CCA CCC G
C CCC CCG CCT CCA CCG C
C CCG CCC CCA CCG CCG A
```

SSCF Permutations:

Original Polymorphic Fragment

1- A CCC CCG CCT CCA CCC G
2- C CCC CCG CCT CCA CCG C

SSCF= 14^2 = 196

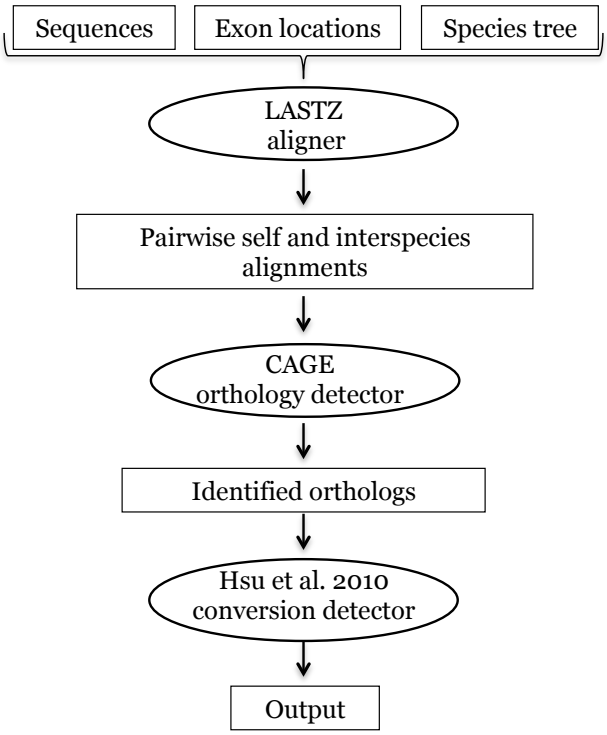
Permuted data # 1

1- CCC CCG A CCT CCA CC G
2- CCC CCG C CCT CCG A CC C

SSCF= $6^2+5^2+3^2$ = 70

P = number of permuted data sets /10000 having an SSCF-value ≥ to that of the observed data

Figure 1.4


Sequences
Exon locations
Species tree
LASTZ aligner
Pairwise self and interspecies alignments
CAGE orthology detector
Identified orthologs
Hsu et al. 2010 conversion detector
Output

REFERENCES:

R.D. Bagnall, K.L. Ayres, P.M. Green, F. Giannelli, Gene conversion and evolution of Xq28 duplicons involved in recurring inversions causing severe hemophilia A, Genome Research 15 (2005) 214-223.

J.B. Bateman, M.D. Fernando, B. von-Bischhoffshaunsen, L. Richter, P.F. Flodman, D. Burch, M.A. Spence, Gene conversion mutation in crystalline, CRYBB2 in a Chilean family with autosomal dominant cataract, Ophthalmology 114 (2007) 425-432.

D. Benovoy, G. Drouin, Ectopic gene conversions in the human genome, Genomics 93 (2009) 27-32.

D. Benovoy, R.T. Morris, A. Morin, G. Drouin, Ectopic gene conversions increase the G + C content of duplicated yeast and Arabidopsis genes, Mol. Biol. Evol. 22 (2005) 1865–1868.

M. Boni, D. Posada, and M. Feldman, An exact nonparametric method for inferring mosaic structure in sequence triplets, Genetics 176 (2007) 1035-1047

J.R. Brooker. *Genetics Analysis and Principles*: McGraw-Hill; 2005.

A.J.R. Carter, S.W. Scherer, Identifying concerted evolution and gene conversion in mammalian gene pairs lasting over 100 million years. BMC Evolutionary Biology 9 (2009) 156-172.

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

S. Das, U. Mohamedy, M. Hirano, M. Nei, N. Nikolaidis, Analysis of the immunoglobulin light chain genes in zebra finch: evolutionary implications, Molecular Biology Evolution (2010) 27:113-120.

G. Drouin, Characterization of the gene conversions between the multigene family members of the yeast genome, J. Mol. Evol. 55 (2002) 14–23.

G. Drouin, F. Prat, M. Ell, G.D.P. Clarke, Detecting and characterizing gene conversions between multigene family members, Mol. Biol. Evol. 16 (1999) 1369–1390.

B. Elliott, C. Richardson, J.Winderbaum, J.A. Nickoloff, M. Jasin, Gene conversion tracts from double-strand break repair in mammalian cells, Mol. Cell. Biol. 18 (1998) 93–101.

K. Ezawa, S. Oota, N. Saitou, Genome-wide search of gene conversions in duplicated genes of mouse and rat, Mol. Biol. Evol. 23 (2006) 927–940.

A. Friaes, A. Toste Rego, J.M. Aragues, L.F. Moura, A. Mirante, M.R. Mascarenhas, T.T. Kay, L.A. Lopes, J.C. Rodrigues, S. Guerra, T. Dias, A.G. Teles, J. Goncalves, CYP21A2 mutations in Portuguese patients with congenital adrenal hyperplasia: IdentiWcation of two novel mutations and characterization of four diVerent partial gene conversions, Molecular Genetics and Metabolism 88 (2006) 58-65.

N. Galtier, G. Piganeau, D. Mouchiroud, and L. Duret. GCcontent evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics (2001) 159:907–911.

D. Graur, W.-H. Li, Fundamentals of molecular evolution, 2nd ed, Sinauer Associates, Sunderland, 2000.

J.E. Haber, G. Ira, A. Malkova and N. Sugawara, Repairing a double-strand chromosome break by homologous recombination: revisiting Robin Holliday's model, Phil. Trans. R. Soc. Lond. 359 (2004) 79-86.

S. Heinen, P. Sanchez-Corral, M. Jackson, L. Strain, J. Goodship, E. Kemp, C. Skerka, T. Jokiranta K. Meyers, E. Wagner, P. Robitaille, J. Esparza-Gordillo, S. Rodriguez De Cordoba, P. Zipfel, T. Goodship, De novo gene conversion in the RCA gene cluster (1q32) causes mutations in complement factor H associated with atypical hemolytic uremic syndrome, Human Mutation 27 (2006) 292-293.

S. Holm, A simple sequential rejective multiple test procedure, Scand. J. Statist 6 (1979) 65-70.

C.H. Hsu, Y. Zhang, R.C. Hardison, Nisc Comparitive Sequence Program, E.D. Green, and W. Miller, An Effective Method for Detecting Gene Conversion Events in Whole Genomes, Journal of Computational Biology 17 (2010) 1281-1297.

S. Keeney, Mechanism and control of meiotic recombination initiation, Curr. Top. Dev. Biol. 52 (2001) 1–53.

G. Kudla, A. Helwak, and L. Lipinski, Gene conversion and GC-content evolution in mammalian Hsp70. Mol. Biol. Evol. (2004) 21:1438–1444.

C.C. Lindegren, Gene conversion in Saccharomyces, Journal of Genetics (1953) 51: 625–637.

R.M. Liskay, A. Letsou, J.L. Stachelek, Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells, Genetics 115 (1987) 161–167.

M.L. Lundqvist, D.L. Middleton, C. Radford, G.W. War, K.E. Magor, Immunoglobulins of the non-galliform birds: antibody expression and repertoire of the duck, Dev Computational Immunology (2006) 30:93-100.

W.T. McCormack, E.A. Hurley, C.B. Thompson, Germ line maintenance of the pseudogene donor pool for somatic immunoglobulin gene conversion in chickens, Loceular Cell Biology (1993) 13:821-830.

J. Meunier, and L. Duret, Recombination drives the evolution of GC-content in the human genome. Mol. biol. Evol. (2004) 21:984–990.

F. Pâques and J. E. Haber, Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae, Mol. Biol. Rev. (1999) 63(2):349-406.

T.D. Petes, C.W. Hill, Recombination between repeated genes in microorganisms, Annu. Rev. Genet. 22 (1988) 147–168.

D. Posada, Evaluation of methods for detecting recombination from DNA sequences: empirical data, Mol. Biol. Evol. 19 (2002) 708–717.

D. Posada, K.A. Crandall, Evaluation of methods for detecting recombination from DNA sequences: computer simulations, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 13757–13762.

G. Santoyo, D. Romero, Gene conversion and converted evolution in bacterial genomes, FEMS Microbial Review (2005) 29:169-183.

S. Sawyer, Statistical tests for detecting gene conversion, Mol. Biol. Evol. 6 (1989) 526–538.

E. Schildkraut, C.A. Miller, J.A. Nickoloff, Gene conversion and deletion frequencies during double strand break repair in human cells are controlled by the distance between direct repeats. Nucleic Acids Res. 33 (2005) 1574–1580.

P. Shen, H.V. Huang, Homologous recombination in Escherichia coli: dependence on substrate length and homology, Genetics 112 (1986) 441–457.

S.K. Shyue, L. Li, B.H. Chang, W.-H. Li, Intronic gene conversion in the evolution of human X-linked color vision genes, Mol. Biol. Evol. 11 (1994) 548–551.

G. Song, C.H. Hsu, C. Riemer, W. Miller, Evaluation of methods for detecting conversion events in gene clusters, BMC Bioinformatics 12 (2011) 1471-2105.

J.L . Slightom, L.Y. Chang, B.F. Koop, M. Goodman, Chimipanzee fetal Gy- and Ay-globin gene nuicleotide sequences provide further evidence of gene conversions in hominine evolution, Molecular Biology and Evolution 2 (1985) 370-389.

F. Stahl, Meiotic recombination in yeast: coronation of the double-strand-break repair model, Cell 87 (1996) 965–968.

J. Szostak, T. Orr-Weaver, R. Rothstein, F. Stahl, The double-strand-break repair model for recombination, Cell 33 (1983) 25–35.

Vanita, V. Sarhadi, A. Reis, M. Jung, D. Singh, K. Sperling, J.R. Singh, J. Burger, A unique form of autosomal dominant cataract explained by gene conversion between B-crystallin B2 and its pseudogene, Medical Genetics 38 (2001) 392-396.

S. Wang, C. Magoulas, D. Hickey, Concerted evolution within a trypsin gene cluster in Drosophila, Mol. Biol. Evol. 16 (1999) 1117–1124.

A.S. Waldman, R.M. Liskay, Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology, Mol. Cell. Biol. 8 (1988) 5350–5357.

Chapter 2.

Gene conversions in the growth hormone gene family of primates: Stronger homogenizing effects in the Hominidae lineage

NOTE: This chapter has been published.

N. Petronella, G. Drouin, Gene conversions in the growth hormone gene family of primates: Stronger homogenizing effects in the Hominidae lineage, Genomics (2011) 98: 173-181

Abstract

In humans, the growth hormone/chorionic somatomammotropin gene family is composed of five highly similar genes. We characterized the gene conversions that occurred between the growth hormone genes of 11 primate species. We detected 48 conversions using GENECONV and others were only detected using phylogenetic analyses. Gene conversions were detected in all species analyzed, their average size (±standard deviation) is 197.8±230.4 nucleotides, the size of the conversions is correlated with sequence similarity and converted regions are significantly more GC-rich than non-converted regions. Gene conversions have a stronger homogenizing effect in Hominidae genes than in other primate species. They are also less frequent in conserved gene regions and towards functionally important genes. This suggests that the high degree of sequence similarity observed between the growth hormone genes of primate species is a consequence of frequent gene conversions in gene regions which are under little selective constraints.

2.1. Introduction

The human growth hormone/chorionic somatomammotropin (hGH/CSH) gene family is located at chromosome 17q22-24. It is comprised of a pituitary growth hormone (GH1), a gene for placental growth hormone (GH2), two chorionic somatomammotropin genes (CSH1 and CSH2) and a chorionic somatomammotropin-like gene (CSHL1; Chen et al., 1989). This gene family is thought to have arisen through three successive duplications. An initial duplication gave rise to a pre-GH and a pre-CSH gene. These two genes were then duplicated to give rise to the GH1, CSH1, GH2 and CSH2 genes. Finally, aCSH1 duplication gave rise to CSHL1 (Chen et al., 1989, Krawczak et al., 1999). Since six genes are present in chimpanzee and rhesus monkey, six genes were likely present before the divergence of Hominidae (the great apes, including human and chimpanzee) and Cercopithecidae, some 30 million years ago and the human lineage lost one (Krawczak et al., 1999, Revol De Mendoza et al., 2004, González et al., 2006, Steiper et al., 2006). Given the age of these genes, one would expect that their nucleotide sequences would be quite different from one another. However, all five human growth hormone genes, and their flanking regions, share from 91 to 99% nucleotide identity (Chen et al., 1989). Previous studies have shown that this unexpectedly high degree of sequence similarity is in large part due to gene conversions between these genes (see below).

The number of growth hormone sequences also varies in other primate species. For example, the spider monkey has six growth hormone genes whereas the marmoset has eight (Revol De Mendoza et al., 2004, González et al., 2006, Wallis et al., 2001). Furthermore, whereas the genomes of higher primates (simians) contain multiple growth hormone genes, those of prosimians (e.g., the bush baby and the slow loris) have a single growth hormone gene (Wallis et al., 2001, Wallis and Wallis, 2002, Adkins et al., 2001). The duplications of higher primate growth hormone genes therefore occurred after the separation of the higher

primate lineage from the prosimian lineage (Adkins et al., 2001, Li et al., 2005, Ye et al., 2005).

Gene conversions are unidirectional recombination events where genetic information is transferred from a donor gene to an acceptor gene (Chen et al., 2007). They often result in the loss of genetic information and in an increase in homogeneity which contributes to maintain sequence similarity and function between duplicated genes (Chen et al., 2007). Detrimental phenotypes can also arise through conversion events where harmful mutations from other sequences are substituted into functional duplicates (Chen et al., 2007, Bischof et al., 2006).

In this context, studying gene conversion between human growth hormone genes is particularly interesting because these different genes are expressed in different tissues and are believed to have different functions. One can therefore hypothesize that conversions between them could change their functions, expression or both. The GH1 gene codes for the main adult growth hormone and is primarily produced in the somatotrophic cells found in the anterior part of the pituitary gland. In contrast, the GH2, CSH1, CSH2 and CSHL1 genes are mainly produced in placental tissue (Chen et al., 1989, Misra-Press et al., 1994, Giordano et al., 1997). These five genes have also shown to be expressed in a wide variety of tissues including reproductive tissues and the retina (Untergasser et al., 1998, Harvey et al., 2003). Furthermore, these 5 different genes have different levels of sequence similarity. For example, while the CSH1 and CSH2 genes code for polypeptides that differ by a single amino acid, they both have 32 amino acid differences with that coded by the GH1 gene (results not shown). Gene conversions between different genes may therefore result in different expression patterns or function.

The human growth hormone gene family has been the subject of numerous evolutionary studies, many of which documented the presence of gene conversions between

the growth hormone genes of humans and other primates (Chen et al., 1989, Krawczak et al., 1999, Revol De Mendoza et al., 2004, González et al., 2006, Li et al., 2005, Ye et al., 2005, Millar et al., 2003, Sedman et al., 2008, Wolf et al., 2009). However, many of these studies focused on gene conversions in the promoter regions and none of them have assessed the characteristics and impact of gene conversions in all currently known primate growth hormone sequences. Here, we present an extensive characterization of the gene conversions that occurred between the growth hormone genes of primates. We also hypothesized that since the GH1 gene codes for the main adult growth hormone in humans, chimpanzee and rhesus monkey (this gene is also called the GHN gene in chimpanzee and rhesus monkey) that, due to selective pressures to preserve its function, this gene will always act as a donor in conversion events, and never as an acceptor. Our results show that gene conversions are more frequent between Hominidae genes than those of other primate species. They also occur less frequently in more conserved gene regions and towards the more functionally important GH1 gene.

2.2. Materials and methods

The human growth hormone sequences, and those of 12 other primate species, were obtained from NCBI (http://www.ncbi.nlm.nih.gov/). The complete list of the GenBank accession numbers of the 56 growth hormone genes analyzed here are listed in Supplementary Table 2.1. All the genes studied are made up of five exons and four introns. Furthermore, all intron and exon lengths among these sequences are similar (Figure 2.1 and results not shown). In all cases, the sequences used started with the ATG initiation codon and ended with the stop codon.

Sequences were manipulated using BioEdit (Hall, 1999). Alignments were performed using ClustalW (Thompson et al., 1994) and PRANK

(http://www.ebi.ac.uk/goldman-srv/webPRANK/) (Loytynoja and Goldman 2010). Phylogenetic trees were constructed with the neighbor joining method implemented in MEGA 4 (Tamura et al., 2007). Default parameters were used: Kimura's two parameter model, maximum composite likelihood with 1st, 2nd, 3rd and non-coding codon positions and 1000 replications for bootstrap analyses. The average amount of sequence differences between all growth hormone genes of each species was also calculated using MEGA 4 using the p-distance model (i.e., the proportion of different nucleotides without any correction for multiple substitutions) and the standard errors of these average values were calculated using 500 bootstrap replicates.

The GENECONV 1.81 program (http://www.math.wustl.edu/sawyer/geneconv/) implementing Sawyer's statistical method for detecting gene conversion was used to identify gene conversion events (Sawyer, 1989). The accuracy of GENECONV has been previously tested and this method was found to be a useful tool to detect gene conversions except when the sequences analyzed are all very similar or when the sequences compared are more than 20% divergent (Drouin et al., 1999, Posada et al., 2001, Posada, 2002). If all sequences are very similar, this method will fail to detect gene conversions (false negative results) because there will not be enough polymorphic sites to permute. Conversely, this method gives false positive results when the sequences being compared are more than 20% divergent (Posada et al., 2001). The growth hormone genes of each species were analyzed using a mismatch penalty of 2. Only the global fragments with simulated P-values of less than 0.05 were considered significant. We did not consider pairwise fragments, even when they had simulated P-values of less than 0.05, because these fragments are not corrected for multiple comparisons.

The distribution of gene conversions along the sequences of growth hormone genes was tested using a chi-square test. For each species, the aligned sequences were divided

into four regions of equal length and the number of conversion events in each region was counted. WebLogo (http://weblogo.berkeley.edu/) was used to assess the degree of conservation of the amino acids along the sequence of the growth hormone protein.

2.3. Results

2.3.1. Gene conversions detected with GENECONV

Using GENECONV, gene conversions were detected between the genomic copies of the growth hormone genes of all 11 species with multiple growth hormone genes (the growth hormone gene is single copy in the bush baby and slow loris genomes; Table 2.1). A total of 48 genomic conversions were detected with the smallest being 14 nucleotides long and the largest 1019 nucleotides long. The average size (±standard deviation) of these 48 conversions is 197.8±230.4 nucleotides. Although this average size does not take into account the variation found in each gene and in different individuals of each species (i.e., it is based on reference sequences obtained from NCBI), this average size is similar to the mean length of 55 to 290 bp measured using sperm analyses (Jeffreys and May, 2004). There are 8 conversions limited to intron sequences and their average size is 42.5±25.3 nucleotides long. There are 13 conversions limited to exon sequences and their average size is 70.6±42.5 nucleotides long. The average size of conversions limited to intron sequences and those limited to exon sequences are not statistically different (t-test, $p=0.07$). In contrast, the average size of the 27 conversions spanning both introns and exons are 309.8±260.4 nucleotides long and are significantly longer than both the size of the conversions limited to introns or the conversions limited to exons (t- tests, $p=0.00001$ and $p=0.00007$, respectively).

The lengths of the converted regions are correlated with the degree of similarity of the genes between which they occur. A positive correlation exists between the lengths of converted regions identified by GENECONV (Table 2.1) and the percent similarity of the

sequences found 100 bp before and after each converted region (r=0.46,p=0.0009).

Converted regions are also more GC-rich than non-converted regions. When considering the 23 conversions longer than 100 bp identified by GENECONV (Table 2.1), the average GC-content (±standard error) of these converted regions (57.17%±0.77) is significantly higher than the non-converted ones (55.44%±0.27; t- test, p=0.004). For this analysis, only conversions larger than 100 bp were considered to minimize the effect of stochastic variation.

For 40 of these 48 conversions, the origin of the converted sequences was determined based on the patterns of nucleotide variation outside the converted regions and the phylogenetic position of the relevant sequences (Figures 2.2 and 2.3). The polarity of the other eight conversions could not be established due to the lack of sequence information (these conversions are 85 nucleotides long or shorter) resulting in either direction being equally likely. In the howler monkey genes, one of the three species where five or more conversions were detected, there is no evidence of a bias in the direction of the conversions. For example, gene 2 of the howler monkey is twice a donor and twice an acceptor and gene 4 is once a donor and once an acceptor. Similarly, the chimpanzee PL-C gene is three times a donor and twice an acceptor. However, the situation is different for the chimpanzee PL-A gene that accepted seven conversions and never converted any other gene. Moreover, in the chimpanzee, the GH1 gene was three times a donor, and the GH2 and PL-D genes were both twice donors, without any of these three genes ever accepting a conversion. In the rhesus monkey, the GH2 gene was twice a donor and never an acceptor.

Surprisingly, the results shown in Table 2.1 seem to suggest that gene conversions are relatively rare in the human genome. In fact, our GENCONV analyses detected only two gene conversions between human genes compared with as many as eleven conversion events between chimpanzee genes (Table 2.1). However, the phylogenetic and sequence

similarity analyses we present below demonstrate that gene conversions are indeed frequent between human growth hormone genes.

2.3.2. Phylogenetic analyses

Phylogenetic trees can be used to detect gene conversion events because conversions between paralogous genes will often cause them to group together rather than with their orthologous gene members in other related species (Drouin et al., 1999, Graur and Li, 2000). We therefore built phylogenetic trees to complement our GENECONV results. Since our GENECONV results clearly indicate that introns contain many converted regions, these trees were built using genomic sequences.

Figure 2.2 shows the phylogenetic tree for the three primate species for which the complete growth hormone gene locus (and genome) has been sequenced. This tree contains two main groups, one composed uniquely of rhesus monkey sequences and one composed of human and chimpanzee sequences. Such a topology would not be expected if no gene conversions were present. Without gene conversions, the orthologous genes of each species (e.g., the GH1 gene) should be grouped together [i.e., (human GH1, chimpanzee GH1), rhesus monkey GH1)]. The fact that all rhesus monkey sequences are grouped together and separated by long branches is likely the result of the fact that the gene conversions found in the rhesus monkey genome are few and short whereas those in the human and chimpanzee genomes are numerous and longer (Table 2.1), hence the clustering of human and chimpanzee genes. The phylogenetic effects of conversions are particularly evident between the human CSH1 and CSH2 where these two paralogous genes are grouped together with a 91% bootstrap value, rather than with their orthologous genes from chimpanzee or rhesus monkey. The strongly supported grouping of these two human genes is likely the result of the 701 nucleotide long gene conversion between them (Table 2.1). Furthermore, the

direction of the conversion was most likely from the human CSH1 gene to the human CHS2 gene because both these genes are a sister group to the chimpanzee PL-D gene. This tree also shows that a gene conversion occurred between the GH1 and GH2 genes in the common ancestor of human and chimpanzee. This explains why these two paralogous genes are grouped together while each still retains its human and chimpanzee orthologous relationship.

The shorter branch lengths connecting the human and chimpanzee sequences, compared with those connecting rhesus monkey sequences, suggest that gene conversions have a stronger homogenization effects between Hominidae sequences than between Cercopithecidae sequences (Figure 2.2). This suggestion is supported by analyses of the average percent nucleotide differences between all the growth hormone genes found in the genomes of all 11 species having multiple growth hormone gene sequences (Table 2.2). These results, whether alignment gaps found in one sequence are removed from all sequences or only between pairs of sequences, show that human and chimpanzee (Hominidae) gene sequences are, on average, more similar to one another than those of any other primate species.

Figure 2.3 shows the phylogenetic tree for the sequences of all primate species for which growth hormone genomic sequences are available. Although some groups most likely represent groups of orthologous genes, such as the black-handed spider monkey GHN and howler monkey 1 sequences, most do not. Here again, the numerous gene conversions have obscured the relationships of the different genes. For example, it is impossible to determine which gene in other species is orthologous to genes 1, 4, 5 and 6 from the northern white-cheeked gibbon because these four genes form a strongly supported group whose sister group is a mixture of human and chimpanzee genes. Furthermore, perhaps excluding gene 1 of the golden snub-nosed monkey, this tree cannot be used to establish which gene from the

other 10 primate species with multiple growth hormone genes is likely orthologous to the rhesus monkey GH1 gene.

The relationships between the human and chimpanzee sequences previously found in Figure 2.2 are also observed in Figure 2.3, providing further evidence for frequent gene conversion in their common ancestor. This tree also provides strong evidence for three gene conversions that had not been detected using the GENECONV method. Genes 4 and 5 of the assamese macaque, genes 1 and 6 of the northern white-cheeked gibbon and genes 2 and 3 of the northern white-cheeked gibbon form three strongly supported pairs of genes which suggest that they are evolving in concert due to frequent gene conversions (Figure 2.3). The fact that the GENECONV method did not detect these conversions is due to the high degree of similarity of these genes over their whole length. For example, genes 4 and 5 of the assamese macaque are 97.6% similar at the nucleotide level (results not shown). The high degree of sequence similarity of these three pairs of genes is also represented by the short length of the branches between the members of each pair (Figure 2.3). Using the scale bar of this figure, one can measure that each pair has less than 5% differences at the nucleotide level (including exons and introns).

2.3.3. Number of gene conversions in different species

The number of gene conversions detected in the 11 species having multiple growth hormone genes is highly variable (Table 2.1). Whereas a single gene conversion was detected in the genomes of the Assamese macaque and the black-handed spider monkey, 11 were detected between chimpanzee genes. Although there is a weak correlation between the number of growth hormone genes in a species and the number of conversions found in that species (r^2=0.25), this correlation is not significant (p=0.14). This suggests that the

number of conversion events is not proportional to the number of genes present in a genome.

2.3.4. Distribution of gene conversions in growth hormone genes

The distribution of gene conversions was analyzed in order to determine whether gene conversions occur with equal frequency in the different regions of growth hormone genes. Since visual inspection of the distribution of gene conversions (Figure 2.1) suggests that gene conversions are less frequent in the last quarter of these sequences, the genes of each species were divided into four regions of equal length. The number of conversions that occurred in each region was tested using a chi-square test. Note that if a conversion spanned two or three gene regions it was counted as having occurred in all of them. This test shows that gene conversions are not evenly distributed throughout growth hormone genes. With 20, 19, 16 and 6 conversions in each of the consecutive regions, there are significantly more conversions in the first three-quarters of the gene than in the last quarter of the gene ($p=0.045$). Interestingly, this uneven distribution reflects the uneven distribution of conserved amino acids along the sequences of growth hormone sequences where amino acids from 140 to 217 are more conserved than those found in positions 1 to 139 (Figure 2.4).

2.4. Discussion

Although several studies previously addressed the impact of gene conversions on the evolution of growth hormone genes, our study is the first to present comprehensive analyses of primate growth hormone genes using genomic sequences of the coding regions. Several previous studies analyzed haplotypes and focused on the occurrence and effects of gene conversion events in the promoter regions of human genes (Krawczak et al., 1999,

Giordano et al., 1997, Millar et al., 2003, Sedman et al., 2008, Wolf et al., 2009). These studies showed that gene conversions are frequent in the promoter regions of growth hormone genes and that the direction of the gene conversions are often from the four placental genes towards the promoter region of the GH1 gene.

Our results are consistent with previous studies. The long gene conversion that we detected between the human CSH1 and CSH2 genes has been observed by at least three previous studies (Table 2.1; Chen et al., 1989, Sedman et al., 2008, Hirt et al., 1987). Similarly, in the chimpanzee genome, the 368 nucleotide long conversion between the PL-C and GH1 (GHN) genes we detected, and the 293 nucleotide long conversion between the PLA and GH2 (GHV) genes, were also detected by Revol de Mendoza et al., (2004) simply on the basis of the unexpectedly high amino acid sequence similarity they observed in these regions. However, these authors did not detect the other 9 conversions we detected in this genome (Table 2.1). Li et al., (2005) analyzed the genomic sequences of the howler monkey, the red-bellied titi and the white-faced saki growth hormone genes and found 4, 3 and 1 conversions in these species, respectively. Even though we found more conversions than them, our results, i.e., the number and lengths of conversions, are largely consistent with prior studies. Any differences are likely due to the fact that they used a variety of mismatch penalties (scale values) when they ran GENECONV and that they reported the longest estimates they obtained from several analyses. In contrast, the fact that the GENECONV analysis of the 6 rhesus monkey growth hormone genes performed by Gonzáles Alvarez et al., (2006) revealed only three gene conversions, compared with our 10 (Table 2.1), is due to the fact that they did not allow for any mismatches within converted regions. Finally, we do not know why Ye et al., (2005) detected only 3 gene conversions among the 21 growth hormone gene sequences they analyzed. However, we suspect it is because they performed a single GENECONV analysis on all 21 aligned sequences from diverse primate species

rather than performing separate GENECONV analyses for each species. This stresses the fact that consistent and appropriate analysis criteria have to be used to perform in depth comparative genomic analyses.

One implication of the high frequency of gene conversion events between growth hormone genes is that it makes it pointless to try to establish the orthologous relationships between genes in different primate species. For example, given that the chimpanzee and rhesus monkey genomes both have 6 growth hormone genes, one would expect that these twelve genes could be grouped into 6 pairs of orthologous genes. However, as demonstrated by the phylogenies shown in Figures 2.2 and 2.3, defining such groups is impossible. In fact, even if the precise chromosomal gene order was known for both species, one could not claim that specific genes are orthologous based on their chromosomal location because, although the locations are conserved, the sequence of each gene no longer reflects their common ancestry but is simply the result of the diverse conversion events that affected each gene. The variable number of genes and the repeated conversions that occurred between them therefore make it impossible to infer the order of duplication or lost events which were responsible for the evolution of these genes in each species.

The fact that the conversions limited to intron and exon sequences are significantly smaller than those spanning both exons and introns is likely due to different factors. Since the lengths of conversions are inversely correlated with the amount of sequence similarity between the donor and acceptor sequences, the lengths of conversion tracts between non-coding sequences are expected to be shorter because non-coding regions evolve faster than coding regions and are therefore more dissimilar (Shen et al., 1986, Waldman and Liskay, 1988, Drouin, 1998, Lukacsovich and Waldman, 1999). In contrast, the small lengths of conversions limited to (more similar) exon sequences likely reflects the action of purifying selection eliminating longer conversions affecting essential amino acids. Therefore, the fact

that the conversions spanning both exons and introns are significantly larger than those limited to introns or exon sequences likely represents conversions that occurred in regions under little selective pressures. This interpretation is consistent with the fact that in the rhesus monkey the three conversions where GH1 (coding for the main adult growth hormone) was the acceptor gene are all limited to intron sequences (Table 2.1). Furthermore, it is also consistent with the fact that among the conversions we detected between chimpanzee growth hormone genes, the GH1 gene was never an acceptor sequence. Therefore, our results are in agreement with those of previous studies that found that gene conversions affecting functional sites are selected against (Shyue et al., 1994, Noonan et al., 2004, Verrelli and Tishkoff, 2004).

The significant correlation ($p=0.0009$) between the length of the conversions and the degree of similarity of the regions flanking these conversions is consistent with previous studies in bacterial, yeast and mammalian genes (Shen and Huang, 1986, Waldman and Liskay, 1988, Drouin, 1998, Elliott et al., 1998, Lukacsovich and Waldman, 1999). Longer conversions therefore tend to occur between more similar genes. As discussed further below, this is particularly relevant to the evolution of Hominidae growth hormone genes because their overall higher average sequence similarities (Table 2.2) are responsible for their higher degree of homogenization. In other words, there is a positive feedback loop between gene conversion frequency and sequence similarity. The high degree of similarity between these sequences allows them to convert one another frequently. In turn, these frequent conversions lead to higher similarity, and so on.

These frequent conversions not only lead to higher sequence similarity, but also to higher GC-content. The fact that the average GC content of converted regions is significantly higher ($p=0.004$) than those of non-converted regions is consistent with the biased gene conversion hypothesis which posits that, due to biases in DNA repair mechanisms, higher

gene conversions frequencies will lead to higher GC-content (Benovoy et al., 2005). This observation demonstrates that gene conversions are so frequent between the growth hormone genes of primates that they affect their GC-content.

The lack of significant correlation between the number of conversions observed in a species and the number of genes present in its genome suggest that the number of genes has no influence on the frequency of gene conversion events. This is in contrast with the situation in bacterial genomes where the number of conversions is correlated with the size of the gene families, a result which likely reflects more frequent conversion events as the number of potential conversion partner increases (Morris and Drouin, 2007). However, our conclusion assumes that all gene conversions were detected. Since the phylogenetic analyses presented above demonstrate that the GENECONV method did not detect all gene conversions present in some genomes, this conclusion is only tentative. Similarly, the fact that we cannot accurately calculate the number of gene conversions that occurred between the human genes, and that we do not know the genomic gene arrangement of the species other than those of rhesus-monkey, chimpanzee and human, precludes us from assessing whether the proximity of the duplicated genes affects the number of gene conversions observed (Goldman and Lichten, 1996).

The fact that the GENECONV method failed to detect all gene conversions in all genomes is the result of the high similarity of the genes found in these genomes (Drouin et al., 1999, Posada and Crandall, 2001, Posada, 2002). Since this method assesses whether the nucleotide differences found between two sequences are randomly distributed along the length of these sequences, it cannot detect if a given region has less differences than its flanking regions if the two sequences are almost identical over their whole length. This stresses the fact that other methods, such as phylogenetic trees, have to be used to identify gene conversions spanning the whole length of genes (Drouin et al., 1999). Phylogenetic

trees are particularly useful in this respect because they not only show the topological clustering of converted genes but also provide a visual representation of the degree of similarity between these sequences. Their main drawback is that they cannot easily be used to define the location of converted regions (Drouin et al., 1999).

Previous studies have shown that conversion between genomic and cDNA copies lead to an excess of conversions in the 3′-end of genes (Fink, 1987, Derr et al., 1991, Mourier and Jeffares, 2003). The fact that gene conversions are less abundant at the 3′-end of growth hormone genes suggests that most conversion events observed in this gene family occur between genomic copies and not between genomic copies and their cDNA molecules (Figure 2.1). Moreover, the fact that the uneven distribution of converted regions reflects the uneven distribution of conserved amino acids along the sequences of growth hormone sequences suggests that the carboxy-end of growth hormone proteins are under higher selective constraints compared to those in their amino-terminal and central regions (Figure 2.4). This biased distribution of conserved amino acids towards the carboxy-end of the growth hormone protein has previously been documented (Watahiki et al., 1989). These authors showed that 13 of the 33 amino acids conserved in 21 vertebrate species are located in the last 31 amino acids of the growth hormone protein. This part of the protein constitutes a conserved domain, called the GD5 domain, and is likely involved in the formation and stabilization of these proteins (Watahiki et al., 1989). X-ray crystallography has shown that the carboxy-end of the growth hormone protein is found buried in the growth hormone and prolactin receptors with which it interacts (Somers et al., 1994). This protein region also includes three residues (Arg 167, Asp 171 and Glu 174, following the numbering of the human growth hormone), which are critical to bind and distinguish between these two receptors (Somers et al., 1994). Therefore, the significantly lower frequency of gene conversion at the 3′-end of primate growth hormone genes is likely the

result of stronger purifying selection at the carboxy-end of growth hormone proteins (Table 2.1, Figure 2.1 and 2.4). This is consistent with other studies, which have shown that functionally important regions are immune to gene conversions (Shyue et al., 1994, Noonan et al., 2004, Verrelli and Tishkoff, 2004).

Our results concerning the directions of gene conversions are also consistent with the suggestion that the biased directionality we observed is the results of purifying selection. This is supported by the fact that gene conversion biases are not observed in most of the species studied. However, in the few instances where gene conversion biases are present, these biases are always from the most functionally important gene (i.e., the GH1 and GH2 genes) to genes of lesser functional importance (e.g., the human CSH genes; Table 2.1). As mentioned above, the GH1 gene (also called the GHN gene in chimpanzee and rhesus monkey) codes for the main adult growth hormone and is primarily transcribed in the pituitary gland. In contrast, the GH2, CSH1, CSH2 and CSHL1 genes are mainly transcribed in placental tissue (Chen et al., 1989, Misra-Press et al., 1994, Giordano et al., 1997, Untergasser et al., 1998, Harvey et al., 2003). The higher functional importance of the human GH1 gene, relative to the four other growth hormone genes, is reflected by the fact that most growth hormone deficiency cases are due to mutations (mainly deletions) in the GH1 gene and never in the three CSH growth hormone genes (Procter et al., 1998, Prager et al., 2003). This is consistent with the existence of a single growth hormone gene in prosimians and most mammals and the fact that there is a single worldwide GH1 variant in human populations (Li et al., 2005, Ye et al., 2005, Sedman et al., 2008). It is also consistent with the fact that no clinical disorders are observed when CSH genes are deleted or when their corresponding hormones are missing (Sedman et al., 2008). Finally, the few cases where growth hormone deficiencies have been observed in the presence of an intact GH1 gene are believed to be due to mutations in cis-acting GH1 regulatory sequences (Procter et

al., 1998, Cacciari et al., 1994). The biased gene conversion direction observed for the GH2 gene of chimpanzee and rhesus monkey, where this gene always acts as a donor and never as an acceptor, also suggests that conversion towards this gene are selected against. Although GH2 likely does not have a direct effect on fetal growth, because it is not detectable in the fetal circulation, it is believed to be involved in driving normal fetal growth (Alsat et al., 1997, McIntyre et al., 2000). This suggests that deleterious conversions in functionally important genes are being eliminated by natural selection whereas neutral conversions are not. This is consistent with the suggestion that the high frequency of gene conversions in the promoter region of the GH1 gene reflects relaxed selective constraints in that region (Krawczak et al., 1999, Sedman et al., 2008).

In conclusion, our results show that gene conversions occurred between the growth hormone genes of all 11 primates we analyzed but that they had a stronger homogenizing effect between the Hominidae genes than between those of other primate species. The fact that conversions are less frequent in conserved gene regions, and towards functionally important genes, suggests that detrimental conversion events are being eliminated by purifying selection. Therefore, the high degree of sequence similarity observed between the growth hormone genes of primate species is a consequence of frequent gene conversions in gene regions that are under little selective constraints.

Table 2.1.
Gene conversion events detected using GENECONV

Species	Sequences	Length	Location	Sequences	Length	Location
Alouatta seniculus (Howler monkey)	5(D);2(A)	273	last 102bp of exon 4 to first 170bp of intron 4	5(D);2(A)	93	first 35bp of exon 5 to last 76bp of exon 5
	2(D);3(A)	651	last 245bp of intron 1 to last 105bp of exon 3	4(D);5(A)	38	first 126bp of intron 1 to last 91bp of intron 1
	1(?);5(?)	85	first bp of exon 1 to last 145bp of intron 1	2(D);4(A)	93	last 86p of intron 2 to last 112bp of exon 3
	3(D);4(A)	106	last 55bp of intron 3 to first 50bp of intron 4			
Hylobates Leucogenys (Northern white-cheeked gibbon)	5(D);4(A)	605	first bp of exon 1 to last 58bp of intron 2	7(?);1 (?)*	57	first 41bp of intron 1 to first 97bp of intron 1
	4(D);7(A)	35	first 6bp of exon 1 to first 30bp of intron 1			
Pithecia pithecia (White-faced saki)	4(D);2(A)	239	last 3bp of exon 2 to first 23bp of exon 3	1(?);3(?)*	29	first 8bp of intron 2 to first 36bp of intron 2
	4(D);3(A)*	115	last 55bp of exon 4 to first 59bp of intron 4			
Rhinopithecus roxellana (Golden snub-nosed monkey)	1(?);3(?)	41	first 47 bp of exon 2 to last 73 of exon 2	4(D);2(A)	223	first 8 of exon 2 to first 70 intron 2
Callicebus moloch (Red-bellied titi)	1(D);2(A)	44	last 4bp of exon 1 to first 39bp of intron 1	1(D);3(A)*	103	last 17bp of exon 3 to last 7bp of intron 3
	1(D);4(A)*	187	first 36bp of exon 5 to last bp of exon 5	3(D)*;2(A)	89	last 56bp of exon 4 to first 32bp of intron 4
Macaca assamensis (Assamese macaque)	2(D);1(A)	120	last 49bp of intron 1 to last 90bp of exon 2			
Pygathrix nemaeus (Red shanked douc langur)	3*(D);2(A)	379	last 181bp of intron 4 to last bp of exon 5	3(?);5(?)	30	first 18bp of exon 3 to first 47bp of exon 3
	3*(D);1(A)	54	first 47bp of exon2 to first 100bp exon2	4(D);2(A)	1019	first bp of exon1 to last bp of exon 4
	3(?);4(?)	42	first 89bp of exon 4 to first 130 of exon 4			
Ateles geoffroyi (Black-handed spider monkey)	GHV(?);GHC(?)	14	first 160bp of intron 1 to last 79bp of intron 1			
Macaca mulatta (Rhesus monkey)	CSH1(D);GH1(A)	87	first 140bp of intron 1 to last 34bp of intron 1	GH1(D);CSH3(A)	56	first 28bp of exon 3 to first 83bp of exon 3
	CSH2(D);CSH4(A)	93	first 12bp of exon 2 to first 104 of exon 2	CSH3(D);CSH2(A)	23	last 32bp of intron 1 to last 10bp of intron 1
	GH2(D);GH1(A)	165	first 186bp of intron 2 to first 22bp of intron 3	CSH2(D);GH1(A)	25	First 31bp of intron 1 to fist 55bp of intron 1
	CSH3(D);CSH4(A)	67	first 106 of exon 4 to first 8bp of intron 4	GH2(D);CSH4(A)	22	first 23bp of exon 3 to last 75bp of exon 3
	GH1(?);GH2(?)	42	first 48bp of exon 2 to first 89bp of exon 2			
Pan troglodytes (Chimpanzee)	GH1(D);PL-C(A)	368	first bp exon 1 to last 78bp of exon 2	PL-B(D);PL-A(A)	293	last 43bp of exon 3 to last bp of exon 4
	GH2(D); PL-A(A)	293	first 16bp of intron 2 to last 36bp of exon 3	PL-C(D);PL-A(A)	67	first 34bp of intron 1 to last 166bp of intron 1
	PL-D(D);PL-A(A)	638	last 40bp of exon 3 to last 11bp exon 5	GH1(D);PL-A(A)	73	last 94bp of exon 3 to last 22bp of exon 3
	PL-D(D);PL-C(A)	675	last 81bp of exon2 to last 15bp of exon 4	GH2(D);PL-A(A)	68	first 15bp of exon 2 to first 82bp of exon 2
	PL-C(D);PL-B(A)	590	first 25bp of intron 2 to first 12bp of intron 4	GH1(D);GH2(A)	114	first 15bp of exon 2 to last 32bp of exon2
	PL-C(D);PL-A(A)	248	first 2bp of intron 3 to last 9bp exon 4			
Human	CSH2(D);CSH1(A)	701	last 76bp exon 2 to last 26bp exon 4	CSHL1(D);CSH2(A)	148	last 11bp of exon 2 to first 136bp of intron 2

Notes. (A) and (D) indicate acceptor and donor sequences respectively. Interrogation points (?) indicate conversions which the polarity could not be determined. Asterisks (*) indicate sequences that are likely pseudogenes.

Table 2.2

Average similarities of the growth hormone genes found in the genome of primate species

	Complete deletion	Pairwise deletion
Species	Mean ± s. e.	Mean ± s. e.
Alouatta seniculus (Howler monkey)	0.120 ± 0.006	0.120 ± 0.005
Hylobates leucogenys (Northern white-cheeked gibbon)	0.068 ± 0.005	0.070 ± 0.005
Pithecia pithecia (White-faced saki)	0.138 ± 0.006	0.139 ± 0.006
Rhinopithecus roxellana (Golden snub- nosed monkey)	0.069 ± 0.005	0.070 ± 0.005
Callicebus moloch (Red-bellied titi)	0.129 ± 0.006	0.129 ± 0.005
Macaca assamensis (Assamese macaque)	0.098 ± 0.005	0.098 ± 0.005
Pygathrix nemaeus (Red shanked douc langur)	0.138 ± 0.006	0.139 ± 0.005
Ateles geoffroyi (Black-handed spider monkey)	0.135 ± 0.007	0.134 ± 0.008
Macaca mulatta (Rhesus monkey)	0.098 ± 0.005	0.098 ± 0.005
Pan troglodytes (Chimpanzee)	0.060 ± 0.004	0.060 ± 0.005
Homo sapiens (Human)	0.060 ± 0.004	0.061 ± 0.004

Notes. Complete deletion refers to average similarities calculated with all gaps deleted from all sequences. Pairwise deletion refers to the average similarities calculated with gaps deleted only from the pairs of sequences being compared. s.e., standard error

Figure Legends:

Figure 2.1. Graphical representation of the distribution of the 48 gene conversions detected between growth hormone genes. The top part represents the exon-intron structure of these genes in all species examined. Whereas the 5 exons of these genes have the same length in all species (black boxes labeled 1 to 5), the lengths of the introns show some variability, as indicated. The lines shown in the bottom part represent the approximate area covered by each conversion, with each species having a different color code, as indicated. The species order in which these conversions are shown corresponds to the order in which they are listed in Table 2.1.

Figure 2.2. Phylogeny of the chimpanzee, macaque and human growth hormone genes. The scale bar represents 1% sequence divergence and the numbers at the nodes are the percent bootstrap support.

Figure 2.3. Phylogeny of the growth hormone genes of all 13 primate species. The scale bar represents 5% sequence divergence and the numbers at the nodes are the percent bootstrap support. Stars (*) indicate sequences which are likely pseudogenes.

Figure 2.4. LOGO representation of the degree of amino acid conservation along the 48 primate growth hormone protein sequences. The height of each letter is proportional to the degree of conservation of each amino acid (in one letter code). Numbers below each amino acid represent their position in the growth hormone sequence.

Figure 2.1

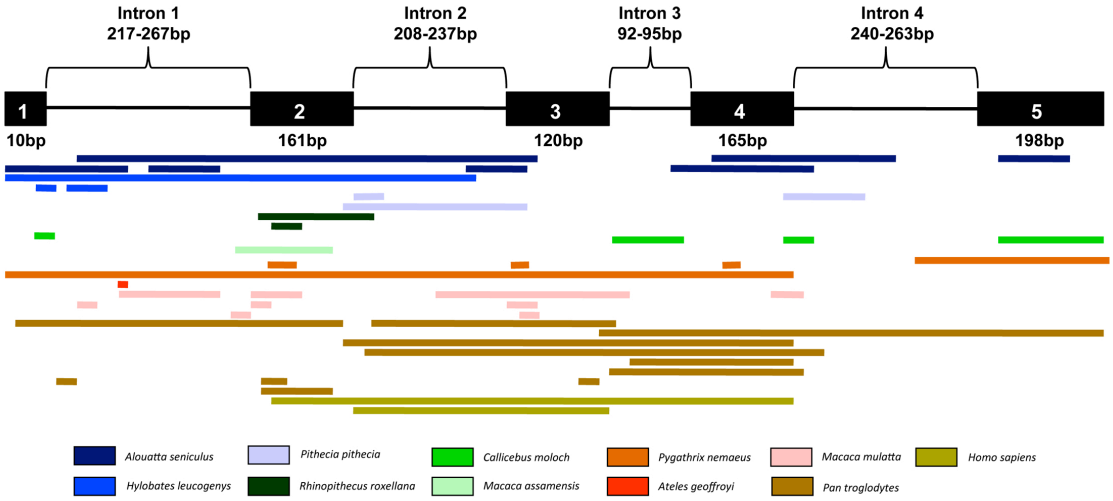


Figure 2.2

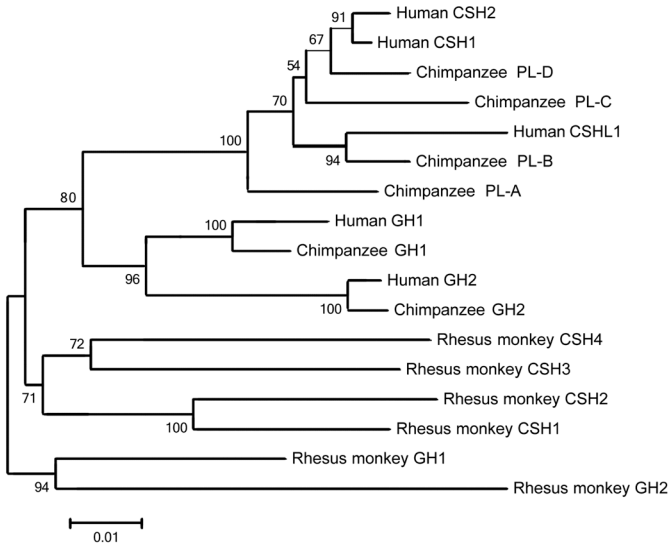


Figure 2.3

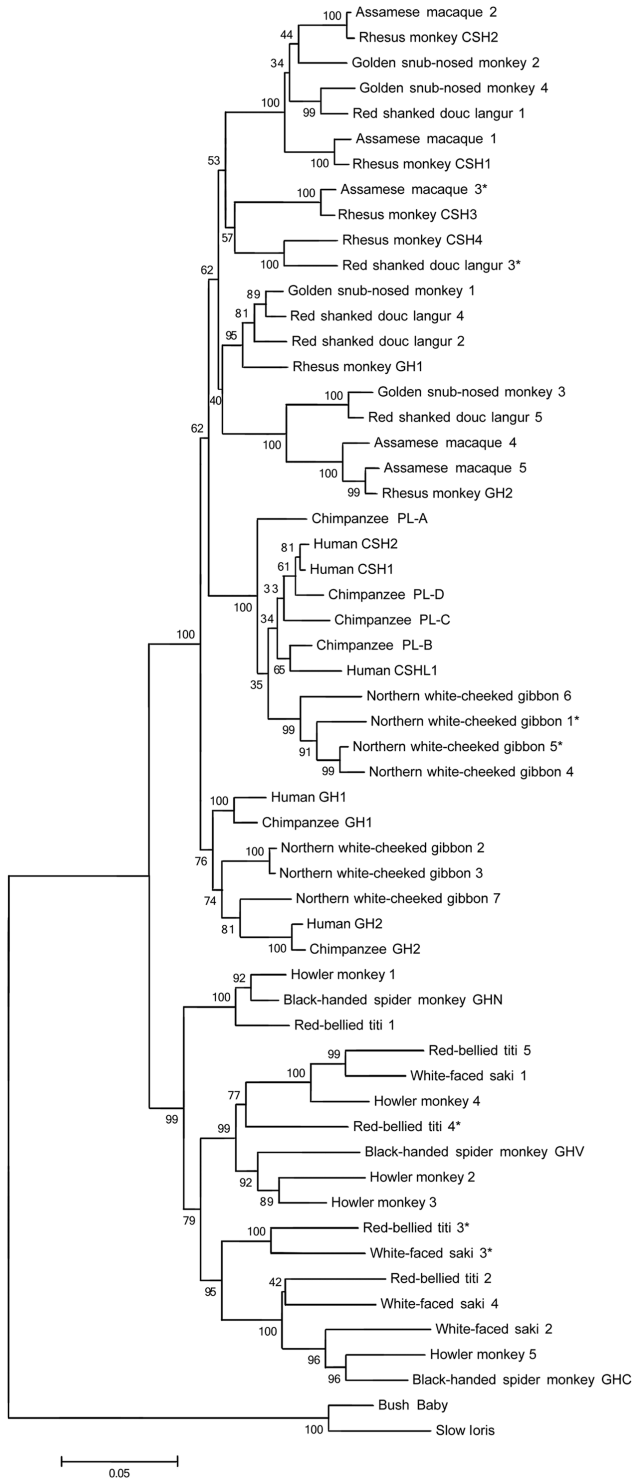
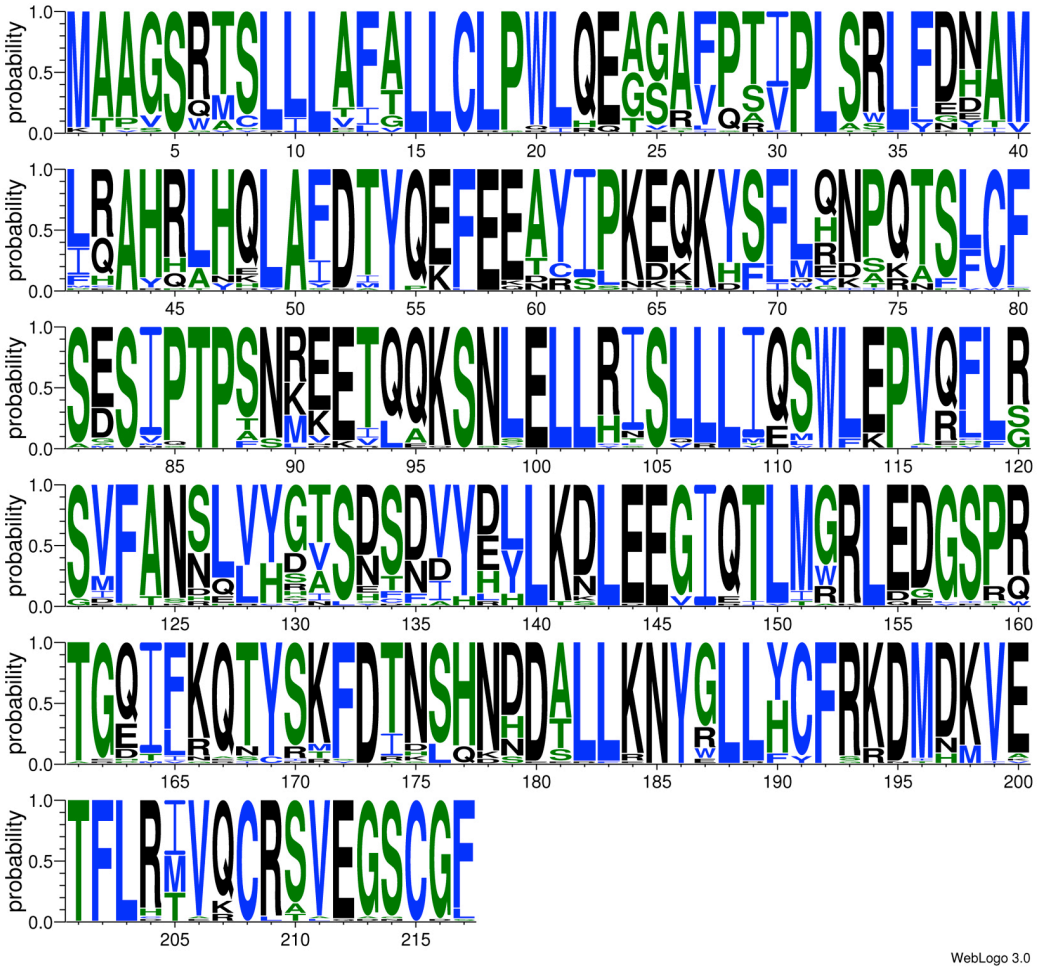

Assamese macaque 2
Rhesus monkey CSH2
Golden snub-nosed monkey 2
Golden snub-nosed monkey 4
Red shanked douc langur 1
Assamese macaque 1
Rhesus monkey CSH1
Assamese macaque 3*
Rhesus monkey CSH3
Rhesus monkey CSH4
Red shanked douc langur 3*
Golden snub-nosed monkey 1
Red shanked douc langur 4
Red shanked douc langur 2
Rhesus monkey GH1
Golden snub-nosed monkey 3
Red shanked douc langur 5
Assamese macaque 4
Assamese macaque 5
Rhesus monkey GH2
Chimpanzee PL-A
Human CSH2
Human CSH1
Chimpanzee PL-D
Chimpanzee PL-C
Chimpanzee PL-B
Human CSHL1
Northern white-cheeked gibbon 6
Northern white-cheeked gibbon 1*
Northern white-cheeked gibbon 5*
Northern white-cheeked gibbon 4
Human GH1
Chimpanzee GH1
Northern white-cheeked gibbon 2
Northern white-cheeked gibbon 3
Northern white-cheeked gibbon 7
Human GH2
Chimpanzee GH2
Howler monkey 1
Black-handed spider monkey GHN
Red-bellied titi 1
Red-bellied titi 5
White-faced saki 1
Howler monkey 4
Red-bellied titi 4*
Black-handed spider monkey GHV
Howler monkey 2
Howler monkey 3
Red-bellied titi 3*
White-faced saki 3*
Red-bellied titi 2
White-faced saki 4
White-faced saki 2
Howler monkey 5
Black-handed spider monkey GHC
Bush Baby
Slow loris
0.05

Figure 2.4


probability
1.0
0.5
0.0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
WebLogo 3.0

Supplementary Table 2.1.

Accession numbers of the sequences used in this study

Species	GenBank No.	Species	GenBank No.
Alouatta seniculus 1 (A.sen1)	AY744451	*Macaca mulatta GH1*	DQ002799
Alouatta seniculus 2 (A.sen2)	AY744452	*Macaca mulatta CSH1*	DQ002801
Alouatta seniculus 3 (A.sen3)	AY744453	*Macaca mulatta CSH2*	DQ002802
Alouatta seniculus 4 (A.sen4)	AY744454	*Macaca mulatta CSH3*	DQ002803
Alouatta seniculus 5 (A.sen5)	AY744455	*Macaca mulatta* CSH4	DQ002804
Hylobates leucogenys 1* (H.leu1)	AY621635	*Macaca mulatta GH2*	DQ002800
Hylobates leucogenys 2 (H.leu2)	AY621636	*Pan troglodytes GH1*	AF374232
Hylobates leucogenys 3 (H.leu3)	AY621637	*Pan troglodytes GH2*	AF374233
Hylobates leucogenys 4 (H.leu4)	AY621638	*Pan troglodytes PL-A*	AY146625
Hylobates leucogenys 5 (H.leu5)	AY621639	*Pan troglodytes PL-B*	AY146626
Hylobates leucogenys 6 (H.leu6)	AY621640	*Pan troglodytes PL-C*	AY146627
Hylobates leucogenys 7 (H.leu7)	AY621641	*Pan troglodytes PL-D*	AY146628
Pithecia pithecia 1 (P.pit1)	AY744461	Human *GH1*	EU421712
Pithecia pithecia 2 (P.pit2)	AY744462	Human *CSHL1*	EU421713
Pithecia pithecia 3* (P.pit3)	AY744463	Human *CSH1*	EU421714
Pithecia pithecia 4 (P.pit4)	AY744464	Human *GH2*	EU421715
Rhinopithecus roxellana 1 (R.rox1)	AY621647	Human *CSH2*	EU421716
Rhinopithecus roxellana 2 (R.rox2)	AY621648	*Nycticebus pygmaeus* GH (Slow Ioris)	AJ297562
Rhinopithecus roxellana 3(R.rox3)	AY621649	*Galago senegalensis* (Bush Baby)	AF292938
Rhinopithecus roxellana 4 (R.rox4)	AY621650	*Pygathrix nemaeus* 1 (P.nem1)	AY621642
*Callicebus moloch*1 (C.mol1)	AY744456	*Pygathrix nemaeus* 2 (P.nem2)	AY621643
Callicebus moloch 2 (C.mol2)	AY744457	*Pygathrix nemaeus* 3* (P.nem3*)	AY621644
Callicebus moloch 3* (C.mol3)	AY744458	*Pygathrix nemaeus* 4 (P.nem4)	AY621645
Callicebus moloch 4* (C.mol4)	AY744459	*Pygathrix nemaeus* 5 (P.nem5)	AY621646
Callicebus moloch 5 (C.mol5)	AY744460	*Ateles geoffroyi* (Spider Monkey GHN)	AF374234
Assamese macaque 1 (M.ass1)	AY621651	*Ateles geoffroyi* (Spider Monkey GHV)	AF374235
Assamese macaque 2 (M.ass2)	AY621652	*Ateles geoffroyi* (Spider Monkey GHC)	AY435434
Assamese macaque 3* (M.ass3)	AY621653		
Assamese macaque 4 (M.ass4)	AY621654		
Assamese macaque 5 (M.ass5)	AY621655		

Note. Stars (*) indicate sequences which are likely pseudogenes.

References:

R.M. Adkins, A. Nekrutenko, W.H. Li, Bushbaby growth hormone is much more similar to nonprimate growth hormone than to rhesus monkey and human growth hormones, Mol. Biol. Evol. 18 (2001) 55–60.

E. Alsat, J. Guibourdenche, D. Luton, F. Frankenne, D. Evain-Brion, Human placental growth hormone, Am. J. Obstet. Gynecol. 177 (1997) 1526–1534.

D. Benovoy, R.T.Morris, A. Morin, G. Drouin, Ectopic gene conversions increase the G+ C content of duplicated yeast and Arabidopsis genes, Mol. Biol. Evol. 22 (2005) 1865–1868.

J.M. Bischof, A.P. Chiang, T.E. Scheetz, E.M. Stone, T.L. Casavant, V.C. Sheffield, T.A. Braun, Genome-wide identification of pseudogenes capable of disease-causing gene conversion, Hum. Mutat. 27 (2006) 545–552.

E. Cacciari, P. Pirazzoli, S. Gualandi, C. Baroncini, L. Baldazzi, B. Trevisani,M. Capelli, S.Zucchini, A. Balsamo, A. Cicognani, F. Bernardi, Molecular study of human growth hormone gene cluster in three families with isolated growth hormone deficiency and similar phenotype, Eur. J. Pediatr. 153 (1994) 635–641.

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

E.Y. Chen, Y.C. Liao, D.H. Smith, H.A. Barrera-Saldaña, R.E. Gelinas, P.H. Seeburg, The human growth hormone locus: nucleotide sequence, biology, and evolution, Genomics 4 (1989) 479–497.

L.K. Derr, J.N. Strathern, D.J. Garfinkel, RNA-mediated recombination in S.cerevisiae, Cell 67 (1991) 355–364.

G. Drouin, Characterization of the gene conversions between the multigene family members of the yeast genome, J. Mol. Evol. 55 (2002) 14–23.

G. Drouin, F. Prat, M. Ell, G.D.P. Clarke, Detecting and characterizing gene conversions between multigene family members, Mol. Biol. Evol. 16 (1999) 1369–1390.

B. Elliott, C. Richardson, J.Winderbaum, J.A. Nickoloff, M. Jasin, Gene conversion tracts from double-strand break repair inmammalian cells, Mol. Cell. Biol. 18 (1998) 93–101.

G.R. Fink, Pseudogenes in yeast? Cell 49 (1987) 5–6.

M. Giordano, C. Marchetti, G.B. Chiorboli, P.M. Richiardi, Evidence for gene conversion in the generation of extensive polymorphism in the promoter of the growth hormone gene, Hum. Genet. 100 (1997) 249–255.

A.S. Goldman, M. Lichten, The efficiency of meiotic recombination between dispersed sequences in Saccharomyces cerevisiae depends upon their chromosomal location, Genetics 144 (1996) 43–55.

R. González Alvarez, A. Revol de Mendoza, D. Esquivel Escobedo, G. Corrales Félix, I. Rodríguez Sánchez, V. González, G. Dávila, Q. Cao, P. de Jong, Y.-X. Fu, H.A. Barrera Saldaña, Growth hormone locus expands and diverges after the separation of New and Old World Monkeys, Gene 380 (2006) 38–45.

D. Graur, W.-H. Li, Fundamentals of molecular evolution, 2nd ed, Sinauer Associates, Sunderland, 2000.

T.A. Hall, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, Nucleic Acids Symp. Ser. 41 (1999) 95–98.

S. Harvey, M. Kakebeeke, A.E.Murphy, E.J. Sanders, Growth hormone in the nervous system: autocrine or paracrine roles in retinal function? Can. J. Physiol. Pharmacol. 81 (2003) 371–384.

H. Hirt, J. Kimelman,M.J. Birnbaum, E.Y. Chen, P.H. Seeburg, N.L. Eberhardt, A. Barta, The human growth hormone gene locus: structure, evolution, and allelic variations, DNA 6 (1987) 59–70.

J.A. Jeffreys, C.A. May, Intense and highly localized gene conversion activity in human meiotic crossover hot spots, Nat. Genet. 36 (2004) 151–156.

M. Krawczak, N.A. Chuzhanova, D.N. Cooper, Evolution of the proximal promoter region of the mammalian growth hormone gene, Gene 237 (1999) 143–151.

Y. Li, C. Ye, P. Shi, X.J. Zou, R. Xiao, Y.Y. Gong, Y.P. Zhang, Independent origin of the growth hormone gene family in New World monkeys and Old World monkeys/hominoids, J. Mol. Endocrinol. 35 (2005) 399–409.

A. Loytynoja and N. Goldman, webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser, BMC Bioinformatics 11 (2010) 579-586.

T. Lukacsovich, A.S. Waldman, Suppression of intrachromosomal gene conversion in mammalian cells by small degrees of sequence divergence, Genetics 151 (1999) 1559–1568.

H.D. McIntyre, R. Serek, D.I. Crane, T. Veveris-Lowe, A. Parry, S. Johnson, K.C. Leung, K.K. Ho, M. Bougoussa, G. Hennen, A. Igout, F.Y. Chan, D. Cowley, A.
Cotterill, R. Barnard, Placental growth hormone (GH), GH-binding protein, and insulin-like growth factor axis in normal, growth-retarded, and diabetic pregnancies: correlations with fetal growth, J. Clin. Endocrinol. Metab. 85 (2000) 1143–1150.

D.S. Millar, M.D. Lewis, M. Horan, V. Newsway, T.E. Easter, J.W. Gregory, L. Fryklund, M. Norin, E.C. Crowne, S.J. Davies, P. Edwards, J. Kirk, K. Waldron, P.J. Smith, J.A. Phillips III, M.F. Scanlon,M. Krawczak, D.N. Cooper, A.M. Procter, Novelmutations of the growth hormone 1 (GH1) gene disclosed by modulation of the clinical selection criteria for individuals with short stature, Hum. Mutat. 21 (2003) 424–440.

A. Misra-Press, N.E. Cooke, S.A. Liebhaber, Complex alternative splicing partially inactivates the human chorionic somatomammotropin-like (hCS-L) gene, J. Biol. Chem. 269 (1994) 23220–23229.

T. Mourier, D.C. Jeffares, Eukaryotic intron loss, Science 300 (2003) 1393.

R.T. Morris, G. Drouin, Ectopic gene conversions in bacterial genomes, Genome 50 (2007) 975–984.

J.P.Noonan, J.Grimwood, J. Schmutz,M. Dickson,R.M.Myers, Gene conversion and the evolution of protocadherin gene cluster diversity, Genome Res. 14 (2004) 354–366.

D. Posada, Evaluation of methods for detecting recombination from DNA sequences: empirical data, Mol. Biol. Evol. 19 (2002) 708–717.

D. Posada, K.A. Crandall, Evaluation of methods for detecting recombination from DNA sequences: computer simulations, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 13757–13762.

S. Prager, H.A.Wollmann, S. Mergenthaler,M.Mavany, K. Eggermann, M.B. Ranke, T.Eggermann, Characterization of genomic variants in CSH1 and GH2, two candidate genes for Silver-Russell syndrome in 17q24-q25, Genet. Test. 7 (2003) 259–263.

A.M. Procter, J.A. Phillips III, D.N. Cooper, Themolecular genetics of growth hormone deficiency, Hum. Genet. 103 (1998) 255–272.

A. Revol De Mendoza, D. Esquivel Escobedo, I. Martínez Dávila, H. Barrera Saldaña, Expansion and divergence of the GH locus between spider monkey and chimpanzee, Gene 336 (2004) 185–193.

S. Sawyer, Statistical tests for detecting gene conversion, Mol. Biol. Evol. 6 (1989) 526–538.

L. Sedman, B. Padhukasahasram, P. Kelgo, M. Laan, Complex signatures of locusspecific selective pressures and gene conversion on human growth hormone/ chorionic somatomammotropin genes, Hum. Mutat. 29 (2008) 1181–1193.

P. Shen, H.V. Huang, Homologous recombination in Escherichia coli: dependence on substrate length and homology, Genetics 112 (1986) 441–457.

S.K. Shyue, L. Li, B.H. Chang, W.-H. Li, Intronic gene conversion in the evolution of human X-linked color vision genes, Mol. Biol. Evol. 11 (1994) 548-551.

W. Somers, M. Ultsch, A.M. De Vos, A.A. Kossiakoff, The X-ray structure of a growth hormone-prolactin receptor complex, Nature 372 (1994) 478–481.

M.E. Steiper, N.M. Young, Primate molecular divergence dates, Mol. Phylogenet. Evol. 41 (2006) 384–394.

K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, Mol. Biol. Evol. 24 (2007) 1596–1599.

J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.

G. Untergasser, M. Hermann, H. Rumpold, P. Berger, Complex alternative splicing of the GH-V gene in the human testis, Eur. J. Endocrinol. 139 (1998) 424–427.

B.C. Verrelli, S.A. Tishkoff, Signatures of selection and gene conversion associated with human color vision variation, Am. J. Hum. Genet. 75 (2004) 363–375.

A.S. Waldman, R.M. Liskay, Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology, Mol. Cell. Biol. 8 (1988) 5350–5357.

O.C. Wallis, M. Wallis, Characterisation of the GH gene cluster in a new-world monkey, the marmoset (Callithrix jacchus), J. Mol. Endocrinol. 29 (2002)
87–89.

O.C. Wallis, Y.P. Zhang, M. Wallis, Molecular evolution of GH in primates: characterisation of the GH genes from slow loris and marmoset defines an episode of rapid evolutionary change, J. Mol. Endocrinol. 26 (2001) 249–258.

M. Watahiki, M. Yamamoto, M. Yamakawa,M. Tanaka, K. Nakashima, Conserved and unique amino acid residues in the domains of the growth hormones. Flounder growth hormone deduced from the cDNA sequence has the minimal size in the growth hormone prolactin gene family, J. Biol. Chem. 264 (1989) 312–316.

A. Wolf, D.S. Millar, A. Caliebe, M. Horan, V. Newsway, D. Kumpf, K. Steinmann, I.S. Chee, Y.H. Lee, A. Mutirangura, G. Pepe, O. Rickards, J. Schmidtke, W. Schempp, N. Chuzhanova, H. Kehrer-Sawatzki, M. Krawczak, D.N. Cooper, A gene conversion hotspot in the human growth hormone (GH1) gene promoter, Hum. Mutat. 30 (2009) 239–247.

C. Ye, Y. Li, P. Shi, Y.P. Zhang, Molecular evolution of growth hormone gene family in old world monkeys and hominoids, Gene 350 (2005) 183–192.

Chapter 3.

Purifying selection against gene conversion in the strongly conserved FOLR1 and FOLR2 genes of the folate receptor gene family of primates

Abstract

The human folate receptor gene family is composed of three functional genes and two pseudogenes. The folate receptor gene families of five other primate species were obtained and through the use of synteny and phylogenetic analysis, human orthologs were identified. Through the use of the Hsu et al., 2010 pipeline, 27 gene conversion events having an average length (± standard deviation) of 519.4 ± 492.9 nucleotides were detected in the folate receptor genes of these studied primates. A significant correlation was found between sequence similarity and the length of conversion events. In addition, the GC-content of converted regions was significantly more GC rich than non-converted regions. Due to their physiological importance and strong conservation, FOLR2 and FOLR1 were converted the least. Contrastingly, the functional FOLR3 gene, the least conserved folate receptor gene, was found to be frequently converted by the pseudogene FOLR3P1. Conversions predominately occurring in FOLR3 and in the other folate receptor pseudogenes are correlated with the importance of the expression of the FOLR2 and FOLR1 genes suggesting that purifying selection is against conversion events occurring within these two highly conserved functional genes.

3.1. Introduction

The human folate receptor gene family is located at chromosome 11q13-14. Folate receptors participate in the binding and transport of folate (a variety of vitamin B9) and the naturally occurring form of folic acid (Heil et al., 1999). Folic acid derivatives have been

found to be crucial for the synthesis of DNA, cell division, tissue growth and DNA methylation (Boyles et al., 2006).

In humans, the folate receptor gene family consists of three functional genes FOLR1, FOLR2 and FOLR3 in addition to two nonfunctional pseudogenes FOLR1P1 and FOLR3P1 (Finell et al., 1998, Sadasivan et al., 1992, Elwood et al., 1997). FOLR1 has been found to be a complex gene that can produce a variety of transcripts through simple or complex alternative splicing. Coding regions are activated in a specific manner in a variety of human tissues including the human cerebellar, kidney and normal lung tissues (Elwood, 1997). FOLR2 is mainly expressed in placental tissue (Sadasivan et al., 1993) while FOLR3 is a secretory molecule expressed in the spleen, thymus, bone marrow, and predominantly in hematopoietic cells (Shen et al., 1993).

Much is known about the expression and organization of these folate receptor genes due to their potential affiliation with disease, specifically neural tube defects (NTDs). Embryonic neural tubes form the brain and spinal cord during the development of the embryo. If the neural tubes fail to close accordingly, the result would be a congenital deformation known as a neural tube defect (NTD). Though well studied, causes for NTDs still remain largely unknown. Many studies, however, highlight the little understood protective effects of folic acid supplementation during pregnancy (Finell et al., 1998, O'Byrne et al., 2010, O'Leary et al., 2003, Boyles et al., 2006). Due to this phenomenon, the folate receptor genes have been recognized as a means to better understand NTDs and their connection has been the focus of past research (Heil et al., 1999, De Marco et al., 2000, Elnakat and Ratnam, 2004).

Gene conversions are a type of homologous recombination that involve the unidirectional movement of genetic material from a donor sequence to an acceptor sequence. Conversion events have been found to be connected to a variety of human

inherited diseases, especially when mutations that accumulate in nonfunctional pseudogenes overwrite essential components of a highly similar functional gene (Chen et al., 2007).

The recent surge in freely available sequence data has facilitated the opportunity to further our understanding of the folate receptor gene family in not only humans but in five other primate species. Here, a comparative genomics study was performed on the effects of gene conversion in the folate receptor gene family of *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Callithrix jacchus* (marmoset), *Nomascus leucogenys* (gibbon), *Pongo abelii* (orangutan), and *Macaca mulatta* (rhesus monkey). In particular, we wanted to study which portions of the sequences exhibited gene conversions and we hypothesize that these regions are likely not crucial to the overall function of the gene or genes.

3.2. Materials and Methods

The folate receptor gene families from 6 primate species (*Homo sapiens*, *Pan troglodytes*, *Callithrix jacchus, Nomascus leucogenys, Pongo abelii,* and *Macaca mulatta*) were downloaded from NCBI. Human FOLR1 and FOLR1P1 sequences were used as queries in BLAST to retrieve all genes belonging to the folate receptor gene family. Genes having greater than 60% sequence identity in their coding regions were retrieved, as well as all pseudogenes in close proximity to functional genes. Reciprocal BLAST searches were performed using the resulting sequences in other species to ensure that all folate receptor gene family members of at least 60% sequence similarity were retrieved.

In order to study the effects of gene conversion in the flanking regions of genes, the complete chromosomal regions containing all folate receptor genes were downloaded using NCBI's primary reference genome assemblies of the 6 studied primate species. These downloaded regions spanned 1000 base pairs before the ATG initiation codon of the first

FOLR gene up until 1000 base pairs after the stop codon of the last FOLR gene. The complete list of NCBI Gene IDs and chromosomal ranges of the 26 folate receptor genes analyzed here are listed in Supplementary Table 3.2.

All alignments were performed using ClustalW (Thompson, 1994) and PRANK (http://www.ebi.ac.uk/goldman-srv/webPRANK/) (Loytynoja and Goldman 2010) and were manipulated using BioEdit (Hall, 1999). The phylogenetic tree was constructed with the maximum likelihood method implemented in PhyML (Guindon et al., 2010). Parameters include: the HKY85 DNA evolution model, 4 substitution rate categories and 500 replications for bootstrap analyses. The sequence similarity between all FOLR genes of each species was calculated using MEGA 4 (Tamura, 2007) and its compute pairwise distance tool (i.e., the number of base substitutions per site from between sequences) and the analyses was conducted using the maximum composite likelihood substitution model.

WebLogo (http://weblogo.berkeley.edu/) was used to assess the degree of conservation of the amino acids along the functional sequences of the FOLR1, FOLR2 and FOLR3 genes of all six primate species.

The conversion detection software used was a pipeline developed by Hsu et al., (2010). The necessary inputs for this pipeline include sequence files containing all genes and their flanking regions for each species, separate text files which identify the positions of all genes and their corresponding exons and a Newick format species tree. The fashion in which the Hsu et al., 2010 pipeline detects conversion events is by examining portions of paralogous sequences that share a higher similarity to each other compared to their corresponding orthologs (Hsu et al., 2010). Various statistical tests are performed to asses whether this higher similarity amongst paralogous genes occurs by chance or is the result of a conversion event. The Hsu et al., 2010 pipeline outperformed all other gene conversion detection software, including GENECONV, exhibiting the best sensitivity and false discovery

rate compared to all other tested conversion detection software (Song et al., 2011).

This method proved useful for this study since it was capable of detecting conversion events not only between genes but also between their flanking regions. Other gene conversion detection software, such as GENECONV, require alignments as input, making the inclusion of flanking regions problematic since they are often difficult to align. Additionally, it has been found that too few variable sites in an alignment cause GENECONV to exhibit a high false negative rate, failing to detect potential conversion events. Not only does the Hsu et al., 2010 pipeline not require any alignments as input but it also differs in that it infers the ancestral relationships between inputted sequences in order to detect conversion events. This allows for conversions to be detected even between highly similar sequences.

3.3. Results

3.3.1. Gene conversions detected with the Hsu et al., 2010 pipeline

Using the pipeline developed by Hsu et al., 2010, gene conversions were detected between the sequences of the folate receptor genes of five of the six primate species studied here. No conversion events were found in gibbon. A total of 27 conversion events spanning both exons and introns were detected with the smallest being 27 nucleotides long and the largest 1727 nucleotides long. The average size (± standard deviation) of these 27 conversions is 519.4 ± 492.9 nucleotides. Figure 3.1 demonstrates the distribution of each detected conversion event along with their approximate lengths. Additionally, a complete list of each detected conversion, their total tract lengths and their respective locations is presented in Supplementary Table 3.1.

Interestingly, one of the reported conversions was reported as occurring internally within the FOLR3P1 marmoset gene. However, upon further review of the alignment of the

marmoset genes, the region detected as a conversion in intron 3 of FOLR3P1 is not found in any other of the marmoset folate receptor genes and thus this reported conversion was ruled as a duplication event.

Converted regions are more GC-rich than non-converted regions. When considering the 20 conversions larger than 100 nucleotides long, the average GC-content (± standard error) of the converted regions is 58.61% ± 0.89 and is significantly larger than that of the non-converted regions (48.36% ± 0.5, t-test, $p= 9.18x10^{-13}$). For this analysis, we only considered conversions larger than 100 bp to minimize the effect of stochastic variation.

3.3.2. Correlation between sequence similarity and conversion tract length

A correlation test was performed in order to examine the effect of overall sequence similarity on the length of conversion events, (similarities between all converted sequences are listed in Supplementary Table 3.1). Similarities between both the coding and full genomic sequences were tested for their potential effects on conversion lengths. Coding sequences consist of the initiation codon, every exon and end with a stop codon whereas genomic sequences contain the same in addition to each gene's intron regions. Positive correlations are found when performing a Spearman correlation test for each species (coding corr = 0.61, 0.81, 0.93, 0.87, 0.61 genomic corr = 0.86, 0.77, 0.93, 0.87, 0.65 for human, chimp, orangutan, rhesus and marmoset respectively) but only the rhesus correlation was found to be significant (coding p-value= 0.005, genomic p-value=0.005). Most species specific correlations are likely not significant due to too few data points. For example, orangutan, human and chimpanzee only have three or four data points respectively. In contrast, when performing the same test on all primate species, the correlation for both coding and genomic sequences and their respective conversion lengths

is highly significant (Spearman coding corr = 0.64 p-value = $3.9x10^{-4}$, Spearman genomic corr = 0.72 p-value = $3.4x10^{-5}$).

3.3.3. Phylogenetic analyses

Phylogenetic trees are useful tools in the detection of conversion events. When conversions are large enough they can cause sequences to be grouped together, breaking them away from their respective orthologous genes (Drouin et al., 1999, Graur and Li, 2000). Figure 3.2 represents the phylogenetic tree for the complete genomic folate receptor sequences used in this study. Surprisingly, despite several conversion events within studied folate receptor sequences, most of the phylogenetic relationships amongst the six primate species remain intact. Only the marmoset displays phylogenetic evidence of gene conversion between FOLR3 and FOLR3P1 genes. This evidence is supported by our gene conversion analyses where numerous long conversions are observed between these two genes (Supplementary Table 3.1)

We determined the orthologous relationships of the non-human folate receptor genes using synteny and the phylogenetic tree presented in Figures 3.1 and 3.2 respectively. All non-human genes were then appropriately renamed with respect to their human ortholog. Four of the six primate species each have 5 folate receptor genes while the marmoset appears to be missing the FOLR1P1 gene and gibbon is missing both FOLR1P1 and FOLR3P1. Furthermore, the orangutan FOLR2 and FOLR1 genes appear to have switched positions when compared to their positions in the human genome.

3.3.4. Number of gene conversions in different species

The number of gene conversions detected in the folate receptor genes in different species somewhat varies. 8 conversions were detected between the marmoset and the rhesus monkey whereas 4 conversions are found in human and chimpanzee, 3 in the orangutan and none in gibbon. Although there is a weak correlation between the number of folate receptor genes in a species and the number of conversions observed (Spearman corr = 0.43), this correlation is not significant (p=0.39). This suggests that the number of conversions is not proportional to the number of genes present in a genome.

3.3.5. Direction of gene conversion events

The number, length and direction of the converted regions were obtained using the Hsu et al., 2010 pipeline and subsequently checked based on the patterns of nucleotide variation outside converted regions and the phylogenetic position of the relevant sequences (Supplementary Table 3.1). In the rhesus monkey genes, one of the species where the most conversions were detected, the FOLR2 gene was twice a donor and never an acceptor. In addition, the human FOLR1 and the chimpanzee FOLR3 genes both converted their respective pseudogenes (FOLR1P1 and FOLR3P1 respectively) without ever being converted themselves. The FOLR1P1 gene took part in ten conversion events. It was exclusively an acceptor in all species excluding the chimpanzee where it was a donor in a conversion spanning 43 base pairs within and 556 base pairs beyond exon 4 (total of 738 bp) of FOLR1. Likewise, the FOLR1 gene was converted an additional two times in rhesus and marmoset where each conversion was under 100 base pairs and occurred in the end of exon 3 and exon 4 respectively. Moreover, the human FOLR1P1 gene was completely converted by the human FOLR1 gene with a conversion spanning 1727 base pairs. FOLR3P1

participated in a total of sixteen conversion events across all primate species acting as a donor in seven and an acceptor in nine. Ten out of those sixteen conversions involved FOLR3. FOLR2 contained the least conversions only being converted twice and acting as a donor only four times. The average size of conversions involving functional genes as acceptors (± standard deviation) is 425.5 ± 429 base pairs whereas those involving pseudogenes as acceptors are 532.3 ± 445.3 base pairs, but these length differences are not statistically significant.

3.3.6. Distribution of gene conversions in folate receptor genes

The distribution of gene conversions varies from gene to gene within each species (Figure 3.1). Conversions detected in the FOLR3 and FOLR3P1 genes tend to be uniformly distributed along the lengths of the genes, especially in marmoset and rhesus monkey. Similarly, when FOLR1P1 is the acceptor in conversion events they occur evenly over the whole length of the gene and even span the entire length of human FOLR1P1. In contrast, conversions were scarce in FOLR2 where only the second exon was partially converted in human and the last exon was converted in marmoset. Additionally, the FOLR1 gene accepted conversions exclusively in its last exon and beyond in chimpanzee and marmoset as well as the second to last exon in rhesus. Overall, no biases in the location of conversion events can be observed since they are either evenly distributed amongst the length of the gene or occur in different regions in each primate species.

3.4 Discussion

Numerous studies have previously reported the effects of mutations in the folate receptor genes of humans, especially in regards to NTDs (Finell et al., 1998, O'Byrne et al., 2010, O'Leary et al., 2003, Boyles et al., 2006, De Marco et al., 2000). However, to our knowledge, none of these previous studies focused of the possible impact of gene conversions between these human genes. Here, we not only performed such a study, but we also further addressed the possible impact of these conversions by studying their characteristics in five other primate species: *Pan troglodytes*, *Callithrix jacchus, Nomascus leucogenys, Pongo abelii,* and *Macaca mulatta*.

The average size (± standard deviation) of the 27 conversions detected using the Hsu et al. (2010) pipeline was 519.4 ± 492.9 nucleotides. This average conversion length is greater than the mean length of 55 to 290 bp measured using sperm analysis (Jeffreys 2004). This is probably due to the fact that our studied sequences display a higher similarity than the average gene.

Although we detected a total of 27 gene conversion events in five of the six primate species we studied, (Supplementary Table 3.1) most of these conversions had little impact on the phylogeny of these sequences (Figure 3.2). This allowed us to establish the orthologous relationships of these genes using phylogenetic analyses and synteny. This was in large part due to the fact that most gene conversions occurred towards the pseudogenes of this gene family, with most of the exons of functional genes being unaffected by gene conversions. Our phylogenetic analysis of the folate receptor genes therefore produced phylogenetic trees with strong bootstrap values for sequences showing syntenic relationships.

Gene conversions have been reported to cause undesirable effects when mutations that have accumulated in pseudogenes are donated to highly similar functional genes, overwriting important regions (Chen et al., 2007). Such is the case in a previous study where evidence of gene conversion in the folate receptor gene family was hypothesized to result in NTDs (De Marco et al., 2000). Evidence was provided for the introduction of mutations originating in the FOLR1P1 pseudogene into the functional FOLR1 gene. All donated mutations affected the last exon of FOLR1 and the 3'-UTR. These results somewhat agree with our detected conversion between human FOLR1P1 and FOLR1 which was found to span this region (Figure 3.1). However, the main difference lies in the polarity of the conversion event, FOLR1 was the donor and not the acceptor. This observed polarity is expected since we used the consensus sequences from NCBI's primary reference assembly of the human genome and we can assume that our used sequence data does not contain diseased genes such as the data used in De Marco et al., (2000). In any event, combined evidence suggests that this region is a gene conversion hot spot and undesirable effects are possible when FOLR1 is the acceptor in a conversion event involving the FOLR1P1 gene containing mutations.

Interestingly, conversions occurring within the FOLR1 were also detected in chimpanzee, marmoset and rhesus monkey (Figure 3.1). The most notable of these conversions is between the chimpanzee FOLR1 and FOLR1P1 genes where the last exon and 3'UTR of the FOLR1 gene were converted. The chimpanzee FOLR1 gene was subsequently compared to the diseased human FOLR1 genes reported in De Marco et al., (2000) and none of the identified mutations within FOLR1 responsible for disease were apparent. In addition, when comparing the chimpanzee FOLR1 gene to its human ortholog, the last exons and 3'-UTR are identical, thus, the identified conversion in chimpanzee would not result in a disease since the pseudogene source does not donate the necessary mutations to cause any

defects. Similarly, in the marmoset, the FOLR3P1 gene converted a small portion of last exon of FOLR1 and no disease causing mutations where identified. The conversions detected in the rhesus FOLR1 gene reside in the second to last exon where no links to disease have been established. In brief, these aforementioned detected conversions occurring in FOLR1 have no found impacts on the function of the gene, which was expected. However, conversions towards the FOLR1 gene do indeed occur in all studied primate species and mutations residing in pseudogenes have the capacity to be transferred to FOLR1 potentially causing detrimental effects.

Overall, the folate receptor gene with the least number of conversions is the FOLR2 gene. Only two conversions were detected, one spans a small portion of the second exon in human and another in the last exon of marmoset FOLR2 (Figure 3.1). When observing how conserved the protein coding amino acid sequences are across all six primate species, the lack of conversion events is not surprising (Figure 3.3). FOLR2 is predominantly expressed in the placental tissues and it is hypothesized that polymorphisms within this gene would essentially result in the death of the embryo (De Marco et al., 2000 and Sadasivan et al., 1994). Additionally, functional studies of the FOLR2 promoter, located at the 5' end of the FOLR2 gene, demonstrated that the promoter region contains a binding sequence crucial to the transcriptional regulation of the gene. For this reason, any conversions carrying polymorphisms to FOLR2, especially in the promoter region, would be subsequently removed by purifying selection (Hao et al., 2003, Elnakat and Ratnam, 2004).

In contrast, the highest number of gene conversions were detected between the FOLR3 and FOLR3P1 genes, which converted each other a total of eleven times. Such a high frequency of conversion events may be explained by the overwhelming high degree of sequence similarity shared by these two sequences. In all primate species with detected conversions, these two genes consistently shared the highest sequence similarity of 95%,

92.2%, 92.8%, 93.9% and 96.4% at the genomic level in human, chimpanzee, orangutan, rhesus and marmoset respectively. All other sequences, except for human FOLR1 and FOLR1P1 having a 91.6% similarity at the genomic level, share a similarity at the genomic level of at most 89%. Since previous research in bacterial, yeast and mammalian genes has shown that there is a correlation between sequence similarity and the frequency of gene conversions, it is not surprising that these two genes often convert one another (Shen et al., 1986, Waldman and Liskay, 1988, Drouin, 2002, Elliott et al., 1998, Lukacsovich and Waldman, 1999). Furthermore, FOLR3 and FOLR3P1 each were converted a total of eight and seven times respectively in all primate species, 11 times by each other and 4 times by other genes.

Moreover, there was no obvious bias with respect to the span of each conversion and their locations within each gene (Figure 3.1). However, many of the detected conversions did occur within the promoter region of FOLR3, which is unexpected. Upon observing the conservation of amino acids of FOLR3 across all primate species, it is evident that the FOLR3 gene is the least conserved amongst the three coding folate receptor genes (Figure 3.3). Likewise, FOLR2 and FOLR3 display a closer similarity than either does to FOLR1, suggesting that FOLR1 diverged earlier than the latter genes and due to FOLR2's known importance, FOLR3 can be considered as an extra copy of FOLR2 (Elwood et al., 1997, Elnakat and Ratnam, 2004). Therefore, the higher conservation of FOLR1 and FOLR2 in addition to the frequent conversions in FOLR3 suggests that FOLR3 has a less important role in binding folate and thus is free to accumulate conversions, whereas conversions in the other genes may result in undesirable effects. However, of the three folate receptor genes, the least has been published about the FOLR3 gene and little is know about its role in folate transport (O'Byrne et al., 2010, Elnakat and Ratnam, 2004). Additionally, FOLR3 is most frequently converted by FOLR3P1, a nonfunctional pseudogene that is free to

accumulate mutations and (as reported in this study) duplications without effect. As a result, FOLR3 has a high affinity for acquiring potentially dangerous genetic material from FOLR3P1 via frequent gene conversion and if it were a crucial folate receptor gene this would not likely happen.

The significant correlations between both coding and genomic sequence similarity (coding p-value = 3.9×10^{-4}, genomic p-value = 3.4×10^{-5}) and the lengths of detected conversions are consistent with previous studies in bacterial, yeast and mammalian genes (Shen et al., 1986, Waldman et al., 1988, Drouin 2002, Elliott et al., 1998, Lukacsovich et al., 1999). Accordingly, longer conversions occur between more similar genes, which was observed here. The largest reported conversion 1727 base pairs occurred between human FOLR1 and FOLR1P1 having a 91.6% sequence similarity at the genomic level. However, as mentioned earlier, FOLR3 and FOLR3P1 genes exhibit the highest degree of similarity in all five primate species leading to the most numerous conversion events occurring between them in addition to the most consistently large conversions detected spanning 354 to 1092 base pairs in length.

The frequent conversions not only lead to higher sequence similarity, but also to higher GC-content. The fact that the average GC-content of converted regions is significantly higher (p-value = 9.18×10^{-13}) than those of non-converted regions is consistent with the biased gene conversion hypothesis which posits that, due to biases in DNA repair mechanisms, higher gene conversion frequencies will lead to higher GC-content (Meunier and Duret, 2004, Galtier et al., 2001, Kudla et al., 2004). This observation demonstrates that gene conversions are so frequent between the folate receptor genes that they affect their GC-content.

In conclusion, our results show that gene conversions occurred between the folate receptor genes in five of the six primate species we analyzed. Gene conversions were scarce

in the functional genes FOLR1 and FOLR2 and therefore have little effect on the coding regions. This likely reflects the fact that purifying selection is selecting against extensive conversion in these genes and their important genetic regions. In turn this lack of extensive conversion signifies that the orthologous relationships remain preserved among the studied primates. The fact that the most detected conversions were between FOLR3, the least important folate receptor gene and the pseudogene, FOLR3P1, further demonstrates that conversion events are free to accumulate in this gene family but are limited to less important genes and genetic regions. If conversions were to occur in such regions, detrimental effects would likely be observed and such conversion events would subsequently be removed via the process of purifying selection.

Figure Legends:

Figure 3.1. Graphical representation of the distribution of the 27 gene conversions detected between the folate receptor genes of human, chimpanzee, orangutan, rhesus and marmoset. Black boxes represent exons and grey boxes are pseudogenes that do not contain coding regions. The vertical lines above the genes indicate the approximate tract lengths of each detected conversion event. The asterisks at the end of conversion tracts are to show that the conversion spans longer than what is represented in the figure. Figure is not drawn to scale. Numbers correspond to the numbering of the conversions in Supplementary Table 3.1.

Figure 3.2. Phylogeny of the human, chimpanzee, orangutan, gibbon, rhesus and marmoset folate receptor genes. The scale bar represents 5% sequence divergence and the numbers at the nodes are the percent bootstrap support.

Figure 3.3. LOGO representation of the degree of amino acid conservation along the a) FOLR1 b) FOLR2 and c) FOLR3 genes of human, chimpanzee, orangutan, gibbon, rhesus and marmoset. The height of each letter is proportional to the degree of conservation of each amino acid (in one letter code). Numbers below each amino acid represent their position in the respective folate receptor sequence.

Figure 3.1

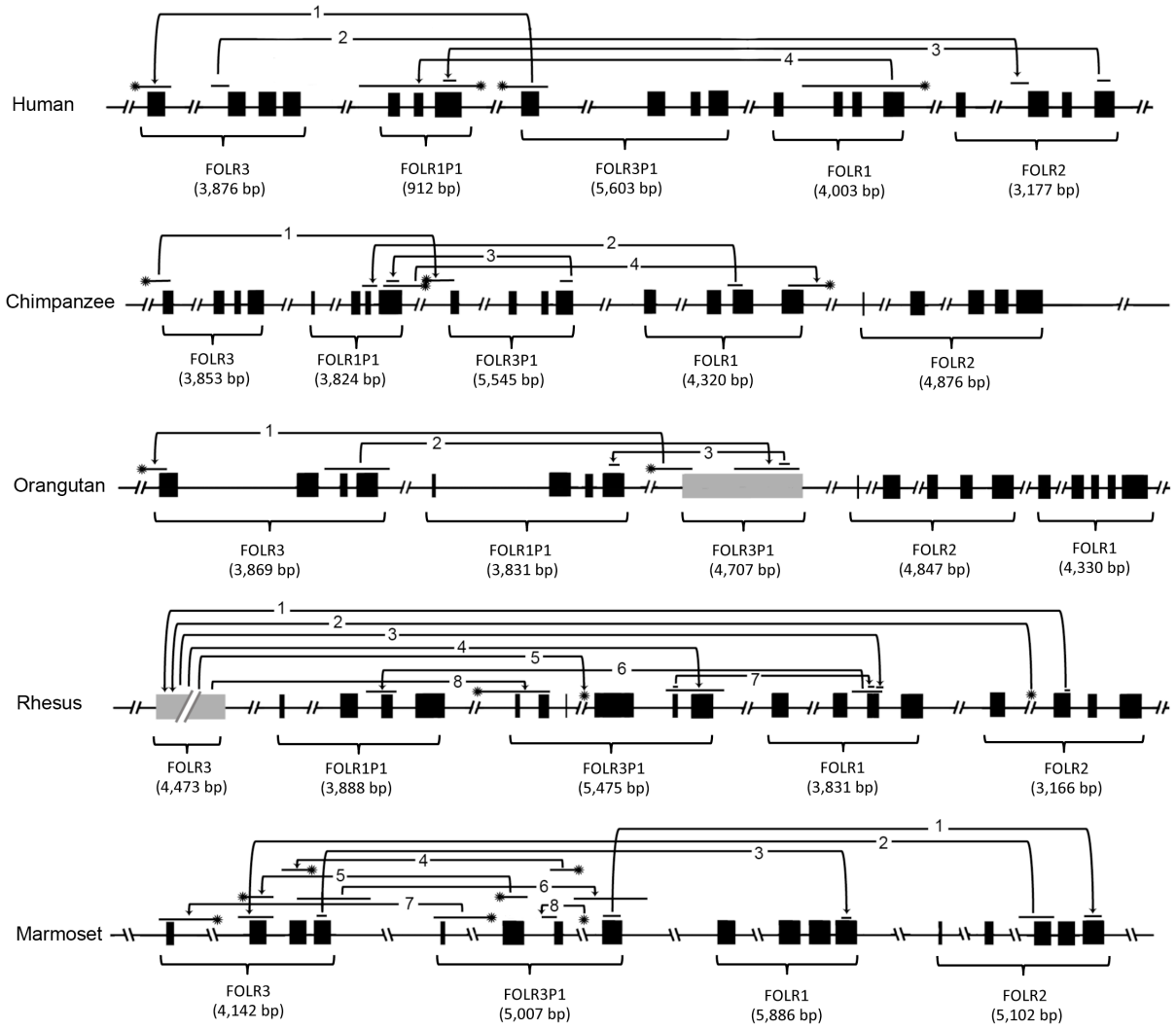

Human
FOLR3 (3,876 bp)
FOLR1P1 (912 bp)
FOLR3P1 (5,603 bp)
FOLR1 (4,003 bp)
FOLR2 (3,177 bp)
Chimpanzee
FOLR3 (3,853 bp)
FOLR1P1 (3,824 bp)
FOLR3P1 (5,545 bp)
FOLR1 (4,320 bp)
FOLR2 (4,876 bp)
Orangutan
FOLR3 (3,869 bp)
FOLR1P1 (3,831 bp)
FOLR3P1 (4,707 bp)
FOLR2 (4,847 bp)
FOLR1 (4,330 bp)
Rhesus
FOLR3 (4,473 bp)
FOLR1P1 (3,888 bp)
FOLR3P1 (5,475 bp)
FOLR1 (3,831 bp)
FOLR2 (3,166 bp)
Marmoset
FOLR3 (4,142 bp)
FOLR3P1 (5,007 bp)
FOLR1 (5,886 bp)
FOLR2 (5,102 bp)

Figure 3.2

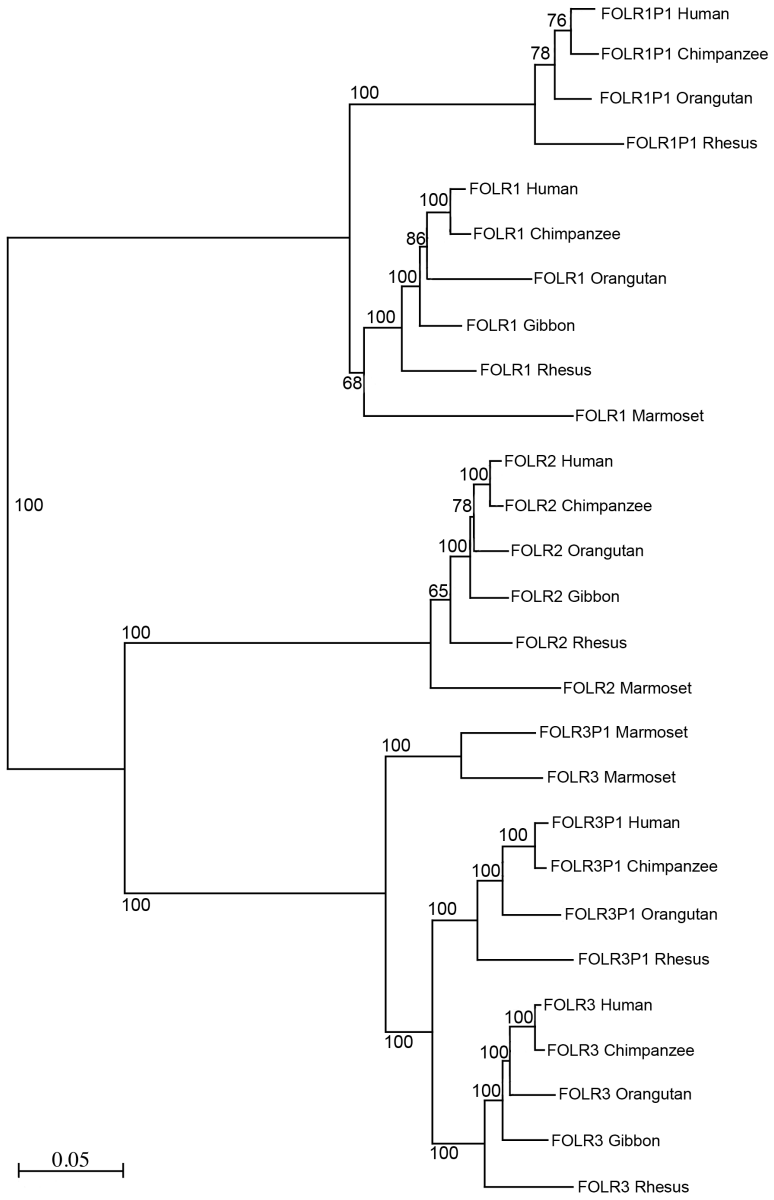
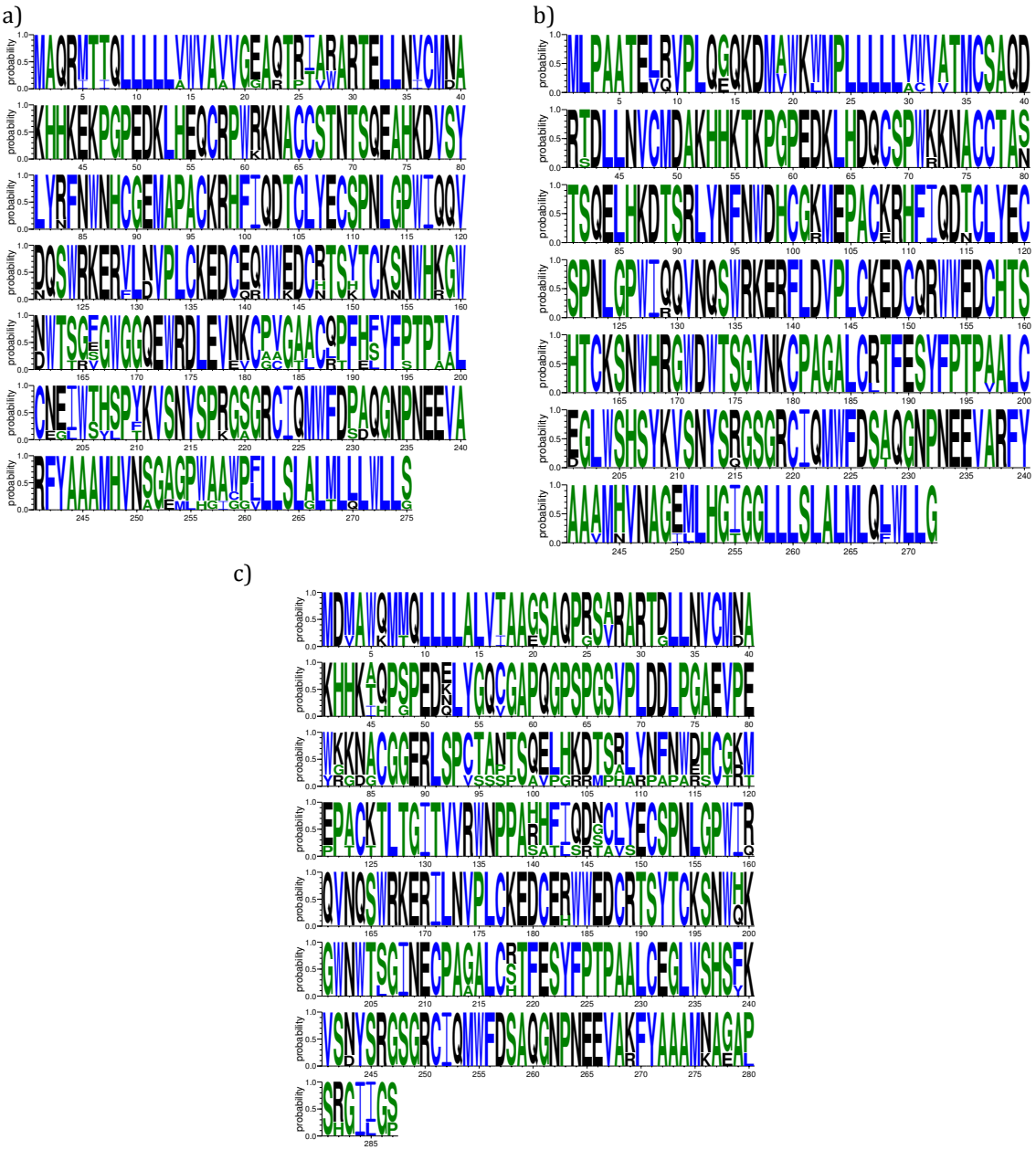

FOLR1P1 Human
FOLR1P1 Chimpanzee
FOLR1P1 Orangutan
FOLR1P1 Rhesus
FOLR1 Human
FOLR1 Chimpanzee
FOLR1 Orangutan
FOLR1 Gibbon
FOLR1 Rhesus
FOLR1 Marmoset
FOLR2 Human
FOLR2 Chimpanzee
FOLR2 Orangutan
FOLR2 Gibbon
FOLR2 Rhesus
FOLR2 Marmoset
FOLR3P1 Marmoset
FOLR3 Marmoset
FOLR3P1 Human
FOLR3P1 Chimpanzee
FOLR3P1 Orangutan
FOLR3P1 Rhesus
FOLR3 Human
FOLR3 Chimpanzee
FOLR3 Orangutan
FOLR3 Gibbon
FOLR3 Rhesus
0.05

Figure 3.3



Supplementary Table 3.1

Gene conversion events detected using Hsu et al. (2010) Pipeline

	Length	Donor	Location	Acceptor	Location	Coding % Sim	Genomic % Sim
Human							
1	1092	FOLR3P1	927bp before exon1 - bp3 intron1	FOLR3	920bp before exon1 - bp3 intron1	96.9	95
2	36	FOLR2	bp 100 - 136 of exon 4	FOLR1P1	bp 100 - 136 of exon 3	70.4	69.1
3	101	FOLR3	bp2815 intron1 - bp74 exon2	FOLR2	bp2106 intron1 - bp74 exon2	88.4	79.2
4	1727	FOLR1	bp2600 of intron1 - 540bp after exon4	FOLR1P1	307bp before exon1 - 379bp after exon3	89.6	91.6
Chimpanzee							
1	908	FOLR3	751bp before exon1 - bp155 exon1	FOLR3P1	768bp before exon1 - bp155 exon1	95.5	92.2
2	274	FOLR1	bp102 intron2 - bp222 exon3	FOLR1P1	bp100 intron2 - bp87 intron3	90.7	89
3	78	FOLR3P1	bp111 - 189 exon4	FOLR1P1	bp105- 183 exon4	71.7	57.8
4	738	FOLR1P1	bp44 exon4 - 502bp after exon4	FOLR1	bp43 exon4 - 556bp after exon4	90.7	89
Orangutan							
1	1116	FOLR3P1	992bp before - bp155	FOLR3	959bp before exon1 - bp155 exon1	93.5	92.8
2	719	FOLR3	bp87 intron2 - 123bp after exon4	FOLR3P1	bp4092 - 4829	93.5	92.8
3	49	FOLR3P1	bp4545 - 4599	FOLR1P1	bp101 - 149 exon4	63.3	59.3
Rhesus							
1	93	FOLR2	bp91 - 184 exon2	FOLR3P1	bp3631 - 3724	87.2	65.5
2	165	FOLR2	bp2080 - 2267 intron1	FOLR3	bp3516 - 3682	87.2	65.5
3	76	FOLR3	bp3991 - 4066	FOLR1	bp65 exon3 - bp5 intron3	77.5	58.2
4	740	FOLR3	bp3771 - 4510	FOLR3P1	bp372 intron4 - 99bp after exon6	97.9	93.9
5	1275	FOLR3	bp459 - 1734	FOLR3P1	bp527 - 1797 intron3	97.9	93.9
6	296	FOLR1	bp39 intron2 - bp44 intron3	FOLR1P1	bp38 intron2 - bp45 intron3	87.5	86.7
7	52	FOLR1	bp64 - 116 exon3	FOLR3P1	bp64 - 116 exon5	76.2	55.1
8	1201	FOLR3	827bp before - bp372	FOLR3P1	777bp before exon1 - bp3 intron2	97.9	93.9
Marmoset							
1	124	FOLR3P1	bp44 - 168 exon4	FOLR2	bp92 - 216 exon5	87.4	79.5
2	263	FOLR2	bp20 intron3 - bp2065 intron2	FOLR3	bp18 intron2 - bp3063 intron1	89.1	82.5
3	99	FOLR3	bp57 - 156 exon4	FOLR1	bp87 - 186 exon4	75.3	69.9
4	354	FOLR3P1	bp63 intron3 - bp333 intron2	FOLR3	bp62 intron3 - bp43 intron2	97.8	96.4
5	371	FOLR3P1	bp27 intron2 - bp2938 intron1	FOLR3	bp28 intron2 - bp2938 intron1	97.8	96.4
6	657	FOLR3	230bp after exon4 - bp76 exon3	FOLR3P1	230bp after exon4 - bp512 intron3	97.8	96.4
7	1282	FOLR3P1	bp1023 intron1 - 90bp before exon1	FOLR3	bp1017 intron1 - 90bp before exon1	97.8	96.4
8*	139	FOLR3P1*	bp3965 intron2 - bp76 exon3	FOLR3P1*	bp649 - bp512 intron3	100	100

Notes. The conversion marked with a "*" , although detected as a conversion, was ruled as a duplication event.

Supplementary Table 3.2

List of sequences used, their chromosomal ranges and their NCBI GENE ID

Species	Chromosomal Range	NCBI GENE ID
Human	chr11: 71824915--71933995	
FOLR3		2352
FOLR1P1		390221
FOLR3P1		100288543
FOLR1		2348
FOLR2		2350
Chimpanzee	chr11: 70443779-70531208	
FOLR3		748291
FOLR1		451401
FOLR1P1		466701
FOLR3P1		100609115
FOLR2		100614013
Orangutan	chr11: 67366868-67521765	
FOLR3		100442083
FOLR1P1		100443179
FOLR1		100444277
FOLR3P1		100443553
FOLR2		100442445
Rhesus	chr14: 70186352-70283873	
FOLR3		718347
FOLR1P1		100429121
FOLR3P1		718378
FOLR1		718388
FOLR2		718405
Marmoset	chr11: 66628983-66703893	
FOLR2		100411452
FOLR1		100411818
FOLR3P1		100396808
FOLR3		100397175
Gibbon	chr?: 18244765-18317033	
FOLR2		100589102
FOLR1		100590223
FOLR3		100589656

Note. The chromosome for the gibbon folate receptor genes was not identified

References:

D. Benovoy, R.T.Morris, A. Morin, G. Drouin, Ectopic gene conversions increase the G+ C content of duplicated yeast and Arabidopsis genes, Mol. Biol. Evol. 22 (2005) 1865–1868.

A.L. Boyles, A.V. Billups, K.L. Deak, D.G. Siegel, L. Mehltretter, S.H. Slifer, A.G. Bassuk, J.A. Kessler, M.C. Reed, H.F. Nijhout, T.M. George, D.S. Enterline, J.R. Gilbert, M.C. Speer, Neural Tube Defects and Folate Pathway Genes: Family-Based Association Tests of Gene–Gene and Gene–Environment Interactions, Environmental Health Perspectives (2006) 1547-1552.

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

G. Drouin, Characterization of the gene conversions between the multigene family members of the yeast genome, J. Mol. Evol. 55 (2002) 14–23.

G. Drouin, F. Prat, M. Ell, G.D.P. Clarke, Detecting and characterizing gene conversions between multigene family members, Mol. Biol. Evol. 16 (1999) 1369–1390.

B. Elliott, C. Richardson, J.Winderbaum, J.A. Nickoloff, M. Jasin, Gene conversion tracts from double-strand break repair inmammalian cells, Mol. Cell. Biol. 18 (1998) 93–101.

H. Elnakat, M. Ratnam, Distribution, functionality and gene regulation of folate receptor isoforms: implications in targeted therapy, Advanced Drug Delivery Reviews 56 (2004) 1067-1084.

R. H. Finnell, K.A. Greer, R.C. Barber, J.A. Piedrahita, G.M. Shaw, E.J. Lammer, Neural Tube and Craniofacial Defects With Special Emphasis On Folate Pathway Genes, Critical Revies in Oral Biology and Medicine 9 (1998) 38-53.

N. Galtier, G. Piganeau, D. Mouchiroud, and L. Duret. GCcontent evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics (2001) 159:907–911.

D. Graur, W.-H. Li, Fundamentals of molecular evolution, 2nd edition, Sinauer Associates, Sunderland, 2000.

S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, Systematic Biology, 59 (2010) 307-21.

T.A. Hall, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, Nucleic Acids Symp. Ser. 41 (1999) 95–98.

H. Hao, H. Qi, M. Ratnam, Modulation of the folate receptor type beta gene by coordinate actions of retinoic acid receptors at activator Sp1/ets and repressor AP-1 sites, Blood 101 (2003) 4551– 4560.

S.G. Heil, N.M.J. van der Put, F.J.M. Trijbels, F.J.M. Gabree ls, H.J. Blom, Molecular genetic analysis of human folate receptors in neural tube defects, European Journal of Human Genetics 7 (1999) 393-396.

C.H. Hsu, Y. Zhang, R.C. Hardison, Nisc Comparitive Sequence Program, E.D. Green, and W. Miller, An Effective Method for Detecting Gene Conversion
Events in Whole Genomes, Journal of Computational Biology 17 (2010) 1281-1297.

J.A. Jeffreys, C.A. May, Intense and highly localized gene conversion activity in human meiotic crossover hot spots, Nat. Genet. 36 (2004) 151–156.

G. Kudla, A. Helwak, and L. Lipinski, Gene conversion and GC-content evolution in mammalian Hsp70. Mol. Biol. Evol. (2004) 21:1438–1444.

A. Loytynoja and N. Goldman, webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser, BMC Bioinformatics 11 (2010) 579-586.

T. Lukacsovich, A.S. Waldman, Suppression of intrachromosomal gene conversion in mammalian cells by small degrees of sequence divergence, Genetics 151 (1999)
1559–1568.

J. Meunier, and L. Duret, Recombination drives the evolution of GC-content in the human genome. Mol. biol. Evol. (2004) 21:984–990.

M.R. O'Byrne, K.S. Au, A.C. Morrison, J.I. Lin, J.M. Fletcher, K.K. Ostermaier, G.H. Tyerman, S. Doebel, H. Northrup, Association of folate receptor (folr1, folr2, folr3) and reduced folate carrier (slc19a1) genes with meningomyelocele, Birth Defects Research Part A: Clinical and Molecular Teratology 88 (2010) 689-694.

V.B. O'Leary, J.L. Mills, P.N. Kirke, A. Parle-McDermott, D.A. Swanson, A. Weiler, F. Pangilinan, M. Conley, A.M. Molloy, M. Lynch, C. Cox, J.M. Scott, L.C. Brodyd, Analysis of the human folate receptor b gene for an association with neural tube defects, Molecular Genetics and Metabolism 79 (2003) 129-133.

P. Shen, H.V. Huang, Homologous recombination in Escherichia coli: dependence on substrate length and homology, Genetics 112 (1986) 441–457.

G. Song, C.H. Hsu, C. Riemer, W. Miller, Evaluation of methods for detecting conversion events in gene clusters, BMC Bioinformatics 12 (2011) 1471-2105.

K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, Mol. Biol. Evol. 24 (2007) 1596–1599.

J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 605 4673–4680.

A.S. Waldman, R.M. Liskay, Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology, Mol. Cell. Biol. 8 (1988) 5350–5357.

Chapter 4.

Strong purifying selection against gene conversion in the trypsin genes of primates

Abstract

The trypsin gene family is composed of members who share a remarkable level of sequence similarity. Mutations occurring within PRSS1 (TRY1), the primary trypsin gene, have been shown to be directly affiliated with pancreatitis. Here, gene conversions occurring within the trypsin gene family in six primate species were studied. A total of 37 conversion events being, on average (± standard deviation), 1526.25 ± 1123.8 nucleotides long, were detected using the Hsu et al., 2010 pipeline and GENECONV. Such frequent gene conversion is likely the result of the high sequence similarity and similar function of the studied trypsin genes. The trypsin genes of these primate species were found to be extremely conserved at the amino acid level and many critical amino acid residues have been identified in past research (i.e. R122, N29, cysteine residues required for forming disulphide bridges and residues that determine trypsin specificity). In this study, none of the detected conversions altered any functionally important amino acid sites providing evidence for strong purifying selection against gene conversions occurring in regions containing functionally important residues in the trypsin gene families of the six primate species studied here.

4.1. Introduction

Trypsinogen, the un-activated form of trypsin, accounts for 30% of pancreatic secretory proteins in humans. Trypsin leads to the activation of all other pancreatic enzymes, including trypsin itself, playing an integral role in food digestion. (Chen et al., 2001).

In total, there are nine trypsin genes in the human genome that have been physically separated into two distinct groups, the Group 1 and the Group 2 trypsin genes. There are six genes belonging to Group 1 each sharing very high sequence similarity of approximately 91% (Chen and Ferec, 2009, Rowen et al., 1996). Five of these six Group 1 genes are located towards the 3' end of the T cell receptor locus and are arranged in tandem on chromosome 7q35. These five genes consist of the primary trypsin producing gene PRSS1, a second trypsin producing gene PRSS2, two pseudogenes TRY5 and TRY7 in addition to an expressed pseudogene TRY6 (Chen and Ferec, 2009, Chen and Ferec, 2000e). The last member of the Group 1 trypsin genes is PRSS3, a trypsin producing gene that has been translocated 15-20 million years ago from chromosome 7q35 to chromosome 9p13 (Rowen et al., 1996, Rowen et al., 2005). The Group 2 trypsin genes consist of three pseudogenes that have evolved more divergently than the members of Group 1 and are located towards the 5' end of the T cell receptor locus on chromosome 7 (Chen and Ferec, 2000e).

Due to their being multiple functional trypsin genes in the genome, no known human diseases have been linked to a deficiency of trypsin since a deficiency in one isoform is compensated by the other genes (Rowen et al., 1996, Chen and Ferec, 2000a). However, though unlikely, a simultaneous loss of all functional trypsin genes would be lethal.

In contrast, it is when an excess of trypsin is present that problems arise and can trigger an inflammation in the pancreas known as pancreatitis (Chen and Ferec, 200b, Whitcomb, 1999). Past studies demonstrate that when the arginine residue located at position 122 in the amino acid sequence of the PRSS1 gene is replaced by a histidine (R122H), a gain-of-function mutation occurs, increasing trypsin production leading to pancreatitis (Chen and Ferec, 2009, Gorry et al., 1997, Ferec et al., 1999, Witt et al., 1999, Teich et al., 2006). This R122H mutation has been found to be the leading mutation causing hereditary pancreatitis (Chen and Ferec, 2009).

This discovery provided further evidence for PRSS1 being the most important gene for the pathway of trypsinogen activation/inactivation. The measured abundance of PRSS1 is the highest of all trypsin genes, accounting for 19% of all pancreatic secretory proteins. Since this R122H gain-of-function mutation was discovered, various other mutations occurring within the PRSS1 gene have been found to be directly responsible for pancreatitis (Chen and Ferec, 2000a, Whitcomb, 1999, Teich et al., 2006, Gorry et al., 1997, Ferec et al, 1999).

The human PRSS2 gene has not been found to contribute to the trypsinogen pathway as heavily as PRSS1, accounting for only 10% of all pancreatic secretory proteins. There are also currently no found PRSS2 mutations responsible for disease (Chen et al., 1999, Whitcomb, 1996, Witt et al., 2006). The PRSS3 gene is thought to be the least important human trypsin gene, encoding a barely detectable isoform of trypsinogen only accounting for >0.5% of all secretory pancreatic proteins. It is even being proposed, from an evolutionary perspective, that PRSS3 is losing its function (Chen et al., 1999, Chen et al., 2009).

Gene conversion is a type of homologous recombination that involve the unidirectional movement of genetic material from a donor sequence to a highly similar acceptor sequence. Conversion events have been found to be connected to a variety of human inherited diseases, especially when mutations that accumulate in nonfunctional pseudogenes overwrite essential components of a highly similar functional gene (Chen et al., 2007). For these reasons, due to their high sequence similarity, the presence of three pseudogenes and the importance of PRSS1, the human Group 1 trypsin genes are ideal candidates for studies on gene conversion (Chen et al., 1999, Idris et al., 2005, Witt et al., 2006, Whitcomb, 1999).

In addition to the important arginine residue found at position 122 in the amino acid sequence of human PRSS1 is an asparagine at position 29. This amino acid is specific to human PRSS1 and a threonine is observed at this site in the PRSS1 (TRY1) sequences of all other species (Chen and Ferec, 2009, Chen and Ferec, 2000e). It was hypothesized in previous research that a gene conversion where PRSS2 converts PRSS1 was responsible for causing pancreatitis by replacing this asparagine by an isoleucine (N29I), causing a gain-of-function mutation similar to R122H (Chen and Ferec, 2009, Chen and Ferec, 2000e, Teich et al., 2005). This N29I mutation is regarded as the second most common mutation known to cause hereditary pancreatitis (Chen and Ferec, 2009). However, due to other species not containing this essential asparagine residue, one would expect that gene conversion in this region would be more frequent in other species.

The recent surge in freely available sequence data has facilitated the opportunity to further our understanding of the trypsin gene family in not only humans but in five other primate species. Here, a comparative genomics study was performed on the effects of gene conversion in the trypsin gene family of *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Callithrix jacchus* (marmoset), *Nomascus leucogenys* (gibbon), *Pongo abelii* (orangutan), and *Macaca mulatta* (rhesus monkey). In particular, we wanted to determine whether gene conversions occurred in all primate species and verify if all gene regions were equally susceptible to conversion events.

4.2. Materials and Methods:

Sequences belonging to the trypsin gene family from 6 primate species (*Homo sapiens*, *Pan troglodytes*, *Callithrix jacchus, Nomascus leucogenys, Pongo abelii,* and *Macaca mulatta*) were downloaded from NCBI. Human PRSS1 and PRSS2 sequences, located at chromosome 7q35, were used as queries in BLAST searches to retrieve sequences from all species. Genes having greater than 50% sequence identity in their coding regions were retrieved, as well as all pseudogenes in close proximity to functional genes. The human PRSS3 gene was not included in this research project since the Hsu et al., 2010 pipeline used for the detection of gene conversion events in this research can only detect conversions between genes that reside on the same chromosome. The human Group 2 trypsin genes were also not included since they are less than 50% similar to the other gene family members. Resulting sequences in other species were then used in further BLAST searches to ensure that all trypsin gene family members of at least 50% sequence similarity were retrieved.

Chromosomal regions containing all resulting genes and their flanking regions were also downloaded. These downloaded regions spanned 1000 base pairs before the ATG initiation codon of the first trypsin gene up until 1000 base pairs after the stop codon of the last gene. All genes were named according to their gene name and description found in NCBI's Gene database. The complete list of NCBI Gene IDs and chromosomal ranges of the 31 trypsin genes analyzed here are listed in Supplementary Table 4.2.

Sequences were manipulated using BioEdit (Hall, 1999). Alignments were performed using ClustalW (Thompson, 1994) and PRANK (http://www.ebi.ac.uk/goldman-srv/webPRANK/) (Loytynoja and Goldman 2010). The phylogenetic tree was constructed with the maximum likelihood method implemented in PhyML (Guindon et al., 2010). Parameters include: the HKY85 DNA evolution model, 4 substitution rate categories and

500 replications for bootstrap analyses. The sequence similarity between all trypsin genes of each species was calculated using MEGA 4 (Tamura 2007) and its compute pairwise distance tool (i.e., the number of base substitutions per site from between sequences) and the analyses was conducted using the maximum composite likelihood substitution model.

In order to detect conversion events, a pipeline developed by Hsu et al., 2010 was used. The pipeline requires a sequence file containing all genes and their flanking regions, a separate text file identifying the positions of all genes and their exons in addition to a Newick format species tree. The pipeline detects conversion events by identifying pieces of sequences that are more similar to each other than to their corresponding orthologs from other inputted species (Hsu et al., 2010). The accuracy of the Hsu et al., 2010 pipeline has been previously tested and outperformed all other tested methods when considering both sensitivity and false discovery rate (Song et al., 2011). The trypsin genes of all species were analyzed using the Hsu et al., 2010 pipeline and P-values less than 0.05 were considered significant.

In addition, due to human TRY7 and PRSS2 genes not being included in the Human primary assembly at NCBI, the GENECONV 1.81 program (http://www.math.wustl.edu/sawyer/geneconv/) implementing Sawyer's statistical method for detecting gene conversion was used to identify gene conversion events involving PRSS2 and TRY7 (Sawyer, 1989). A mismatch penalty of 2 was used to analyze these two human trypsin genes. Only the global fragments with simulated P-values of less than 0.05 were considered significant. We did not consider pairwise fragments, even when they had simulated P-values of less than 0.05, because these fragments are not corrected for multiple comparisons.

The Hsu et al., 2010 pipeline proved useful for this study since it was capable of detecting conversion events not only between genes but also between their flanking regions

where most of the detected conversions extended. Furthermore, GENECONV requires an alignment as input and its performance is highly dependent on the quality of the inputted alignment as well as the degree of similarity that is shared amongst the inputted sequences. If there are too few variable sites in the alignment, GENECONV exhibits a high false negative rate, failing to detect potential conversion events (Posada and Crandall, 2001, Posada, 2002).

GENECONV was only used to detect conversion events occurring within the human trypsin genes and detected far less conversions compared to the Hsu et al., 2010 pipeline. This can be explained by the fact that human trypsin genes share a similarity of ~91% leading to a high false negative rate in GENECONV. Not only does the Hsu et al., 2010 pipeline require no alignments as input but it also differs in that it infers the ancestral relationships between inputted sequences in order to detect conversion events. This allowed for conversions to be detected even between highly similar sequences such as the studied trypsin genes.

4.3. Results

4.3.1. Gene conversions detected with the Hsu et al., 2010 pipeline and GENECONV

Figure 4.1 shows the number and organization of trypsin genes in the five primate species with multiple trypsin genes (marmoset has a single trypsin gene).

A total of 32 conversion events spanning both exons and introns were detected with the smallest being 318 nucleotides long and the largest 4972 nucleotides long. Since human TRY7 and PRSS2 are not included in NCBI's primary genetic assembly, GENECONV was used in addition to the Hsu et al., 2010 pipeline to detect conversions between human trypsin genes. A total of four conversions were found using GENECONV with the largest being 211 nucleotides long and the shortest being 38 nucleotides long. The average size (± standard

deviation) of all 36 conversions detected by both the Hsu et al., 2010 pipeline and GENECONV is 1526.25 ± 1123.8 nucleotides. Figure 4.1 shows the distribution of each detected conversion event along with their approximate lengths. A complete list of each detected conversion, their total tract lengths and their respective locations can be found in Supplementary Table 4.1.

Interestingly, one of the reported conversions was reported as occurring internally within the TRY6(2) rhesus gene. However, upon further review of an alignment of the rhesus genes, the region detected as a conversion in intron 3 of TRY6(2) is not found in any other of the rhesus trypsin genes and thus this reported conversion was ruled as a duplication event.

The members of the trypsin gene family of chimpanzee, orangutan, gibbon and rhesus each contain numerous trypsin genes. The names of these genes were assigned based on their identified functions according to their GenBank files. Due to multiple genes in each species having similar identified functions, each repeated gene name was numbered in no particular order of importance. As a result of each species containing numerous functionally similar genes, it would be expected for the majority of conversion events to occur between such pairs. However, only four of the 36 detected conversions occur between these pairs. For example, a single conversion of 3156 base pairs in length was detected between the TRY1(1) and TRY1(2) genes in chimpanzee.

Converted regions were, on average, found to be more GC-rich than non-converted regions but this difference was not found to be statistically significant. When considering the 33 conversions larger than 100 nucleotides long, the average GC-content (± standard error) of the converted regions is 51.37% ± 0.64 and is not significantly larger than that of the non-converted regions (50.02% ± 0.56; t-test, p= 0.13). For this analysis, we only considered conversions larger than 100 bp to minimize the effect of stochastic variation.

4.3.2. Correlation between the number of genes present and sequence similarity with conversion tract length

The number of gene conversions detected in the trypsin genes of the six studied primate species varies somewhat. The rhesus monkey exhibited nineteen conversion events, the most out of all six studied species. In comparison, seven were found in human, six in chimpanzee, three in orangutan and one in gibbon. Although there is a weak correlation between the number of trypsin genes in a species and the number of conversions observed (Spearman correlation test = 0.12), this correlation is not significant (p=0.49). This suggests that the number of conversions is not proportional to the number of genes present in a genome.

In order to examine the effect of overall sequence similarity and the length of conversion events, a correlation test was performed (similarities between all converted sequences are listed in Supplementary Table 4.1). Similarities between both the coding and full genomic sequences were tested for their potential effects on conversion lengths. Coding sequences consist of the initiation codon, every exon and end with a stop codon whereas genomic sequences contain the same in addition to each gene's intron regions. When performing a Spearman correlation test, the following correlations are found for each species: coding = -0.68, 0.17, 0.17, 0.13, genomic = -0.18, 0.79, -0.35, 0.01 for human, chimp, pongo, gibbon and rhesus respectively. None of the aforementioned correlations were found to be significant. Additionally, no significant correlations were observed when performing the same test on all conversions (Spearman coding correlation = 0.13 p-value = 0.44, Spearman genomic correlation = 0.12 p-value = 0.49).

Sequence similarity between the studied primate trypsin genes seems to determine whether or not conversion events will occur. For example, in chimpanzee, the TRY3 gene shared a similarity between 53 and 54% at the genomic level with the other chimpanzee

sequences and was never involved in a conversion event. In contrast, the other chimpanzee sequence similarities never dropped below 89.8% and were all involved in conversions. Additionally, in gibbon, the TRY3 gene shares between 54 and 56% sequence similarity at the genomic level and never participates in a single conversion event. The only detected conversion event in gibbon was between the TRY1(1) and TRY1(3) genes who share the highest similarity amongst all gibbon trypsin sequences (94.4%). Moreover, in the rhesus monkey, the species with the most detected conversions, the similarity between its trypsin genes never reaches below 85.7% and all its trypsin genes included in this study were found to participate in a gene conversion event.

Furthermore, our results are similar to past studies in that the distance between trypsin genes may affect the frequency of gene conversion events (Schildkraut et al, 2005, Benovoy and Drouin, 2009). The aforementioned TRY3 genes in chimpanzee and in gibbon are located approximately half a million base pairs away from the other trypsin genes and no conversions were found to occur within these genes.

4.3.3. Phylogenetic analyses

Phylogenetic trees are useful tools in the detection of conversion events. When conversions are large enough they can cause sequences to be grouped together, breaking them away from their respective orthologous genes (Drouin et al., 1999, Graur and Li, 2000). Figure 4.2 represents the phylogenetic tree for the complete genomic trypsin sequences of all six primate species included in this study. Due to duplication events creating multiple copies of genes in select species, one would expect such gene copies to be grouped together in a phylogenetic tree based on their similarity. Moreover, we would expect to see some conservation of the expected ancestral relationships between each primate species. Here, we see that the tree contains a grouping of the rhesus monkey trypsin genes surrounded by

the remaining trypsin genes of the other primates resulting in a loss of the expected ancestral relationships. Such a topology would not be expected if no gene conversions were present. For example, the topology of this tree does not provide any evidence that the rhesus gene is orthologous to the human PRSS1. This can be explained by the very large and frequent detected conversion events (19 conversion events ranging 506 – 3228 nucleotides) causing all rhesus genes to break away from their expected phylogenetic relationships. This is can also be seen when observing the three rhesus TRY6 genes. Instead of being paired with each other, these sequences group with those whom they have participated in large conversion events with. For instance, TRY6(2) and TRY3(1) are grouped together due to a 2841 nucleotide long conversion.

Although some groups in other species most likely represent groups of orthologous genes, such as those surrounding human PRSS1, most do not. Here again, the numerous gene conversion events have obscured the relationships of the different genes. For instance, a large conversion between the chimpanzee TRY2 and TRY6(3) genes was successful in pulling away TRY6(3) from its corresponding TRY6 orthologs. This phylogenetic tree also suggests that the TRY3 genes in gibbon and chimpanzee are evolving independently from the other trypsin genes.

4.3.4. Number and direction of gene conversions in different species

The number, length and direction of the converted regions were obtained using the Hsu et al., 2010 pipeline and subsequently checked based on the patterns of nucleotide variation outside converted regions and the phylogenetic position of the relevant sequences (Supplementary Table 4.1). For those conversions found using GENECONV, the polarity could not be established for three of the four detected conversions due to the lack of sequence information (these conversions are 79 nucleotides long or shorter) resulting in

either direction being equally likely. A bias in the direction of conversion events was found in the studied trypsin genes of some species. In chimpanzee, the TRY6(3) and TRY6(1) pseudogenes each accepted a conversion once without ever converting another gene. Additionally, the human TRY6 accepted a total of three conversions without ever converting another gene. Also, no conversions were found in the human pseudogene TRY7. Conversely, in the rhesus monkey, the TRY6(2) pseudogene was exclusively a donor in two conversion events, never accepting a single conversion. Furthermore, the rhesus PRSS2(1) gene acted exclusively as a donor, only accepting one conversion in its flanking region before exon 1. Comparatively, each PRSS genes in rhesus exhibited very different characteristics with respect to conversion polarity. The PRSS1(1) gene was converted five times, the most of any gene in rhesus, whereas PRSS1(2), PRSS2(2) and PRSS2(3) were converted twice, once and twice respectively. By contrast, the PRSS1(1) gene only acted as a donor once and the PRSS1(2), PRSS2(1), PRSS2(2) and PRSS2(3) converted two, two, zero and one genes respectively. Similarly, in human, the PRSS1 gene was detected as being once a donor, converting the TRY6 pseudogene and once an acceptor, being converted by the TRY5 pseudogene.

4.3.5. Distribution of gene conversions in trypsin genes

Due to the large number of detected conversion events, most conversions are uniformly distributed along the lengths of the genes with few exceptions (Figure4.1). The primary exception is that at least one copy of either PRSS1 or TRY1 in all studied primates remains either completely unconverted (TRY1(1) and TRY1(2) in orangutan and gibbon, respectively) or exhibits conversion only in the first half of the gene (PRSS1, TRY1 and PRSS1(2) in human, chimpanzee and rhesus, respectively). A second exception is as stated

above, where no conversions were detected in the TRY3 genes of gibbon and chimpanzee as well as TRY7 in human in addition to the PRSS2(1) gene in rhesus being exclusively a donor.

4.4. Discussion

Due to their close proximity on the chromosome, having an overall sequence similarity greater than 91%, and known links to pancreatitis, previous studies have addressed the impact of gene conversions in the five Group 1 trypsin genes in human (Chen and Ferec, 2009, Rowen et al., 1996, Teich et al., 2005, Chen and Ferec, 2000c, Chen and Ferec 2000d). Our study, however, is the first to present a comprehensive analysis focusing on gene conversion events occurring within the trypsin gene families of 6 primate species (*Homo sapiens*, *Pan troglodytes*, *Callithrix jacchus, Nomascus leucogenys, Pongo abelii,* and *Macaca mulatta*).

The average size (± standard deviation) of the 36 conversions detected by both the Hsu et al., 2010 pipeline and GENECONV is 1526.25 ± 1123.8. This average conversion length is greater than the mean length of 55 to 290 bp measured using sperm analysis (Jeffreys, 2004). This is probably due to the fact that our studied sequences display a higher similarity than an average gene.

Figure 4.1 displays all 36 detected conversion events using the Hsu et al., 2010 pipeline (and GENECONV for human PRSS2) in the 6 primate trypsin gene families having more than one trypsin gene. The tract lengths of these detected conversions were found to be very long (38 to 4972 nucleotides long) often spanning into the flanking regions of the genes. The primary reason as to why so many conversion events would be allowed to occur must reside in the fact that there are numerous, very similar (in sequence similarity and in function), trypsin genes in all studied primate species. To support this, previous studies have demonstrated that deficiencies in one trypsin gene may be compensated for by

another (Rowen et al., 1996, Chen and Ferec, 2000a). Thus, if a conversion were to have detrimental effects on one gene, another trypsin gene could effectively take its place resulting in no negative effects on the overall trypsin production for the species.

One implication of these large and frequent conversion events is that it makes it pointless to try and establish the orthologous relationships between genes in different species, especially in the trypsin genes of the rhesus monkey. The orthologous relationships between the rhesus trypsin genes in the phylogenetic tree in Figure 4.2 have been completely lost due to frequent gene conversion. This agrees with the results of the Hsu et al., 2010 pipeline which detected nineteen gene conversion events in rhesus, the most of all studied primates.

Additionally, Figure 4.2 shows that the TRY3 gene of chimpanzee and gibbon seem to be evolving independently of the other trypsin genes. This gene in both chimpanzee and gibbon is located approximately half a million nucleotides away from the rest of the trypsin genes and exhibits a very low sequence similarity, each being less than 55% similar compared to all other of their trypsin genes. Furthermore, no conversions were detected in the TRY3 gene of chimpanzee and gibbon. Therefore, it is likely that the large distance between the TRY3 genes and the other trypsin genes results in a lower similarity, not allowing them to undergo gene conversion and thus causing the TRY3 genes to evolve independently from the other trypsin genes.

The primary fashion in which gene conversions can cause undesirable effects is when mutations that have accumulated in pseudogenes are donated to highly similar functional genes, overwriting important regions (Chen et al., 2007). The PRSS1 gene has been identified as the most important trypsin gene not only because of its crucial role in trypsin production but also due to mutations within it causing pancreatitis (Teich et al., 2005, Chen and Ferec, 2009, Whitcomb, 1999, Gorry et al., 1997, Witt et al., 1999).

Furthermore, past studies show that no other human trypsin genes have any associations with causing pancreatitis (Chen et al., 1999, Idris et al., 2005, Witt et al., 2006, Whitcomb, 1999, Chen and Ferec, 2009).

The R122 residue, located in the third exon of human PRSS1 and PRSS2, has been found to be the most important site of functional mammal trypsinogen genes and is very strictly conserved in all the PRSS1 genes of all vertebrate species (Chen et al., 2001, Chen and Ferec, 2009). It has been determined in previous research that the primary mutation responsible for causing hereditary pancreatitis is when this arginine is replaced by a histidine (R122H) causing a gain-of-function mutation in PRSS1 (Chen et al., 2000c, Witt et al., 1999, Ferec et al., 1999, Gory et al., 1997, Whitcomb, 1999, Chen and Ferec, 2009). Gene conversion was found to be responsible for carrying over this mutation into the PRSS1 gene in a study done by Chen et al., (2000c). The PRSS1 gene was found to be converted by either one of two pseudogenes TRY6 or TRY7, causing PRSS1 to contain the R122H mutation. Here, no conversion events were found to occur between human PRSS1 and the TRY6 and TRY7 pseudogenes likely due to the fact that non-pancreatitis patient sequences were used in this research. This arginine (R122) can be seen in grey in Figure 4.3.

In addition to the important R122 amino acid, there is an asparagine located at position 29 (N29) exclusively in the human PRSS1 sequence that is of critical importance to this primary trypsin gene (see Figure 4.3). Previous research has shown that the second most common cause for hereditary pancreatitis is when this asparagine is replaced by an isoleucine (N29I), causing a gain-of-function mutation leading to pancreatitis (Chen and Ferec, 2009, Chen and Ferec, 2000e, Teich et al., 2005). Interestingly, a conversion event was detected in our study where the pseudogene TRY5 converted PRSS1 in this region containing N29, however, no substitution for the N29 residue was observed here. Similarly, apart from the previously mentioned amino acid in addition to the residue at position 27 in

Figure 4.3, where a leucine is in the place of a valine, all amino acids in the converted region of PRSS1 are identical to those in the other functional trypsin sequences of the alignment in Figure 4.3. This provides support of purifying selection acting against gene conversion events that would result in the modification of the PRSS1 gene, especially in this region. Conversions appear to be free to accumulate as long as no critical residues (seen highlighted in Figure 4.3) of the PRSS1 gene are affected. This negative selection is further supported by a gene conversion identified in past research between PRSS1 and PRSS2 that caused the N29I substitution. The PRSS2 gene was found to convert PRSS1 in the region containing N29 and as a result caused the N29I pancreatitis inducing mutation (Chen and Ferec, 2000e, Chen and Ferec ,2000d, Teich et al., 2005). No such conversion occurring between the human PRSS1 and PRSS2 genes was detected here, however, the isoleucine responsible for the N29I mutation can clearly be seen in Figure 4.3 at the N29I highlighted position.

This type of negative selection is further supported by the lack of amino acid variation in the highlighted residues of Figure 4.3 in all of the other studied primate species. For example, in the rhesus genes where nineteen conversion events were detected, all critical residues are completely conserved except for PRSS2(1) and PRSS2(2) at the R122 position. Here, a histidine in PRSS2(1) and a tyrosine in PRSS2(2) can be observed. However, the PRSS2(1) gene only acted as a donor twice in two detected conversion events, once converting exon 1 and exon 2 of PRSS1(2) and once converting the first exon of a pseudogene TRY1. Each conversion however resulted in an identical exchange of genetic information between the two genes. In addition, the PRSS2(2) gene was not found to be the donor in any conversions. The fact that nineteen conversion events were detected in rhesus and that all critical amino acid residues remain conserved, save for the previously mentioned cases, provides strong evidence that conversions are free to occur as long as the identified essential amino acid residues are conserved.

Moreover, in chimpanzee where six conversions were detected, no variation in the highlighted amino acids of Figure 4.3 can be seen, apart from the TRY6(2) gene, where a histidine is in the place of an arginine at highlighted position R122. However, once again, similar to the rhesus PRSS2(2) gene, TRY6(2) in chimpanzee is not a donor in a single detected conversion event.

Comparatively to chimpanzee, almost all critical amino acid residues are conserved in the orangutan except for the TRY1(2) and the TRY6(2) gene. The TRY1(2) gene contains a difference in the second of the four residues that determine trypsin specificity and an arginine can be seen in the place of a glutamine. Additionally, three differences are observed in the TRY6(2) gene. One in the N29 position as well as two additional amino acid differences occurring in the first two of the six cysteine residues responsible for forming disulphide bridges. With respect to the N29 position, it can be seen in Figure 4.3 that unlike the human trypsin genes, all other primates possess a highly conserved threonine in this position. However, the orangutan TRY6(2) gene contains a valine instead of this threonine in addition to an alanine and a tryptophan in the place of the first two of six cysteine residues highlighted in green in Figure 4.3. However, similar to rhesus PRSS2(1), the orangutan TRY1(2) and TRY6(2) genes were found to be the donor in one detected conversion event each but these conversions did not involve any of the previously mentioned important regions and resulted, once again, in an identical exchange of genetic information between two genes. Furthermore, TRY6(2) was converted by TRY1(1) from its third exon until the end of the gene. Therefore, this evidence suggests that conversions involving TRY6(2) as a donor are allowed to occur as long as the aforementioned residue changes are not involved. Conversions of this type would alter important conserved amino acids and would most likely be selected against. Moreover, it is possible that conversions such as TRY1(1) converting TRY6(2) are having a conservation effect on the 3'-end of the

orangutan TRY6(2) gene, maintaining and possibly repairing its remaining essential residues.

The amino acids in the gibbon trypsin genes adhere to the same conservation as in the previously mentioned primates. The only exceptions being in the TRY3 and TRY1(2) genes. TRY3 contains amino acids that differ in the R122 position as well as the second to last column of residues that determine trypsin specificity (highlighted in blue in Figure 4.3). In the place of the R122 amino acid is a glutamine and an alanine can be observed in the second to last amino acid site determining trypsin specificity. The TRY1(2) gene contains a glycine in the fourth cysteine region responsible for forming disulphide bridges. Once again, similar to all previously mentioned genes, TRY3 and TRY1(2) were not found to be involved in any conversion events. Interestingly however, the gibbon TRY3 and TRY1(2) genes, in addition to the orangutan TRY1(2) gene, are the only sequences where an amino acid residue is found to be different in the 3'-end of the trypsin genes of all studied primates. In the alignment of Figure 4.3 it can be observed that the conservation in the highlighted amino acid sites is strongest in all studied functional trypsin genes from amino acid 200 onward, except for the differences in the aforementioned genes of gibbon and orangutan. This may be a potential reason as to why conversion events involving these genes are seldom detected or not detected at all.

Further explanation for the non-existence of detected conversions in human PRSS1 can be explained by its other conserved structures seen in pink in Figure 4.3. These pink residues are critical in humans for the interaction between PRSS1 and SPINK1 (Chen and Ferec, 2009). This interaction between these two genes is essential in the regulation of the production of trypsin and if hindered, an overproduction of trypsin can occur, leading to pancreatitis (Chen and Ferec, 2009, Teich et al, 2006, Whitcomb, 1999). In Figure 4.3 it can be seen that these regions in pink are positioned in the latter portion of the PRSS1 gene in

conjunction with other identified important amino acid residues (see Figure 4.3). The conservation of these residues further supports the hypothesis that amino acid changing conversions are selected against.

As a result of all the previously demonstrated evidence, strong purifying selection in the conversion events of the trypsin gene family must be present due to the fact that no detected conversions altered any previously identified functionally important amino acid sites (all the colored sites in Figure 4.3). Residues that do differ in these important regions were not found to be converted or to convert any other genes. Also, detected conversions involving pseudogenes (i.e. rhesus TRY6(1) converting TRY3(1)) have minimal effects on the amino acid composition of converted genes and no essential residues were found to be altered by conversions involving pseudogenes. All of this evidence strongly suggests that conversions are free to occur amongst trypsin genes as long as the identified essential amino acid residues remain conserved. Conversion events that do not adhere to this rule would most likely have detrimental effects and subsequently be removed from the population via purifying selection.

No significant correlations were found between both coding and genomic sequence similarity and the lengths of detected conversions. The reason for this lack of significance is due to most converted genes (in both coding and non-coding sequences) sharing a similarity of 90% not allowing any correlations to be determined (see Figure 4.4). However, as found in past studies in bacterial, yeast and mammalian genes, there is undoubtedly a connection between the high sequence similarity shared amongst studied trypsin genes and the large and frequent detected conversion events (Shen et al., 1986, Waldman et al., 1988, Drouin, 2002, Elliott et al., 1998, Lukacsovich et al., 1999).

Frequent conversion events not only lead to higher shared sequence similarity, but also to higher GC-content. The converted regions of the studied trypsin genes were found,

on average, to be more GC-rich than non-converted regions, which agrees with the biased gene conversion hypothesis. This hypothesis posits that, due to biases in DNA repair mechanisms, higher gene conversion frequencies will lead to higher GC-content (Meunier and Duret 2004, Galtier et al., 2001, Kudla et al., 2004). Converted regions may have been more GC-rich on average but this result was not found to be significant. This lack of significance may be due to all the trypsin genes being so similar that they probably convert one another repeatedly over their whole length. Thus, the conversions we detected likely represent the most recent conversions, not the fact that previous conversion never previously occurred in the other regions. The fact that both converted and non-converted regions have GC-contents which are higher than the genome wide GC-content of the human genome (41%) supports this interpretation (International Human Genome Sequencing Consortium, 2001).

The lack of significant correlation between the number of conversions observed in a species and the number of genes present in its genome suggest that the number of genes has no influence on the frequency of gene conversion events. This is in contrast with the situation in bacterial genomes where the number of conversions is correlated with the size of the gene families, a result which likely reflects more frequent conversion events as the number of potential conversion partner increases (Morris and Drouin, 2007). However, our conclusion assumes that all gene conversions were detected. As previously mentioned, our detected conversions most likely only represent the most recent conversions. Additionally, due to the very high sequence identities and large detected conversions, additional conversion events could have occurred inside of the long tract lengths of already detected events but are only counted as one conversion event.

In conclusion, our results show that gene conversion occurred between the trypsin genes of all studied primate species containing more than one trypsin gene. Detected

conversion events were very frequent with large tract lengths extending into the flanking regions of genes. Conversion events were found to directly affect the phylogenetic relationships of genes resulting in a complete loss of the orthologous relationships of genes, especially in the rhesus monkey. Such frequent gene conversion is likely the result of the high sequence similarity and similar function of the studied trypsin genes. Upon viewing all of the conserved sites (those marked in color in Figure 4.3) of these trypsin genes in all studied primate species it is evident that there is strong conservation amongst the trypsin genes of these studied primates at the amino acid level. This suggests that all detected conversion events most likely result in one section of a gene being replaced by an identical portion of another, thus having no consequences, allowing conversion events to freely accumulate in the trypsin gene family. If essential amino acid differences do occur, they only do so in a single copy of the species' trypsin gene where another intact gene can be found for each non-conserved one. Additionally, such genes exhibiting changes of critically conserved residues do not partake in any detected conversion events. This suggests that selection is preventing such conversions from occurring. If essential amino acids were to be replaced (i.e. N29 or R122) via gene conversion in PRSS1 (TRY1), the most important trypsin gene, the effects would be detrimental and thus removed by selection.

Figure Legends:

Figure 4.1. Graphical representation of the distribution of the 36 gene conversions detected between the trypsin genes of human, chimpanzee, orangutan, gibbon and rhesus monkeys. Black boxes represent exons and grey boxes are pseudogenes that do not contain coding regions. The vertical lines above the genes indicate the approximate tract lengths of each detected conversion event. The asterisks at the end of conversion tracts are to show that the conversion spans longer than what is represented in the figure. Genes marked with a "*" are known pseudogenes. Figure is not drawn to scale. Numbers correspond to the numbering of the conversions in Supplementary Table 4.1.

Figure 4.2. Phylogeny of the human, chimpanzee, orangutan, gibbon, rhesus and marmoset trypsin genes. The scale bar represents 5% sequence divergence and the numbers at the nodes are the percent bootstrap support. Genes marked with a "*" are known pseudogenes.

Figure 4.3. Amino acid alignment of the PRSS1 and PRSS2 (try1 and try2) trypsin genes of all studied primate species. Portions of sequences that are considered critical for trypsin function and structure and therefore conserved throughout each studied primate species are highlighted. In blue are the four residues that determine trypsin specificity. Represented in green are the six residues responsible of forming disulphide bridges. In pink are the amino acid residues of PRSS1 that interact with the SPINK1 gene allowing trypsin production to be regulated (Chen and Ferec, 2009). Human PRSS1 and PRSS2 exon boundaries are represented by the red lines and further identified below by an exon number. The strictly conserved R122 residue in addition to the R122H and N29I pancreatitis causing mutations are highlighted in gray.

Figure 4.4. Comparison between sequence similarity and conversion length between all detected conversion events. Panel a) represents the similarities shared between sequence coding regions and panel b) shows the similarities of between full genomic sequences.

Figure 4.1

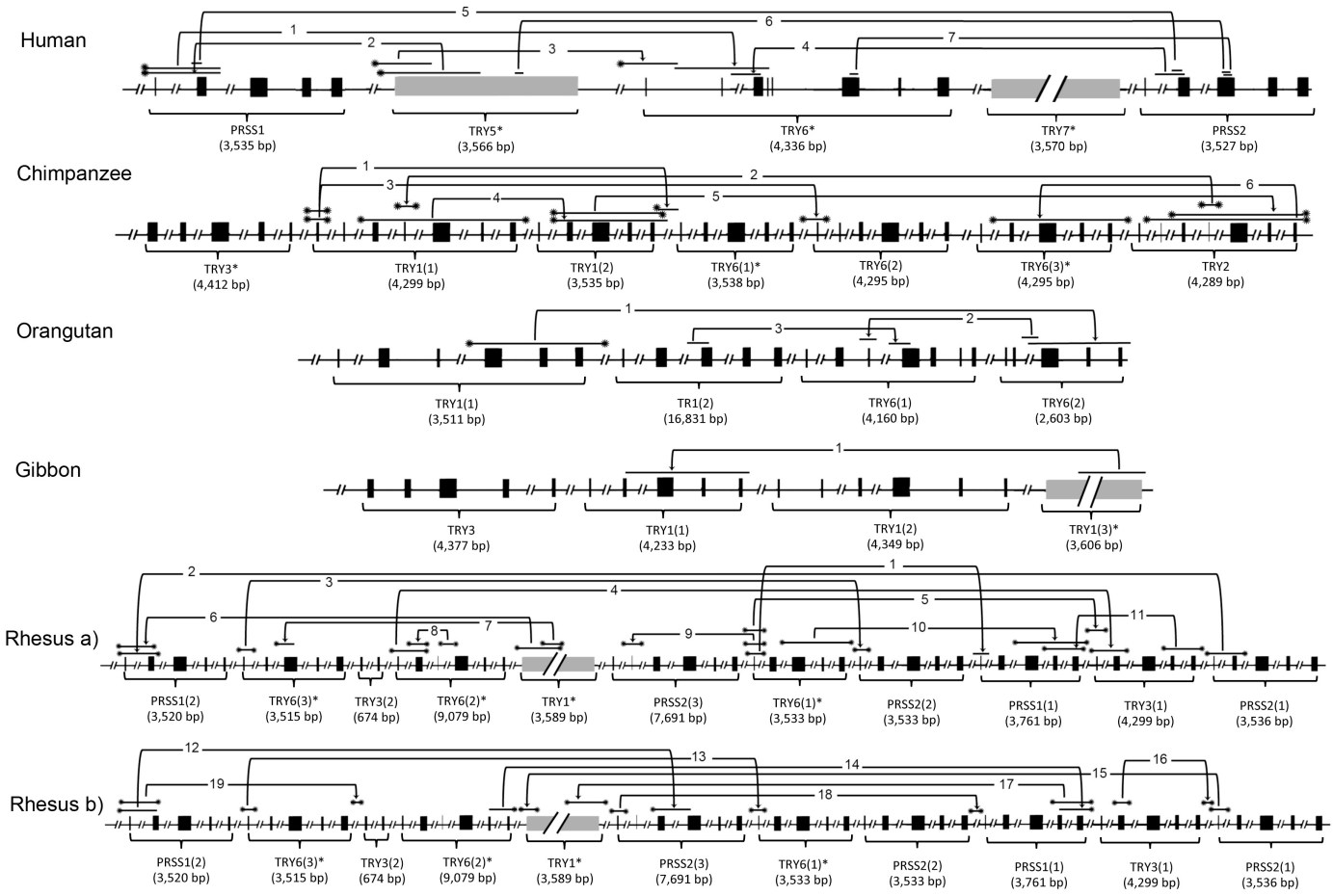

Human
PRSS1 (3,535 bp)
TRY5* (3,566 bp)
TRY6* (4,336 bp)
TRY7* (3,570 bp)
PRSS2 (3,527 bp)
Chimpanzee
TRY3* (4,412 bp)
TRY1(1) (4,299 bp)
TRY1(2) (3,535 bp)
TRY6(1)* (3,538 bp)
TRY6(2) (4,295 bp)
TRY6(3)* (4,295 bp)
TRY2 (4,289 bp)
Orangutan
TRY1(1) (3,511 bp)
TR1(2) (16,831 bp)
TRY6(1) (4,160 bp)
TRY6(2) (2,603 bp)
Gibbon
TRY3 (4,377 bp)
TRY1(1) (4,233 bp)
TRY1(2) (4,349 bp)
TRY1(3)* (3,606 bp)
Rhesus a)
PRSS1(2) (3,520 bp)
TRY6(3)* (3,515 bp)
TRY3(2) (674 bp)
TRY6(2)* (9,079 bp)
TRY1* (3,589 bp)
PRSS2(3) (7,691 bp)
TRY6(1)* (3,533 bp)
PRSS2(2) (3,533 bp)
PRSS1(1) (3,761 bp)
TRY3(1) (4,299 bp)
PRSS2(1) (3,536 bp)
Rhesus b)
PRSS1(2) (3,520 bp)
TRY6(3)* (3,515 bp)
TRY3(2) (674 bp)
TRY6(2)* (9,079 bp)
TRY1* (3,589 bp)
PRSS2(3) (7,691 bp)
TRY6(1)* (3,533 bp)
PRSS2(2) (3,533 bp)
PRSS1(1) (3,761 bp)
TRY3(1) (4,299 bp)
PRSS2(1) (3,536 bp)

Figure 4.2

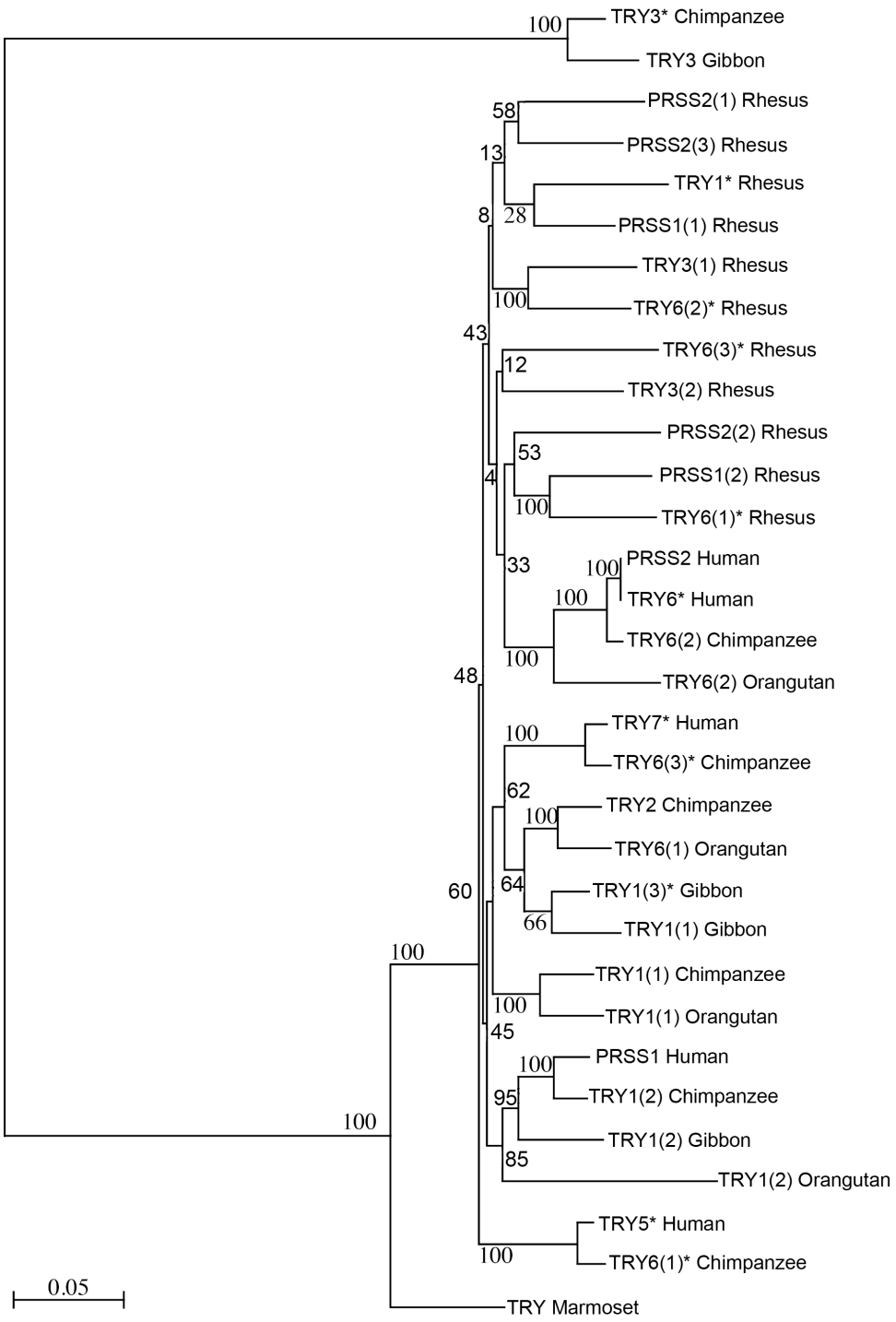

TRY3* Chimpanzee
TRY3 Gibbon
PRSS2(1) Rhesus
PRSS2(3) Rhesus
TRY1* Rhesus
PRSS1(1) Rhesus
TRY3(1) Rhesus
TRY6(2)* Rhesus
TRY6(3)* Rhesus
TRY3(2) Rhesus
PRSS2(2) Rhesus
PRSS1(2) Rhesus
TRY6(1)* Rhesus
PRSS2 Human
TRY6* Human
TRY6(2) Chimpanzee
TRY6(2) Orangutan
TRY7* Human
TRY6(3)* Chimpanzee
TRY2 Chimpanzee
TRY6(1) Orangutan
TRY1(3)* Gibbon
TRY1(1) Gibbon
TRY1(1) Chimpanzee
TRY1(1) Orangutan
PRSS1 Human
TRY1(2) Chimpanzee
TRY1(2) Gibbon
TRY1(2) Orangutan
TRY5* Human
TRY6(1)* Chimpanzee
TRY Marmoset
0.05

Figure 4.3

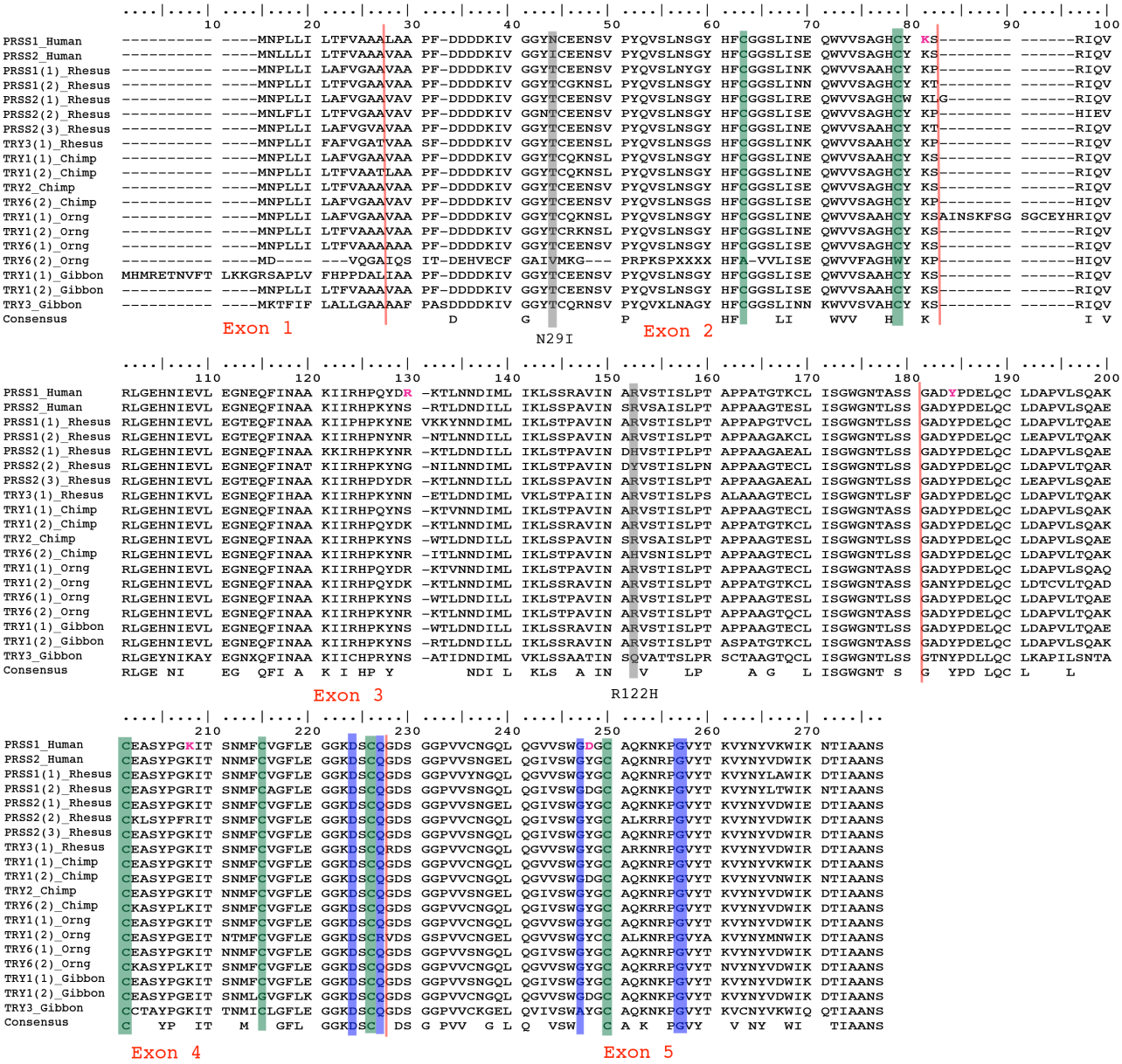
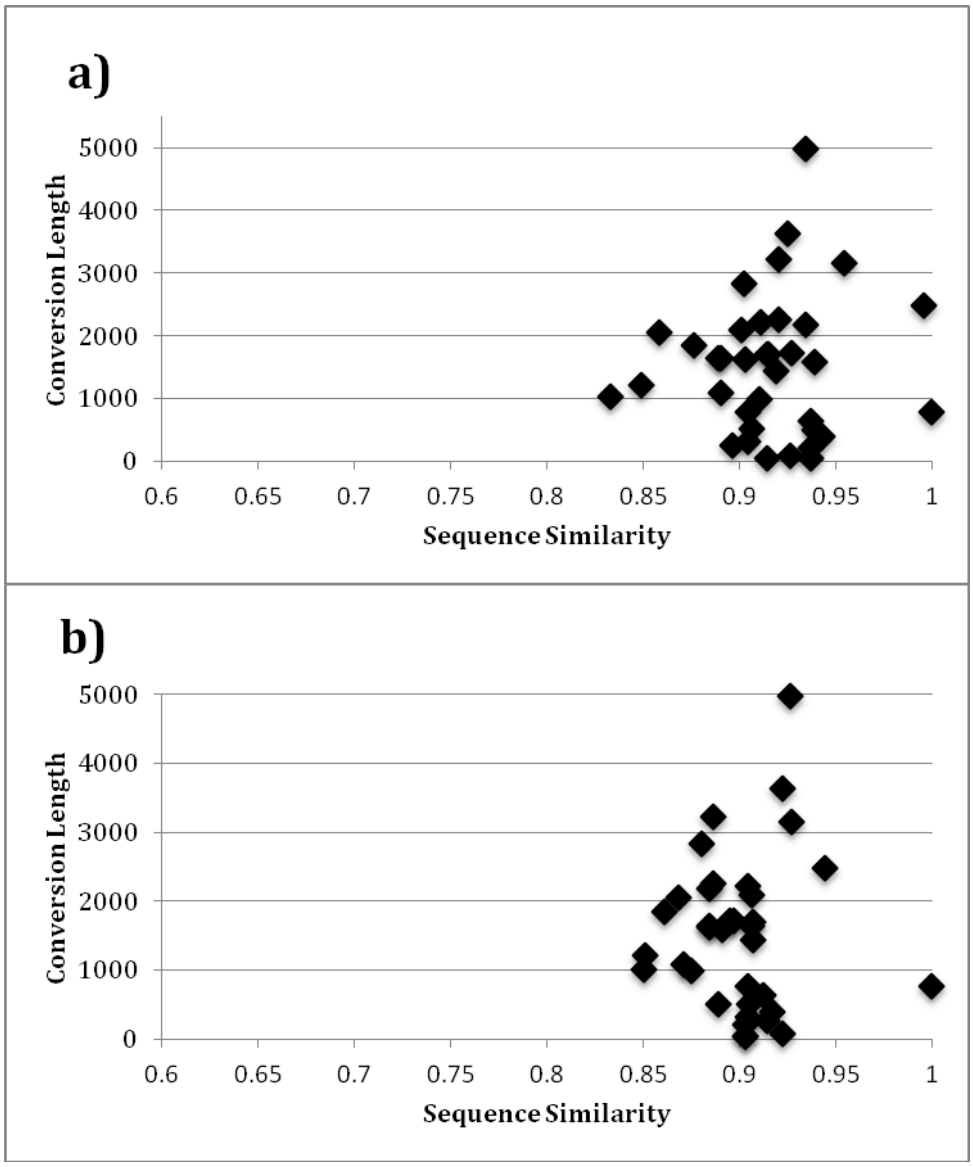


Figure 4.4


a)
Conversion Length
5000
4000
3000
2000
1000
0
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Sequence Similarity
b)
Conversion Length
5000
4000
3000
2000
1000
0
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Sequence Similarity

Supplementary Table 4.1

Gene conversion events detected using Hsu et al. (2010) Pipeline and GENECONV

	Length	Donor	Location	Acceptor	Location	Coding % Sim	Genomic % Sim
Human							
1	1704	PRSS1	198bp before exon1 - bp275 intron2	TRY6*	bp496 intron1 - bp77 intron4	91.4	90.7
2	1644	TRY5*	177bp before exon1 - bp236 intron2	PRSS1 (A)	177bp before exon1 - bp236 intron2	88.9	90.6
3	2222	TRY5*	1512bp before exon1 - bp668 intron1	TRY6*	738bp before exon1 - bp678 intron2	91.1	90.4
4 GCNV	211	TRY6* (A)	bp977 intron2 – bp157 exon3	PRSS2 (D)	bp970 intron1 - bp151 exon2	93.7	90.3
5 GCNV	79	PRSS1 (?)	bp997 intron1 - bp44 exon2	PRSS2 (?)	bp989 intron1 - bp38 exon2	92.6	92.2
6 GCNV	38	PRSS2 (?)	bp98 exon3 - bp135 exon3	TRY5* (?)	bp103 exon3 - bp140 exon3	91.4	90.3
7 GCNV	42	PRSS2 (?)	bp167 exon3 - bp208 exon3	TRY6* (?)	bp173 exon6 - bp214 exon6	93.7	90.3
Chimpanzee							
1	776	TRY1(1)	517bp before exon1 - bp176 intron1	TRY6(1)*	1286 - 515bp before exon1	90.4	90.4
2	397	TRY2	bp278 intron3 - bp279 intron4	TRY1(1)	bp280 intron3 - bp279 intron4	94.3	91.7
3	639	TRY1(1)	451bp before exon1 - bp105 intron1	TRY6(2)	457bp before exon1 - bp106 intron1	93.7	91.2
4	3156	TRY1(1)	bp817 intron2 - 487bp after exon7	TRY1(2)	bp824 intron1 - 490bp after exon5	95.4	92.7
5	3631	TRY1(2)	bp842 intron1 - 977bp after exon5	TRY2	bp838 intron2 - 960bp after exon7	92.5	92.2
6	4972	TRY2	bp283 intron1 - 1037bp after exon7	TRY6(3)*	bp283 intron1 - 1041bp after exon5	93.4	92.6
Orangutan							
1	2099	TRY1(1)	bp292 intron3 - 496bp after exon6	TRY6(2)	bp696 intron2 - 499bp after exon5	90.1	90.6
2	260	TRY6(2)	bp242 - 505 intron2	TRY6(1)	bp241 intron2 - bp108 intron3	89.6	91.4
3	318	TRY1(2)	bp803 intron2 - bp63 exon3	TRY6(1)	bp894 intron3 - bp63 exon4	90.4	90.4
Gibbon							
1	2486	TRY1(3)*	bp1151 - 44bp after	TRY1(1)	bp50 exon2 - 101bp after exon5	99.6	94.4
Rhesus (a)							
1	2265	TRY6(1)**	1223bp before exon1 - bp1001 intron1	PRSS1(1)	1195bp before exon1 - bp998 intron1	92.0	88.6
2	2176	PRSS2(1)	269bp before exon1 - bp684 intron2	PRSS1(2)	269bp before exon1 - bp681 intron2	93.4	88.4
3	2062	TRY6(3)*	1030bp before exon1 - bp991 intron1	PRSS2(2)	1030bp before exon1 - bp1009 intron1	85.8	86.8
4	2841	TRY6(2)*	803bp before exon1 - bp39 intron2	TRY3(1)	806bp before exon1 - bp39 intron2	90.2	88.0
5	1094	TRY6(1)*	873bp before exon1 - bp180 intron1	TRY3(1)	95bp before exon1 - bp918 intron1	89.0	87.1
6	1629	TRY1*	546bp before - bp1074	PRSS1(2)	558bp before exon1 - bp4 exon2	90.3	88.4
7	1020	TRY1*	bp1471 - 2490	TRY6(3)*	bp243 intron2 - bp202 exon3	83.3	85.0
8 !	776	TRY6(2)*	bp673 - 1451 intron3	TRY6(2)*	bp1692 intron1 - bp551 intron2	100	100
9	516	TRY6(1)*	215bp before exon1 - bp261 intron1	PRSS2(2)	bp4028 intron1 - bp260 intron2	90.6	88.9
10	3228	TRY6(1)*	bp471 intron2 - 1404bp after exon5	PRSS1(1)	bp710 intron2 - 1433bp after exon5	92.0	88.6
11	1709	TRY3(1)	bp156 intron3 - 873bp after exon5	PRSS1(1)	bp155 intron3 -874bp after exon5	91.5	89.7
Rhesus (b)							
12	1448	PRSS1(2)	272bp before exon1 - bp109 exon2	PRSS2(3)	bp3970 intron1 - bp109 exon3	91.9	90.7
13	1207	TRY6(3)*	198bp before exon1 - bp968 intron1	TRY6(1)*	198bp before exon1 - bp988 intron1	84.9	85.1
14	1582	TRY6(2)*	bp45 exon5 - 1040bp after exon6	PRSS1(1)	bp45 exon4 - 1023bp after exon5	93.9	89.1
15	1856	PRSS2(1)	1064bp before exon1 - bp782 intron1	TRY1*	1031bp before - bp823	87.6	86.1
16	998	TRY3(1)	bp1606 intron1 - bp683 intron2	PRSS2(1)	9623bp 8625 before exon1	91	87.5
17	1635	PRSS1(1)	bp274 intron3 - 925bp after exon5	TRY1*	bp2817 - 863bp after	89	88.4
18	506	PRSS2(3)	80bp before exon1 - bp343 intron1	PRSS1(1)	843bp before exon1 - 342bp before exon1	93.9	90.5
19	1723	PRSS1(2)	547bp before exon1 - bp109 exon2	TRY3(2)	2977 - 1264bp before exon1	92.7	89.5

Notes. Genes marked with a "*" are pseudogenes. The conversions labeled "GCNV" were detected using GENECONV. Rhesus (a) and (b) correspond to the respective figure label. Conversion 8 in Rhesus (a) marked with a "!" although detected as a conversion, was ruled as a duplication event.

Supplementary Table 4.2
List of sequences used, their chromosomal ranges and their NCBI GENE ID

Species	Chromosomal Range	NCBI GENE ID
Human	chr7: 142456319-142483417	
PRSS1		5644
PRSS2 (alternate assembly)		5645
TRY5*		168330
TRY6*		154754
TRY7*		207148
Chimpanzee	chr7: 143752943-144321250	
TRY3*		100616521
TRY1(1)		100615987
TRY1(2)		742453
TRY6(1)*		742403
TRY6(2)		100615846
TRY6(3)*		100608379
TRY2		100615529
Orangutan	chr7: 139631560-140284025	
TRY1(1)		100443026
TRY1(2)		100444247
TRY6(1)		100443647
TRY6(2)		100452346
Rhesus	chr3: 180289143-180419703	
PRSS2(1)		699238
TRY3(1)		716882
PRSS1(1)		698983
PRSS2(2)		698612
TRY6(1)*		100428303
PRSS2(3)		698352
TRY1*		698859
TRY6(2)*		699109
TRY3(2)		100429044
TRY6(3)*		699357
PRSS1(2)		698729
Marmoset	chr8: 103403794-103407296	
TRY		100395240
Gibbon	chr?: 2597078-3108172	
TRY3		100591545
TRY1(1)		100592228
TRY1(2)		100592562
TRY1(3)*		100591887

Note. The chromosome for the gibbon trypsin genes was not identified

References:

D. Benovoy, G. Drouin, Ectopic gene conversions in the human genome, Genomics 93 (2009) 27-32.

J.M. Chen, M.P. Audrezet, Mercier B, C. Quere, C. Ferec, Exclusion of anionic trypsinogen and mesotrypsinogen involvement in hereditary pancreatitis without cationic trypsinogen gene mutations, Scand. J. Gastroenterol (1999) 34:831–32.

J.M. Chen, C. Ferec, Genes, cloned cDNAs, and proteins of human trypsinogens and pancreatitis-associated cationic trypsinogen mutations. Pancreas (2000a) 21:57–62.

J.M. Chen, C. Ferec, Molecular basis of hereditary pancreatitis. Eur J Hum Genet (2000b) 8:473–479.

J.M. Chen, C. Ferec, Gene conversion-like missense mutations in the human cationic trypsinogen gene and insights into the molecular evolution of the human trypsinogen family. Mol Genet Metab (2000d) 71:463–469.

J.M. Chen, C. Ferec, Chronic Pancreatitis: Genetics and Pathgenesis, Annual Review Genomics Human Genetics (2009) 10:63-87.

J.M. Chen, C. Ferec, Orgin and implication of the hereditary pancreatitis-associated N21I mutation in the cationic trypsinogen gene, Human Genetics (2000e) 106:125-126.

J.M. Chen, O. Raguenes, C. Ferec, P.H. Deprez, C. Verellen-Dumoulin A CGC>CAT gene conversion-like event resulting in the R122H mutation in the cationic trypsinogen gene and its implication in the genotyping of pancreatitis. J Med Genet (2000c) 37:E36.

C. Ferec, O. Raguenes, R. Salomon, C. Roche, J.P. Bernard, M. Guillot, I. Quere, C. Faure, B. Mercier, M.P. Audrezet, P.J. Guillausseau, C. Dupont, A. Munnich, J.D. Bignon, L. Le Bodic, Mutations in the cationic trypsinogen gene and evidence for genetic heterogeneity in hereditary pancreatitis. J Med Genet (1999) 36: 228–232.

N. Galtier, G. Piganeau, D. Mouchiroud, and L. Duret. GCcontent evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics (2001) 159:907–911.

M.C. Gorry, D. Gabbaizedeh, W. Furey, L.K. Gates Jr, R.A. Preston, C.E. Aston, Y. Zhang, C. Ulrich, G.D. Ehrlich, D.C. Whitcomb, Mutations in the cationic trypsinogen gene are associated with recurrent acute and chronic pancreatitis. Gastroenterology (1997) 113:1063–1068.

T.A. Hall, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, Nucleic Acids Symp. Ser. 41 (1999) 95–98.

C.H. Hsu, Y. Zhang, R.C. Hardison, Nisc Comparitive Sequence Program, E.D. Green, and W. Miller, An Effective Method for Detecting Gene Conversion Events in Whole Genomes, Journal of Computational Biology 17 (2010) 1281-1297.

M.M. Idris, S. Bhaskar, D.N. Reddy, K.R. Mani, G.V. Rao, Mutations in anionic trypsinogen gene are not associated with tropical calcific pancreatitis. Gut (2005) 54:728–29

International Human Genome Sequencing Consortium, E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, et al., Initial sequencing and analysis of the human genome, Nature (2001) 409:860-921.

J.A. Jeffreys, C.A. May, Intense and highly localized gene conversion activity in human meiotic crossover hot spots, Nat. Genet. 36 (2004) 151–156.

G. Kudla, A. Helwak, and L. Lipinski, Gene conversion and GC-content evolution in mammalian Hsp70. Mol. Biol. Evol. (2004) 21:1438–1444.

A. Loytynoja and N. Goldman, webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser, BMC Bioinformatics 11 (2010) 579-586.

J. Meunier, and L. Duret, Recombination drives the evolution of GC-content in the human genome. Mol. biol. Evol. (2004) 21:984–990.

R.T. Morris, G. Drouin, Ectopic gene conversions in bacterial genomes, Genome 50 (2007) 975–984.

D. Posada, Evaluation of methods for detecting recombination from DNA sequences: empirical data, Mol. Biol. Evol. 19 (2002) 708–717.

D. Posada, K.A. Crandall, Evaluation of methods for detecting recombination from DNA sequences: computer simulations, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 13757–13762.

L Rowen, B.F. Koop, L. Hood, The complete 685-kilobase DNA sequence of the human β T cell receptor locus. Science (1996) 272:1755–1762.

L. Rowen, E. Williams, G. Glusman, E. Linardopoulou, C. Friedman, Interchromosomal segmental duplications explain the unusual structure of PRSS3, the gene for an inhibitor-resistant trypsinogen. Mol. Biol. Evol. (2005) 22:1712–20.

S. Sawyer, Statistical tests for detecting gene conversion, Mol. Biol. Evol. 6 (1989) 526–538.

G. Song, C.H. Hsu, C. Riemer, W. Miller, Evaluation of methods for detecting conversion events in gene clusters, BMC Bioinformatics 12 (2011) 1471-2105.

E. Schildkraut, C.A. Miller, J.A. Nickoloff, Gene conversion and deletion frequencies during double strand break repair in human cells are controlled by the distance between direct repeats. Nucleic Acids Res. 33 (2005) 1574–1580.

K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, Mol. Biol. Evol. 24 (2007) 1596–1599.

N. Teich, Z. Nemoda, H. Kohler, W. Heinritz, J. Mossner, V. Keim, M. Sahin-Toth, Gene conversion between functional trypsinogen genes PRSS1 and PRSS2 associated with chronic pancreatitis in a six year-old girl, Human Mutation (2005) 25:343-347.

N. Teich, J. Rosendahl, M. Tóth, J. Mössner, M. Sahin-Tóth, Mutations of human cationic trypsinogen (PRSS1) and chronic pancreatitis, Human Mutation (2006) 27:721-730.

J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 605 4673–4680.

DC. .Whitcomb, Hereditary pancreatitis: new insights into acute and chronic pancreatitis. Gut (1999) 45:317–322.

H. Witt, W. Luck, M. Becker, A signal peptide cleavage site mutation in the cationic trypsinogen gene is strongly associated with chronic pancreatitis. Gastroenterology (1999) 117:7–10.

H. Witt, M. Sahin-Toth, O. Landt, J.M. Chen, T. Kahne, A degradation-sensitive anionic trypsinogen (*PRSS2*) variant protects against chronic pancreatitis, Nat. Genet. (2006) 38:668–73.

Chapter 5

General Conclusion

In this research project, various gene conversion events occurring within the growth hormone, folate receptor and trypsin gene families of primates were characterized. Although various gene families were studied, some common themes and characteristics about the various conversion events occurring within them were observed. A general summary of the major conclusions and results can be seen in tables 5.1 and 5.2.

Firstly, gene conversion events were detected in most of the included primate species whose gene families contained multiple gene members. In the growth hormone gene family, a total of 48 conversion events were detected in eleven primates containing multiple growth hormone genes and their average size (±standard deviation) was 197.8±230.4 nucleotides. In the folate receptor gene family a total of 27 conversion events were detected in five of the six primate species having an average length (±standard deviation) of 519.4 ± 492.9. Lastly, in the trypsin gene family, a total of 37 conversion events being, on average (± standard deviations), 1526.25 ± 1123.8 nucleotides long, were detected in five of the six studied primates. It can be observed that the average detected conversion lengths increase with each study. One potential reason for this increase may be the gene conversion detection method used. In the growth hormone study, GENECONV was used, whereas in the other two studies, the Hsu et al., 2010 pipeline was implemented. When sequences contain a high degree of sequence similarity, such as the growth hormone gene family, GENECONV is known to exhibit a high false

negative rate (Posada and Crandall, 2002, Posada, 2001). The Hsu et al., 2010 pipeline on the other hand circumvents this issue by detecting conversion events by inferring the ancestral relationships between inputted sequences, making comparisons amongst them and examining if portions of sequences within species are more similar to each other than to their respective ancestral counterpart. Not only was this method of gene conversion detection found to demonstrate the highest sensitivity of all other techniques but it also allows for conversions to be detected in flanking regions of genes, further increasing the detection area (Song et al., 2011). Therefore, for these reasons, studies implementing the Hsu et al., 2010 pipeline report larger detected conversion events. Unfortunately, the Hsu et al., 2010 pipeline requires a specific form of input that was not available for the growth hormone sequences and in addition, it was released after completing the growth hormone study. It is for these reasons that it was not implemented in the growth hormone study.

Although the Hsu et al., 2010 pipeline was used in both the folate receptor and the trypsin gene family studies, the average gene conversion length was still much larger in the latter. This is most likely due to the fact that the trypsin genes share a higher degree of sequence similarity (being 84.4 – 90.2% similar) than those in the folate receptor gene family (being 75.6 – 81.5% similar). Studies involving a variety of organisms from yeast to mammals demonstrated that the frequency and size of conversion events depends on the degree of sequence similarity of involved genes (Shen and Huang, 1986, Wang et al., 1999, Ezawa et al., 2006, Drouin, 2002, Liskay et al., 1987, Waldman and Liskay, 1988, Elliott et al., 1998). This

characteristic of gene conversion was observed in all presented studies where genes that shared the highest sequence similarity often had the largest and most frequent gene conversion events. In addition, significant positive correlations were found in the growth hormone and folate receptor gene families between sequence similarity and the length of conversion events. The fact that no significant correlations of this nature were found between the trypsin sequences is most likely due to almost all converted trypsin genes sharing a similarity of 90% not allowing any correlations to be determined (see Figure 4.4).

Additionally, conversions were not found to be more frequent between directly adjacent genes. In all studied gene families, conversions were found to occur between all genes and no biases were observed with respect to their proximities. The only exception was in the TRY3 gene in the chimpanzee and gibbon trypsin gene family. These genes were found to be located a half million base pairs away from the remainder of its gene family and exhibited a much lower sequence similarity and participated in no conversion events.

Interestingly, no positive correlations between the number of gene family members and the number of detected conversion events were observed in any of the three studies, suggesting that the number of conversions is not proportional to the number of genes present in a genome. This is in contrast with the situation in bacterial genomes where the number of conversions is correlated with the size of the gene families, a result which likely reflects more frequent conversion events as the number of potential conversion partners increases (Morris and Drouin, 2007). The fact that we did not find such a positive correlation might be due to the limited

number of genes in our data set.

In each gene family study, phylogenetic trees were produced to assess the impact of conversion events on the orthologous relationships of sequences. In both the growth hormone and trypsin gene families, the expected ancestral orthologous relationships were completely lost due to frequent gene conversion. However, in the folate receptor gene family, the orthologous relationships were somewhat maintained. This may due to conversion events predominantly occurring towards pseudogenes (where the functional genes are the donors) as well as being localized in the introns and flanking regions of genes, leaving the coding regions intact.

Gene conversions are also known to affect the GC-content of converted sequences. The biased gene conversion hypothesis states that due to biases in DNA repair mechanisms, higher gene conversion frequencies will lead to higher GC-content (Meunier and Duret, 2004, Galtier et al., 2001, Kudla et al., 2004). When examining the GC-content of converted regions versus non-converted ones, the growth hormone and folate receptor gene's converted regions displayed a significantly higher GC-content. The trypsin genes did not exhibit this characteristic, however, the lack of significance may be due to all the trypsin genes being so similar that they probably convert one another repeatedly over their whole length. Thus, the conversions we detected likely represent the most recent conversions, not the fact that previous conversion never previously occurred in the other regions. The fact that both converted and non-converted regions have GC-contents that are higher (51.37% and 50.02%, respectively) than the genome wide GC-content of the human genome (41%) supports this interpretation (International Human Genome

Sequencing Consortium et al., 2001).

When comparing the number of gene conversions detected in each primate species, the primate containing the most conversion events differs amongst studies. In the growth hormone gene family, gene conversions were predominantly detected in chimpanzee and were found to have a stronger homogenizing effect in hominoids. In both the folate receptor and trypsin gene families, the rhesus monkey contained the most detected conversion events.

The main observation common to our three studies is how gene conversions were found to be less frequent in conserved gene regions and towards functionally important genes. Although multiple gene conversions were detected in the gene families of each study, the majority of these conversions were found to occur towards pseudogenes and within regions that are of little importance with respect to the functionality of the gene. For example, in the growth hormone gene family, the fact that the conversions spanning both exons and introns were found to be significantly larger than those limited to introns or exon sequences likely represents conversions that occurred in regions under little selective pressures. Moreover, the primary growth hormone gene was involved in the fewest detected conversion events.

In the folate receptor gene family, detected gene conversions were scarce in the functional genes FOLR1 and FOLR2 and therefore had little effect on the coding regions. The fact that the most detected conversions were between FOLR3, the least important folate receptor gene and the pseudogene, FOLR3P1, further demonstrates

that conversion events are free to accumulate in the folate receptor gene family but are limited to less important genes and genetic regions.

Lastly, in the trypsin gene family, strong conservation amongst the trypsin genes at the amino acid level in the studied primates was observed. However, this gene family was found to have the most frequent and largest conversion events. Amino acid residues integral to the functionality of the genes were identified in past studies and genes in various primate species, which had sites that differed from these conserved regions, were not found to convert any other genes. This suggests that all conversions between these trypsin genes most likely result in one section of a gene being replaced by an identical portion of another, thus having no consequences, allowing conversion events to freely accumulate in the trypsin gene family.

All of the aforementioned observations strongly suggest that the gene conversions occurring in all these studied gene families are under strong selective pressures. Conversions are free to accumulate amongst gene family sequences but they are restricted to the less functionally important genes and gene regions that are under little selective constraints. However, just because they may not be detrimental to the function of genes does not mean they do not occur. In fact, our results show that they are actually quite frequent.

Table 5.1
General summary of characteristics, major conclusions and results from each study

	Number of species included	Average % similarity	Conversion detection method used	Species with most conversions	Average conversion length	Conversions more frequent between adjacent genes
GH	13	94%	GENECONV	Homonidae (chimpanzee and human)	197.8 ± 230.4 (with 48 detected conversions)	NO (can only make this claim about human GH genes since the orientation of the GH genes in other species is still not determined)
FOLR	6	81.5%	HSU et al. 2010 pipeline	Rhesus monkey	519.4 ± 492.9 (with 27 detected conversions)	NO
PRSS	6	91%	HSU et al. 2010 pipeline & GENECONV	Rhesus monkey	1526.25 ± 1123.8 (with 37 detected conversions)	NO (exception: TRY3 gene in chimpanzee and gibbon located 0.5 million bp away from cluster contained no conversions)

Table 5.2
General summary of characteristics, major conclusions and results from each study

	Gene conversions observed in tree	Orthologous relationships lost due to conversions	Significant increase in GC content in converted regions	% similarity significantly correlates with conversion tract length	Lower frequency of conversions in conserved genes or gene regions
GH	YES	Completely lost	YES	YES	YES (specifically in GH1)
FOLR	YES	Somewhat maintained	YES	YES	YES (specifically in FOLR1 and FOLR2)
PRSS	YES	Completely lost	NO (may be due to an overall high GC content over the entirety of genes due to ongoing conversion)	NO (no significant correlations could be drawn since all genes are approximately 90% similar to one another)	YES (specifically in the conserved amino acid regions of all trypsin genes)

References:

G. Drouin, Characterization of the gene conversions between the multigene family members of the yeast genome, J. Mol. Evol. 55 (2002) 14–23.

B. Elliott, C. Richardson, J.Winderbaum, J.A. Nickoloff, M. Jasin, Gene conversion tracts from double-strand break repair in mammalian cells, Mol. Cell. Biol. 18 (1998) 93–101.

K. Ezawa, S. Oota, N. Saitou, Genome-wide search of gene conversions in duplicated genes of mouse and rat, Mol. Biol. Evol. 23 (2006) 927–940.

N. Galtier, G. Piganeau, D. Mouchiroud, and L. Duret. GCcontent evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics (2001) 159:907–911.

C.H. Hsu, Y. Zhang, R.C. Hardison, Nisc Comparitive Sequence Program, E.D. Green, and W. Miller, An Effective Method for Detecting Gene Conversion
Events in Whole Genomes, Journal of Computational Biology 17 (2010) 1281-1297.

International Human Genome Sequencing Consortium, E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, et al., Initial sequencing and analysis of the human genome, Nature (2001) 409:860-921.

G. Kudla, A. Helwak, and L. Lipinski, Gene conversion and GC-content evolution in mammalian Hsp70. Mol. Biol. Evol. (2004) 21:1438–1444.

R.M. Liskay, A. Letsou, J.L. Stachelek, Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells, Genetics 115 (1987) 161–167.

J. Meunier, and L. Duret, Recombination drives the evolution of GC-content in the human genome. Mol. biol. Evol. (2004) 21:984–990.

R.T. Morris, G. Drouin, Ectopic gene conversions in bacterial genomes, Genome 50 (2007) 975–984.

D. Posada, Evaluation of methods for detecting recombination from DNA sequences: empirical data, Mol. Biol. Evol. 19 (2002) 708–717.

D. Posada, K.A. Crandall, Evaluation of methods for detecting recombination from DNA sequences: computer simulations, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 13757–13762.

P. Shen, H.V. Huang, Homologous recombination in Escherichia coli: dependence on substrate length and homology, Genetics 112 (1986) 441–457.

G. Song, C.H. Hsu, C. Riemer, W. Miller, Evaluation of methods for detecting conversion events in gene clusters, BMC Bioinformatics 12 (2011) 1471-2105.

A.S. Waldman, R.M. Liskay, Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology, Mol. Cell. Biol. 8 (1988) 5350–5357.

S. Wang, C. Magoulas, D. Hickey, Concerted evolution within a trypsin gene cluster in Drosophila, Mol. Biol. Evol. 16 (1999) 1117–1124.