Sexy Gene Conversions: Locating Gene Conversions on the X-Chromosome

Mark J. Lawson[1], Liqing Zhang[1,2*]
Department of Computer Science, Virginia Tech
[2]Program in Genetics, Bioinformatics, and Computational Biology
*To whom correspondence should be addressed; E-mail: lqzhang@vt.edu

April 3, 2009

Abstract

Gene conversion can have a profound impact on both the short-term and long-term evolution of genes and genomes. Here we examined the gene families that are located on the X-chromosomes of human, chimp, mouse, and rat for evidence of gene conversion. We identified seven gene families (WD repeat protein family, Ferritin Heavy Chain family, RAS-related Protein RAB-40 family, Diphosphoinositol polyphosphate phosphohydrolase family, Transcription Elongation Factor A family, LDOC1 Related family, Zinc Finger Protein ZIC, and GLI family) that show evidence of gene conversion. Through phylogenetic analyses and synteny evidence, we show that gene conversion has played an important role in the evolution of these gene families and that gene conversion has occured independently in both primates and rodents. Comparing the results with those of two gene conversion prediction programs (GENECONV and Partimatrix), we found that both GENECONV and Partimatrix have very high false negative rates (i.e. failed to predict gene conversions), leading to many undetected gene conversions. The combination of phylogenetic analyses with physical synteny evidence exhibits high power in the detection of gene conversions.

1 Introduction

Gene conversions are the exchange of genetic information between two genes, initiated by a double-strand break in one gene (acceptor) followed by the repair of this gene through the copying of the sequence of a similar gene (donor). This process has important implications for the evolution of an organism. On the one hand, gene conversion between genes can lead to homogenization of these genes, thus preventing their divergence and leading to evolutionary conservation of the genes. On the other hand, gene conversion between alleles can lead to increased polymorphism and genetic diversity.

The X-chromosome (along with the Y-chromosome) plays an important role in sex determination in humans and other mammals. It contains many genes that are important for reproduction such as genes that are testis-specific [1]. In addition it also contains an abundance of genes that are linked to mental development and cognitive ability [2, 3] and is considered highly conserved in size, gene content, and gene order [4]. The X-chromosome contains a disproportionately high number of Mendelian diseases (a summary lists 168 diseases explained by 113 X-linked genes [1] and a recent search through OMIM listed 1049 inheritable disorders that originated from the X-chromosome [5]). This large number is explained through the fact that males only have one X-chromosome and thus males express recessive phenotypes (which is in many cases the disease-causing phenotype).

Many cases of gene conversion between genes on the X chromosome have been reported. A well-studied gene conversion is the one that occurs between *OPN1LW* and *OPN1MW* (the red and green opsin genes, respectively). These genes are located only 24 kb apart from each other and a gene conversion between the two can lead to color blindness, also known as blue cone monochromacy [6, 7]. Additionally, analysis of *OPN1LW* has revealed a large amount of alleles created through gene conversions on exon 3 [8]. Another interesting gene conversion was discovered between the genes *Pgk* and *Pdha* [9]. These genes are both located on the X-chromosome but are not tandemly arranged, as is the case with the majority of gene conversions. Instead they are in "widely separated chromosome locations". It has been proposed that there have been two major recombination (i.e. gene conversion) events: one around the placental-marsupial split and one around the primate-rodent split.

Gene conversions on the X-chromosome have also been shown to be linked with genetic diseases and disorders. For instance, intrachromosomal gene conversions occurring on intron 1 or intron 22 of the *F8* gene lead to sequence inversion [10], which in turn lead to severe hemophilia A. The gene family *SPANX* (which consists of five genes) owes its high amount of diversity to gene conversions [11, 12]. This high amount of variation leads to haplotypes with an increased susceptibility to prostate cancer.

Due to the importance of the X-chromosome and the amount of important gene conversions discovered on it, we undertook a wide-scale analysis of the X-chromosome in order to identify more gene conversions. Starting with known gene families (as gene conversions typically occur between duplicated genes) we attempt to identify gene conversions through a variety of methods including using the existing programs designed for identification of gene conversion and combining phylogenetic tree analyses with evidence of gene physical linkage. Our results show the limitations of existing software for gene conversion prediction, and the need to incorporate a variety of information to increase the power of gene conversion detection.

2 Materials and Methods

We selected the human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), chimp (*Pan troglodytes*), and rhesus monkey (*Macaca mulatta*) genomes for our X-chromosome gene conversion analysis. We selected these genomes as they are relatively well-studied and annotated. Furthermore, this set of genomes allows us to draw conclusions based on biological order, i.e. primates vs. rodentia.

Gene conversions typically occur between genes that have a high level of sequence similarity and are a major mechanism of concerted evolution within gene families. Since our focus is on gene conversions between genes on the X chromosome, we started our search with gene families that are located on the X-chromosome. We used PANTHER [13], a gene family database that combines sequence similarity information with expert knowledge on protein function to cluster genes into families.

This provides us with gene families that are similar in both functions and sequences, allowing us to make assessment about the functional importance of any potential gene conversions. The only shortcoming with using the PANTHER database is that they only have gene families for the human, mouse, and rat genomes.

Extracting only those families that have most of their genes on the X-chromosome, we picked those subfamilies that had at least two genes in humans. We then took a three-pronged attack to find orthologs of these genes in other species. The first way we established orthology is by looking at genes that are in the same subfamily in another species (this is limited to mouse and rat genes). The second way was by using listings of orthologs from Evola [14], HomoloGene [5], and Ensembl [15], which allowed us to include chimp and rhesus monkey genes. The third and final way was to select genes that have the same name as those original human genes in the other species. This was done because orthology can be obscured through concerted evolution.

The coding sequences of the gene families were aligned based on the alignments of protein sequences. Pairwise distance matrices were computed using the HKY model [16]. Phylogenetic trees were constructed using the neighbor-joining method with 1000 bootstraps in Phylip [17]. Gene conversions were then identified by visually inspecting all the trees. Instances where genes were grouped with other genes within the same species (or the same biological order) instead of with their known orthologs were marked as having gene conversions. The idea here is that gene conversions have kept these sequences more similar to each other within the species due to gene conservation and concerted evolution.

We retrieved the locations of these genes from Ensembl to establish synteny for these genes across the species. We included genes around and in between those genes exhibiting conversions. These figures provide visual evidence that the genes are located in the same syntenic regions across all species, thus giving evidence for their conservation across these species.

To provide additional evidence of gene conversions, we used two programs that are designed to identify genetic recombination. The first program is GENECONV, a program specifically designed to identify gene conversions [18]. It identifies highly similar sequence subsets within a set of aligned sequences and then determines their statistical significance by computing a global p-value and a local p-value. In addition, it lists data such as the length and location of these subsets (which we list as the gene conversion lengths in Tables 1 and 2). Those with low p-values (default is less than 0.05) are considered to be indicative of gene conversion events. As global p-value is calculated based on consideration of all sequences in the alignments, whereas the pairwise p-value is based on consideration of only the pair of sequences of interest, global comparison is considered to be more conservative than pairwise [18].

The second program is Partimatrix [19]. Partimatrix takes a set of aligned sequences and then creates partitions of one or more sequences that are equivalent to branches in a phylogenetic tree. These partitions are then given a support score for a recombination event having occurred between the sequences sharing this partition. This support score is balanced with a conflict score that indicates how unique this partition is. So partitions with a high support score and low conflict score provide evidence for a gene conversion event.

3 Results

Altogether, we identified 7 gene families that exhibit signs of gene conversions. They are (using the PANTHER IDs): PTHR14754, PTHR19860:SF2, PTHR11431:SF14, PTHR11708:SF146, PTHR12629, PTHR15503, and PTHR19818:SF18. In the following sections we show the results of gene conversion analyses for each of these families.

3.1 PTHR14754

This family is referred to as "Transcription Elongation Factor A," and the protein products of these genes are involved in mRNA transcription elongation.

Figure 1 shows the phylogeny of the human *TCEAL3*, *TCEAL5*, and *TCEAL6* genes and their orthologs across the other four species. Again we have an example of gene conversions occurring within biological

order (primates and rodentiae) as the mouse and rat genes are more similar to each other, and the human/chimp/rhesus genes show similar pattern. Within mouse there may even be a stronger level of gene conversion between *Tceal3* and *Tceal6*, however we were unable to find *Tceal6* in the rat genome.

Figure 2 shows the locations of these genes together with upstream and/or downstream neighboring genes in human and mouse (other species show similar information and for brevity are not shown here). The synteny strongly suggests that these genes arose before the speciation of primates and murids.

Global GENECONV analysis revealed gene conversion between *TCEAL3* and *TCEAL5* in rat, betwen *TCEAL5* and *TCEAL6* in mouse, between *TCEAL3* and *TCEAL5* in mouse, between *TCEAL5* and *TCEAL6* in human, and between *TCEAL5* and *TCEAL6* in chimp (Table 1). The converted regions in these gene conversions range from 140 to 313 base pairs. The pairwise analysis shows similar results (Table 2). The main difference is that the pairwise analysis detected an additional gene conversion between *TCEAL3* and *TCEAL6* in human, and did not detect any conversion between *TCEAL3* and *TCEAL5* in rat. Partimatrix analysis shows high support scores and rather low conflict scores for mouse *TCEAL3*, *TCEAL5*, and *TCEAL6*, suggesting gene conversion beween these three genes in mouse (Table 3). Similarly, according to the Partimatrix analysis, rat *TCEAL3* and *TCEAL5* and human *TCEAL3* and *TCEAL6* appear to have undergone gene conversion.

3.2 PTHR19860:SF2

This family contains the human *WDR40B* and *WDR40C* genes and their subsequent orthologs across the other species. These genes are part of the WD repeat protein family and thought to be involved in a variety of cellular processes such as cell cycle progression, signal transduction, apoptosis, and gene regulation [5].

The phylogenetic tree for these genes is shown in Figure 3. The predicted rat genes are presumbly the equivalent orthologs of *WDR40B* and *WDR40C* as they are within the same subfamily in PANTHER. This is further supported by the high similarity between them and the mouse genes. An ortholog of *WDR40B* was not found in the rhesus monkey. Figure 3 suggests evidence for gene conversion between biological orders. The rodentiae (mouse and rat) genes are more similar to each other than to their orthologs in the primates (human, chimp, rhesus). This gives rise to the theory that these genes have maintained similarity within their order in an effort to conserve their sequences. Evidence of gene conversion is further strongly supported by the synteny information of these genes (see Supplementary Materials).

In the global analysis, GENECONV detected gene conversion between *WDR40B* and *WDR40C* in chimps and the length of gene conversion is 74 base pairs (Table 1). The pairwise analysis of GeneConv detected gene conversions between *WDR40B* and *WDR40C* in chimp, human, and mouse, with gene conversion lengths of 74, 55, and 30 base pairs, respectively (Table 2). The Partimatrix analysis shows that the two mouse and rat genes have the highest support scores and very low conflict scores, reflecting gene conversion between the two mouse paralogous genes and the two rat paralogous genes (Table 3). The evidence of gene conversion in the human and chimp genes is not as strong as suggested by the "weak" conflict and support scores in Partimatrix.

3.3 PTHR11431:SF14

This family is also referred to as the "Ferritin Heavy Chain" family and consists of multiple sequences under which *FTHL17* and *FTHL19* are the most prominent members (the remaining ones have unclear or unassigned annotations). These genes are involved in cation transport and homeostasis activities, while also functioning as storage proteins.

Figure 4 shows the phylogenetic tree for the genes. The *FTHL17* gene is more closely related to *FTHL19* in the primates than it is to the mouse *FTHL17* gene. This gives evidence for a gene conversion among biological order. Furthermore, the mouse gene *EG436193* is part of the same subfamily and is assumed to be an equivalent to FTHL19, giving further evidence for gene conversion. The synteny comparison shows that these genes arose before the speciation of primates and murids (see the Supplementary Materials).

GENECONV found no evidence of gene conversion for both pairwise and global analysis (Tables 1 and 2). Partimatrix analysis shows weak evidence of gene conversion between *FTHL17* and *FTHL19* in the

mouse (Table 3).

3.4 PTHR11708:SF146

This subfamily is referred to as the "RAS-related Protein RAB-40" family and is part of the "RAS-related GTPase" family.

Figure 5 shows the phylogenetic tree for the *RAB40A* and *RAB40AL* genes. Here we can clearly see evidence for a gene conversion event as the human *RAB40A* gene and the *RAB40AL* gene have a closer relation to each other than to their orthologs. In fact, it would seem that *RAB40AL* has donated part of its sequence to *RAB40A* as the human RAB40A is located within the RAB40AL genes subtree. The detailed synteny graph of the region around these genes (see the Supplementary Materials) indicates that this is a highly conserved region where gene order has been well kept including a few other gene families such as the *BEX* and *TCEAL* genes.

Global GENECONV analysis did not detect any gene conversion (Table 1). Pairwise GENECONV analysis detected a gene conversion between RAB40A and RAB40AL genes in the chimp with the length of 103 base pairs (Table 2). Partimatrix detected no gene conversions.

3.5 PTHR12629

Referred to as the "Diphosphoinositol polyphosphate phosphohydrolase" family, genes belonging to this family create phosphotase and are involved in phospholipid metabolism. *NUDT10* and *NUDT11* are phosphohydrolases that preferentially attack diphosphoinositol polyphosphates [20] and are strongly expressed in testis and brain [21]. These two genes are highly similar having six nucleotide differences in humans [21]. Due to this high similarity and their close proximity, it was assumed that they are the result of a recent gene duplication . However, as can be seen in Figure 6, we identified orthologs in mouse and rat, leading us to speculate that this duplication occurred before the speciation event between primates and rodents and that this high similarity is maintained between these genes by gene conversion.

In addition, the synteny graph in the Supplementary Material indicates that an inversion event has occurred in the mouse and rat X-chromosomes at this location (the gene order in rat is similar to that of the mouse). The highly conserved synteny further supports that *NUDT10* and *NUDT11* arose from a duplication that predates the split of rodents and primates, and the high sequence similarity between the two genes is due to frequent gene conversion.

Global GENECONV analysis detected no gene conversions. Pairwise GENECONV analysis shows that there have been gene conversions between *NUDT10* and *NUDT11* in human, mouse, and rat of lengths 228, 495, and 395, respectively (Table 2). Partimatrix analysis reveals gene conversions between *NUDT10* and *NUDT11* in both mouse and rat (Table 3).

3.6 PTHR15503

This family is the "LDOC1 Related" family and involved in cell proliferation and differentiation. *LDOC1* has been shown to be highly expressed in both brain and thyroid tissues, while being downregulated in cancer cell lines [22].

The phylogenetic tree for this family can be seen in Figure 7. Important to note is that the *Cxx1* genes from mouse are considered to be homologous to the *FAM127* genes of human. HomoloGene has *Cxx1a*/*FAM127C* and *Cxx1c*/*FAM127A* listed as orthologs. More exact homology seems to be obscured by the high sequence similarity between these genes due most likely to gene conversion events. Here again we can see a clear grouping by biological order. Furthermore, the *Cxx1* set of genes in mouse appear highly conserved as they are grouped closer to the rat genes that were listed as belonging to this family as well. Synteny for this family can be seen in the Supplementary Material.

Global GENECONV analysis detected no gene conversion (Table 1). However, pairwise GENECONV analysis shows that there is gene conversion between *Cxx1c* and *Cxx1a*, and between *Cxx1c* and *Cxx1b* in mouse (Table 2). Thus, *Cxx1c* seems to have gene conversion with two of its family members, and

interestingly, the converted regions appear to overlap when we look at the more detailed GENECONV output. Partimatrix shows supporting evidence for gene conversion among *Cxx1c*, *Cxx1a*, and *Cxx1b* in mouse, and gene conversion between the two rat genes (Table 3).

3.7 PTHR19818:SF18

This subfamily is part of the "Zinc Finger Protein ZIC and GLI" family in PANTHER. Its molecular function consists of creating the zinc finger transcription factor that can regulate mRNA transcription. Two X-linked gene members of the subfamily, *ZXDA* and *ZXDB*, derive from a very ancient and highly conserved gene duplication event and are subject to X-inactivation [23].

In the phylogenetic tree in Figure 8, we can see that again we have a split by biological order. Furthermore, the genes seem to have very high similarity which is indicated by the very short branch lengths. This high similarity could account for why the rat *Zxda* gene is considered to be closer related to the mouse *Zxda* gene. Unfortunately, we could not find a definitive rat *Zxda* gene.

The synteny graph in the Supplementary Material shows the locations of *ZXDA* and *ZXDB* for human and mouse. All species show similar locations for the genes. Furthermore, these genes are relatively isolated across all species with other genes located further downstream and upstream by relatively large margins.

Global and pairwise GENECONV analyses showed similar results; both identified gene conversions between *Zxda* and *Zxdb* genes in human, chimp, and mouse (Tables 1 and 2). Interestingly, predicted gene conversion lengths are quite long, with human and chimp having more than 450 base pairs, and mouse having more than 1000 base pairs (pairwise analysis) and 2000 base pairs (global detection). The results of Partimatrix are consistent with the findings of GENECONV, also suggesting gene conversions in the three species (Table 3).

4 Discussion

At its most fundamental level, gene conversion is mutagenic. The mutagenic nature of gene conversion, among the many roles that gene conversion can play, is the most easily neglected aspect in the current research [24]. Because gene conversion involves double-strand breaks and repairing double-strand breaks can introduce nucleotide mutations [25, 24], frequent gene conversion can increase the amount of DNA variation and lead to rapid rates of evolution. It has been shown that some mutation hot spots in the human genome may actually be the results of biased gene conversions rather than of adaptation [26]. In the short-term, gene conversion can shape the short-range patterns of linkage disequilibrium (LD) in a genome [27]. Disregarding the effect of gene conversion can lead to spurious inferences of a species population history and effective population size [28]. In addition, gene conversion between recently duplicated genes can create novel haplotypes in the population [29]. The process of gene conversion becomes especially important to consider in studies of human populations due to large-scale segmental duplications in the human populations [30, 31].

The best known effect of gene conversion is on sequence evolution of gene families. Ample cases of gene conversion on the X-chromosome have been identified. A case of gene conversion that is especially relevant to the current study is the region between 100.5 and 102.8 Mb (mega base pairs). Within this region there are 20 genes (encompassing three families: *BEX*, *WEX*, and *GASP*) that have undergone concerted evolution in humans. While the coding regions within these families are highly diverged, the 5'-UTR regions are highly conserved. A more extensive analysis of the *BEX* family across multiple species confirmed gene conversion in the 5'-UTR regions and suggested evidence of positive selection on these genes [32]. Interestingly, our study reveals additional gene conversion activities in the region surrounding the *BEX* genes. *RAB40AL* is located 124360bp upstream of *BEX1* and *RAB40A* is located 188733 downstream of *BEX2*. *TCEAL6* is located 10742bp upstream of *BEX5*, *TCEAL5* is 32482bp upstream of *BEX2*, and *TCEAL3* is located 296434 downstream of *BEX2*. Therefore this large syntenic region seems to be ripe with gene conversion and gene conversion has been ongoing independently in multiple mammalian species. Further interesting to note is the similarity of these genes' functions and expressions in relation to the brain. The BEX gene family is known

to be highly expressed in brains, for example, accounting for more than 12% of the expressed sequenced tags in the rat brain [33]. The RAB40AL and RAB40A genes create a small GTPase that regulates many intracellular processes such as cytoskeletal organization and secretion. It is also involved in some biological processes, such as receptor mediated endocytosis and general vesicle transport. Research has shown that defects in the RAB40AL gene can lead to mental retardation [34], indicating that these genes play an important role in mental development.

Our study demonstrates the effect of gene conversion on deterring the sequence divergence of duplicated genes. However, gene conversion can also promote genetic diversity, as in the case of immunoglobulin genes [35]. The relative contribution that gene conversion makes to sequence diversity and conservation is not known. It is likely that depending on the functional requirement of the genes, gene conversion can be utilized to achieve both diversity and conservation. Our recent extensive survey of literature on gene conversion ("A Pattern Analysis of Gene Conversion Literature" - submitted) indicates that among the 2478 papers, 551 of them focus on the role of gene conversion in boosting genetic diversity of genes and 402 of them on the role of gene conversion to deter divergence in certain gene families. These raw numbers seem to suggest that the two consequences are equally likely manifested in the literature. Whether this reflects the relative contribution of gene conversion to diversity and conservation in nature needs further investigation.

The balance between gene conversion and mutations is an intricate and dynamic process. Constant gene conversion can slow down the divergence of two duplicated genes. However, because frequency of gene conversion is a function of sequence identity, rates of gene conversion will, presumably, decrease as mutation erodes sequence identity and will eventually stop when the sequence identity reaches a certain threshold. It is unclear what threshold value of sequence identity is required for gene conversion. Analysis of pathogenic gene conversion events in humans suggests that at least 92% sequence identity might be required for gene conversion to occur [36]. However, the identities could be the result of current detection methods which fail to identify more ancient gene conversions. To our knowledge, no theoretical studies are available to examine when the process of gene conversion will eventually stop. In our study, it is also possible that mutation rates in these genes happen to be very slow, which enables gene conversion to occur frequently and relatively recently in both primates and rodents. If this is the case, we expect that these X-chromosome orthologous genes in human and mouse should have lower sequence divergences than the autosomal orthologs that belong to the same gene family. We therefore obtained the autosomal genes for the gene families and calculated and compared the sequence identity of autosomal orthologs and X-chromosome orthologs between human and mouse. We found no suggestions of slower rate of evolution in the X-linked genes than autosome-linked genes (Table 4).

Much confusion and errors can be caused by gene conversion. For example, in our study, there is the case of NUDT10 and NUDT11. From the synteny map and sequence information, it is clear that the annotation of the two human genes is wrong. Furthermore, Ensembl lists two mouse genes as being Nudt11 with no Nudt10 being listed. We determined which gene is the Nudt10 gene based on the surrounding genes. In the case of Tceal3 and Tceal6, the highly conserved synteny map across species suggests that these genes arose at least before the split of primates and rodents. However, the literature considers them to be the product of a recent gene duplication. Besides the erroneous gene annotation caused by gene conversion, gene conversion can also lead to artificially inflated estimates of rates of gene duplication by several orders of magnitude [37].

Characterizing the process of gene conversion has been challenging. Our recent analysis of literature shows that most studies of gene conversion mechanisms have been done in yeast("A Pattern Analysis of Gene Conversion Literature" - submitted), due to the convenience of observing the direct product of gene conversion. Identification of gene conversion in sequence evolution is another challenging problem. Our results and methodology clearly demonstrate the difficulty. Using the phylogenetic trees alone to identify gene conversion is insufficient, especially in the case of independent gene conversions in multiple species. In this case, taking the tree literally will lead to the conclusion of independent duplications instead of multiple gene conversions in multiple species. As we know little about the relative frequencies of gene conversion and gene duplication, which likely varies from gene to gene or even species to species, it is hard to draw a definite conclusion on which scenario is more likely. Our study shows that the strongest evidence of gene conversion comes from the combination of phylogeny and synteny maps. For all the seven gene families that

we examined, well-conserved synteny maps shared between the rodents and primates offers strong evidence for the existence of regions and therein genes in the regions before the split of the two groups of species.

The usage of computer programs is also far from sufficient, as illustrated by the results of GENECONV and Partimatrix analyses, both of which can have as high as 100% false negative rates (e.g., GENECONV finds no gene conversions in PTHR11431:SF14). Our previous study shows that although GENECONV and Partimatrix are to some extent complementary in terms of sequence divergence, with GENECONV more powerful in predicting highly identical sequences and Partimatrix performing more steadily with respect to sequence divergence, combining the two programs' predictions together using a boosting algorithm still does not raise the accuracy significantly [38]. Taken together, we believe that many incidences of gene conversion have gone undetected due to the technical difficulty in identifying these events. This might be one of the evolutionary issues that have "blind spots:" gene conversion has been a wide-spread and frequent process that affects the divergence of duplicated genes, yet using the existing programs to predict gene conversion in large-scale may give one only the "prominent" gene conversion cases, while many of less "prominent" gene conversion cases might go undetected.

Although largely neglected, the role of gene conversion in the evolutionary conservation of noncoding regions, has begun to emerge from recent studies. It has been shown that the 5'-UTR of the *BEX* gene family achieved high conservation due to gene conversions in the region [39, 40] and that a functional motif was discovered in the region, suggesting that selection for motif conservation may keep gene conversion ongoing [39]. Another example is the evolution of the RNAase noncoding regions. A study of the RNAase noncoding regions (including 5'-UTR, intron, and 3'-UTR) in leaf-eating African and Asian Colobine Monkeys has revealed a high level of homogenization through gene conversion in the regions. Interestingly, gene conversion seems to be largely restricted to the noncoding regions of the gene, whereas coding regions display much higher divergence than that of noncoding regions. This has been considered to be the result of selection to prevent gene conversion in coding regions in order to keep two functionally distinct pRNase proteins in these species [41]. Clearly, a complete understanding of the functional significance of gene conversion in noncoding regions calls for more studies like this one, which will benefit from the development of algorithms that identify gene conversion specifically in noncoding regions.

5 Acknowledgements

We thank the anonymous reviewers for their ideas and comments.

6 Tables

6.1 GENECONV Results

Table 1: Global GENECONV p-values

Family	**Gene 1**	**Gene 2**	**Sim p-value**	**KA p-value**	**GC Length**
PTHR14754	RN_TCEAL5	RN_TCEAL3	0.0000	0.00000	313
	MM_Tceal5	MM_Tceal6	0.0000	0.00000	244
	MM_Tceal5	MM_Tceal3	0.0000	0.00000	222
	HS_TCEAL5	HS_TCEAL6	0.0064	0.03491	130
	PT_TCEAL5	PT_TCEAL6	0.0065	0.03537	140
PTHR19860:SF2	PT_WDR40C	PT_WDR40B	0.0089	0.02628	74
PTHR19818:SF18	MM_Zxdb	MM_Zxda	0.0000	0.00000	2107
	PT_ZXDB	PT_ZXDA	0.0000	0.00018	456
	HS_ZXDB	HS_ZXDA	0.0001	0.00022	462

Table 2: Pairwise GENECONV p-values

Family	**Gene 1**	**Gene 2**	**Sim p-value**	**KA p-value**	**GC Length**
PTHR14754	HS_TCEAL5	HS_TCEAL6	0.0001	0.00063	130
	PT_TCEAL5	PT_TCEAL6	0.0006	0.00092	95
	HS_TCEAL6	HS_TCEAL3	0.0173	0.03593	141
	MM_Tceal5	MM_Tceal3	0.0000	0.00000	124
	MM_Tceal5	MM_Tceal6	0.0000	0.00000	244
PTHR19860:SF2	PT_WDR40C	PT_WDR40B	0.0007	0.00073	74
	HS_WDR40B	HS_WDR40C	0.0073	0.01019	55
	MM_Wdr40c	MM_Wdr40b	0.0078	0.01129	30
PTHR11708:SF146	PT_RAB40A	PT_RAB40AL	0.0109	0.02073	103
PTHR12629	HS_NUDT11	HS_NUDT10	0.0158	0.09260	228
	RN_Nudt10	RN_Nudt11	0.0015	0.24290	495
	MM_Nudt11	MM_Nudt10	0.0288	0.19226	395
PTHR15503	MM_Cxx1c	MM_Cxx1a	0.0420	0.05943	136
	MM_Cxx1c	MM_Cxx1b	0.0364	0.05324	136
PTHR19818:SF18	HS_ZXDB	HS_ZXDA	0.0000	0.00001	462
	PT_ZXDB	PT_ZXDA	0.0000	0.00000	456
	MM_Zxdb	MM_Zxda	0.0000	0.00000	1184

6.2 Partimatrix Results

6.3 Sequence Identities

Table 3: Partimatrix values

Family	Gene Combination	Support Score	Conflict Score
PTHR14754	MM_Tceal5/MM_Tceal3/MM_Tceal6/RN_Tceal5/RN_Tceal3	59.5	2.53
	MM_Tceal5/RN_Tceal5	35.5	0.64
	MM_Tceal3/MM_Tceal6/RN_Tceal3	35.0	1.60
	HS_TCEAL5/PT_TCEAL5/MM_Tceal5/RN_Tceal5	34.5	3.35
	HS_TCEAL5/PT_TCEAL5	13.0	0.11
	HS_TCEAL6/HS_TCEAL3/PT_TCEAL6/RM_TCEAL6	11.5	0.77
PTHR19860:SF2	MM_Wdr40C/RN_RGD1560768/MM_Wdr40b/RN_RGD1563873	82.5	2.02
	MM_Wdr40b/RN_RGD1563873	75.5	0.80
	MM_Wdr40C/RN_RGD1560768	47.5	1.22
	HS_WDR40B/PT_WDR40B	44.5	0.46
	HS_WDR40B/PT_WDR40B/HS_WDR40C/RM_WDR40C	20.0	2.27
PTHR11431:SF14	MM_EG436193/RN_RGD1563161	44.0	1.09
	MM_EG436193/RN_RGD1563161/MM_FTHL17	35.0	1.93
	HS_FTHL19/RM_FTHL19	16.5	0.59
	MM_EG436193/RN_RGD1563161/PT_FTHL19	13.0	3.03
	MM_FTHL17/MM_EG436193	10.5	3.05
	HS_FTHL17/PT_FTHL17/RM_FTHL17	10.0	0.43
PTHR11708:SF146	HS_RAB40A/PT_RAB40A/RM_RAB40A	21.0	2.28
	HS_RAB40A/PT_RAB40A	7.5	1.40
PTHR12629	RN_Nudt10/RN_Nudt11/MM_Nudt10/MM_Nudt11	53.0	1.05
PTHR15503	MM_Ldoc1/RN_Ldoc1/HS_LDOC1	23.0	0.95
	MM_Ldoc1/RN_Ldoc1/HS_LDOC1/PT_LDOC1	19.0	1.57
	MM_Ldoc1/RN_Ldoc1	16.0	0.44
	MM_Cxx1a/MM_Cxx1b/MM_Cxx1c/RN_LOC678880/RN_LOC679038	12.5	1.09
PTHR19818:SF18	MM_Zxda/MM_Zxdb/RN_Zxdb	274.5	1.90
	MM_Zxda/MM_Zxdb	117.5	1.23
	HS_ZXDA/PT_ZXDA	20.0	1.41
	HS_ZXDA/PT_ZXDA/HS_ZXDB/PT_ZXDA	15.0	3.07
	MM_Zxda/RN_Zxdb	14.0	5.48
	HS_ZXDA/PT_ZXDA/RM_ZXDA	13.5	2.87
	HS_ZXDB/PT_ZXDA	10.0	0.18

Table 4: Sequence Identities Between Human and Mouse Orthologs

Family	**X-chromosome Orthologs**	**% Ident**	**Autosome Orthologs**	**% Ident**
PTHR11431:SF14	HS_FTHL17-MM_FTHL17	59.2	HS_FTH1-MM_FTH1	61.0
	HS_FTHL19-MM_FTHL19	51.6	HS_FTMT-MM_Ftmt	52.4
PTHR12629	HS_NUDT11-MM_Nudt11	69.2	HS_NUDT4-MM_Nudt4	45.6
	HS_NUDT10-MM_Nudt10	68.1	HS_NUDT3-MM_Nudt3	45.5
PTHR15503	HS_LDOC1-MM_Ldoc1	65.4	HS_PEG10-MM_Peg10	54.1
	HS_FAM127A-MM_Cxx1c	51.6	HS_LDOC1L-MM_Ldoc1l	48.6
	HS_FAM127B-MM_Cxx1b	61.5		
	HS_FAM127C-MM_Cxx1a	40.5		
PTHR19818:SF18	HS_ZXDB-MM_Zxdb	68.9	HS_ZXDC-MM_Zxdc	61.8
	HS_ZXDA-MM_Zxda	32.1		
PTHR19860:SF2	HS_WDR40B-MM_Wdr40b	63.7	HS_WDR40A-MM_Wdr40a	39.1
	HS_WDR40C-MM_Wdr40c	39.7		

References

[1] Ross, M., Grafham, D., Coffey, A., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G., Burrows, C., Bird, C., Frankish, A., Lovell, F., Howe, K., Ashurst, J., Fulton, R., Sudbrak, R., Wen, G., Jones, M., Hurles, M., Andrews, T., Scott, C., Searle, S., Ramser, J., Whittaker, A., Deadman, R., Carter, N., Hunt, S., Chen, R., Cree, A., Gunaratne, P., Havlak, P., Hodgson, A., Metzker, M., Richards, S., Scott, G., Steffen, D., Sodergren, E., Wheeler, D., Worley, K., Ainscough, R., Ambrose, K., Ansari-Lari, M., Aradhya, S., Ashwell, R., Babbage, A., Bagguley, C., Ballabio, A., Banerjee, R., Barker, G., Barlow, K., Barrett, I., Bates, K., Beare, D., Beasley, H., Beasley, O., Beck, A., Bethel, G., Blechschmidt, K., Brady, N., Bray-Allen, S., Bridgeman, A., Brown, A., Brown, M., Bonnin, D., Bruford, E., Buhay, C., Burch, P., Burford, D., Burgess, J., Burrill, W., Burton, J., Bye, J., Carder, C., Carrel, L., Chako, J., Chapman, J., Chavez, D., Chen, E., Chen, G., Chen, Y., Chen, Z., Chinault, C., Ciccodicola, A., Clark, S., Clarke, G., Clee, C., Clegg, S., Clerc-Blankenburg, K., Clifford, K., Cobley, V., Cole, C., Conquer, J., Corby, N., Connor, R., David, R., Davies, J., Davis, C., Davis, J., Delgado, O., DeShazo, D., Dhami, P., Ding, Y., Dinh, H., Dodsworth, S., Draper, H., Dugan-Rocha, S., Dunham, A., Dunn, M., Durbin, K., Dutta, I., Eades, T., Ellwood, M., Emery-Cohen, A., Errington, H., Evans, K., Faulkner, L., Francis, F., Frankland, J., Fraser, A., Galgoczy, P., Gilbert, J., Gill, R., Glockner, G., Gregory, S., Gribble, S., Griffiths, C., Grocock, R., Gu, Y., Gwilliam, R., Hamilton, C., Hart, E., Hawes, A., Heath, P., Heitmann, K., Hennig, S., Hernandez, J., Hinzmann, B., Ho, S., Hoffs, M., Howden, P., Huckle, E., Hume, J., Hunt, P., Hunt, A., Isherwood, J., Jacob, L., Johnson, D., Jones, S., deJong, P., Joseph, S., Keenan, S., Kelly, S., Kershaw, J., Khan, Z., Kioschis, P., Klages, S., Knights, A., Kosiura, A., Kovar-Smith, C., Laird, G., Langford, C., Lawlor, S., Leversha, M., Lewis, L., Liu, W., Lloyd, C., Lloyd, D., Loulseged, H., Loveland, J., Lovell, J., Lozado, R., Lu, J., Lyne, R., Ma, J., Maheshwari, M., Matthews, L., McDowall, J., McLaren, S., McMurray, A., Meidl, P., Meitinger, T., Milne, S., Miner, G., Mistry, S., Morgan, M., Morris, S., Muller, I., Mullikin, J., Nguyen, N., Nordsiek, G., Nyakatura, G., O'Dell, C., Okwuonu, G., Palmer, S., Pandian, R., Parker, D., Parrish, J., Pasternak, S., Patel, D., Pearce, A., Pearson, D., Pelan, S., Perez, L., Porter, K., Ramsey, Y., Reichwald, K., Rhodes, S., Ridler, K., Schlessinger, D., Schueler, M., Sehra, H., Shaw-Smith, C., Shen, H., Sheridan, E., Shownkeen, R., Skuce, C., Smith, M., Sotheran, E., Steingruber, H., Steward, C., Storey, R., Swann, R., Swarbreck, D., Tabor, P., Taudien, S., Taylor, T., Teague, B., Thomas, K., Thorpe, A., Timms, K., Tracey, A., Trevanion, S., Tromans, A., d'Urso, M., Verduzco, D., Villasana, D., Waldron, L., Wall, M., Wang, Q., Warren, J., Warry, G., Wei, X., West, A., Whitehead, S., Whiteley, M., Wilkinson, J., Willey, D., Williams, G., Williams, L., Williamson, A., Williamson, H., Wilming, L., Woodmansey, R., Wray, P., Yen, J., Zhang, J., Zhou, J., Zoghbi, H., Zorilla, S., Buck, D., Reinhardt, R., Poustka, A., Rosenthal, A., Lehrach, H., Meindl, A., Minx, P., Hillier, L., Willard, H., Wilson, R., Waterston, R., Rice, C., Vaudin, M., Coulson, A., Nelson, D., Weinstock, G., Sulston, J., Durbin, R., Hubbard, T., Gibbs, R., Beck, S., Rogers, J., and Bentley, D. (2005) The dna sequence of the human x chromosome. *Nature,* **434**, 325–337.

[2] Skuse, D. (APR 15, 2005) X-linked genes and mental functioning. *HUMAN MOLECULAR GENETICS,* **14**(Sp. Iss. 1), R27–R32.

[3] Zechner, U., Wilda, M., Kehrer-Sawatzki, H., Vogel, W., Fundele, R., and Hameister, H. (DEC, 2001) A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution?. *TRENDS IN GENETICS,* **17**(12), 697–701.

[4] Graves, J., Koina, E., and SankoviC, N. How the gene content of human sex chromosomes evolved. (2006) Current Opinion In Genetics & Development.

[5] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G.,

Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2008) Database resources of the national center for biotechnology information. *Nucleic Acids Research,* **36**, D13–D21.

[6] Reyniers, E., Vanthienes, M., Meire, F., Boulle, K., Devries, K., Kestelijn, P., and Willems, P. (1995) Gene conversion between red and defective green opsin gene in blue cone monochromacy. *Genomics,* **29**, 323–328.

[7] Hayashi, T., Motulsky, A. G., and Deeb, S. S. (1999 May) Position of a 'green-red' hybrid gene in the visual pigment array determines colour-vision phenotype.. *Nat Genet,* **22**(1), 90–93.

[8] Verrelli, B. and Tishkoff, S. (2004) Signatures of selection and gene conversion associated with human color vision variation. *American Journal of Human Genetics,* **75**, 363–375.

[9] Fitzgerald, J., Dahl, H., Jakobsen, I., and Easteal, S. (1996) Evolution of mammalian x-linked and autosomal pgk and pdh e1 alpha subunit genes. *Molecular Biology and Evolution,* **13**, 1023–1031.

[10] Bagnall, R., Ayres, K., Green, P., and Giannelli, F. (2005) Gene conversion and evolution of xq28 duplicons involved in recurring inversions causing severe hemophilia a. *Genome Research,* **15**, 214–223.

[11] Kouprina, N., Mullokandov, M., Rogozin, I., Collins, N., Solomon, G., Otstot, J., Risinger, J., Koonin, E., Barrett, J., and Larionov, V. (2004) The spanx gene family of cancer/testis-specific antigens: Rapid evolution and amplification in african great apes and hominids. *Proceedings of the National Academy of Sciences of the United States of America,* **101**, 3077–3082.

[12] Kouprina, N., Pavlicek, A., Noskov, V., Solomon, G., Otstot, J., Isaacs, W., Carpten, J., Trent, J., Schleutker, J., Barrett, J., Jurka, J., and Larionov, V. (2005) Dynamic structure of the spanx gene cluster mapped to the prostate cancer susceptibility locus hpcx at xq27. *Genome Research,* **15**, 1477–1486.

[13] Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M., Kitano, H., and Thomas, P. (2005) The panther database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research,* **33**, D284–D288.

[14] Matsuya, A., Sakate, R., Kawahara, Y., Koyanagi, K. O., Sato, Y., Fujii, Y., Yamasaki, C., Habara, T., Nakaoka, H., Todokoro, F., Yamaguchi, K., Endo, T., Oota, S., Makalowski, W., Ikeo, K., Suzuki, Y., Hanada, K., Hashimoto, K., Hirai, M., Iwama, H., Saitou, N., Hiraki, A. T., Jin, L., Kaneko, Y., Kanno, M., Murakami, K., Noda, A. O., Saichi, N., Sanbonmatsu, R., Suzuki, M., Takeda, J.-I., Tanaka, M., Gojobori, T., Imanishi, T., and Itoh, T. (2008) Evola: Ortholog database of all human genes in h-invdb with manual curation of phylogenetic trees. *Nucleic Acids Research,* **36**, D787–D792.

[15] Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., and Searle, S. (2008) Ensembl 2008. *Nucleic Acids Research,* **36**, D707–D714.

[16] Hasegawa, M., Kishino, H., and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.. *J Mol Evol,* **22**(2), 160–74.

[17] Felsenstein, J. (2005) Phylip (phylogeny inference package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle,*.

[18] Sawyer, S. (1989) Statistical tests for detecting gene conversion. *Molecular Biology and Evolution,* **6**(5), 526–538.

[19] Jakobsen, I. B., Wilson, S. R., and Easteal, S. (1997) The partition matrix: Exploring variable phylogenetic signals along nucleotide sequence alignments. *Molecular Biology and Evolution,* **14**(5), 474–484.

[20] Hidaka, K., Caffrey, J. J., Hua, L., Zhang, T., Falck, J. R., Nickel, G. C., Carrel, L., Barnes, L. D., and Shears, S. B. (2002) An adjacent pair of human nudt genes on chromosome x are preferentially expressed in testis and encode two new isoforms of diphosphoinositol polyphosphate phosphohydrolase.. *J Biol Chem,* **277**(36), 32730–32738.

[21] Leslie, N. R., McLennan, A. G., and Safrany, S. T. (2002) Cloning and characterisation of haps1 and haps2, human diadenosine polyphosphate-metabolising nudix hydrolases.. *BMC Biochem,* **3**, 20.

[22] Nagasaki, K., Manabe, T., Hanzawa, H., Maass, N., Tsukada, T., and Yamaguchi, K. (1999) Identification of a novel gene, ldoc1, down-regulated in cancer cell lines.. *Cancer Lett,* **140**(1-2), 227–234.

[23] Greig, G. M., Sharp, C. B., Carrel, L., and Willard, H. F. (1993) Duplicated zinc finger protein genes on the proximal short arm of the human x chromosome: isolation, characterization and x-inactivation studies.. *Hum Mol Genet,* **2**(10), 1611–1618.

[24] Marais, G. (2003) Biased gene conversion: implications for genome and sex evolution.. *Trends Genet,* **19**(6), 330–8.

[25] Birdsell, J. A. (2002) Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution.. *Mol Biol Evol,* **19**(7), 1181–97.

[26] Galtier, N. and Duret, L. (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution.. *Trends Genet,* **23**(6), 273–277.

[27] Ardlie, K., Liu-Cordero, S. N., Eberle, M. A., Daly, M., Barrett, J., Winchester, E., Lander, E. S., and Kruglyak, L. (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion.. *Am J Hum Genet,* **69**(3), 582–9.

[28] Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., and Di Rienzo, A. (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels.. *Am J Hum Genet,* **69**(4), 831–43.

[29] Lazzaro, B. P. and Clark, A. G. (2001) Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the Attacin genes of Drosophila melanogaster.. *Genetics,* **159**(2), 659–71.

[30] Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002) Recent segmental duplications in the human genome.. *Science,* **297**(5583), 1003–7.

[31] Zhang, L. Q., Lu, H. H. S., Chung, W. Y., Yang, J., and Li, W. H. (2005) Patterns of segmental duplication in the human genome. *Molecular Biology and Evolution,* **22**(1), 135–141.

[32] Zhang, L. (2008 Jun) Adaptive evolution and frequent gene conversion in the brain expressed x-linked gene family in mammals.. *Biochem Genet,* **46**(5-6), 293–311.

[33] Alvarez, E., Zhou, W., Witta, S. E., and Freed, C. R. (2005) Characterization of the Bex gene family in humans, mice, and rats.. *Gene,* **357**(1), 18–28.

[34] Saito-Ohara, F., Fukuda, Y., Ito, M., Agarwala, K. L., Hayashi, M., Matsuo, M., Imoto, I., Yamakawa, K., Nakamura, Y., and Inazawa, J. (2002) The xq22 inversion breakpoint interrupted a novel ras-like gtpase gene in a patient with duchenne muscular dystrophy and profound mental retardation.. *Am J Hum Genet,* **71**(3), 637–645.

[35] Maizels, N. (2005) Immunoglobulin gene diversification. *Annual Review of Genetics,* **39**, 23–46.

[36] Chen, J. M., Cooper, D. N., Chuzhanova, N., Ferec, C., and Patrinos, G. P. (2007) Gene conversion: mechanisms, evolution and human disease.. *Nat Rev Genet,* **8**(10), 762–75.

[37] Pan, D. and Zhang, L. (2007) Quantifying the major mechanisms of recent gene duplications in the human and mouse genomes: a novel strategy to estimate gene duplication rates.. *Genome Biol,* **8**(8), R158.

[38] Lawson, M. J., Heath, L. S., Ramakrishnan, N., and Zhang, L. (2008) Using cost-sensitive learning to determine gene conversions. In Huang, D.-S., II, D. C. W., Levine, D. S., and Jo, K.-H., (eds.), *ICIC (2)*, Vol. 5227, of Lecture Notes in Computer Science pp. 1030–1038.

[39] Winter, E. and Ponting, C. (2005) Mammalian bex, wex and gasp genes: Coding and non-coding chimaerism sustained by gene conversion events. *Bmc Evolutionary Biology,* **5**.

[40] Zhang, L. (2008) Adaptive evolution and frequent gene conversion in the brain expressed X-linked gene family in mammals.. *Biochem Genet,* **46**(5-6), 293–311.

[41] Schienman, J. E., Holt, R. A., Auerbach, M. R., and Stewart, C. B. (2006) Duplication and divergence of 2 distinct pancreatic ribonuclease genes in leaf-eating African and Asian colobine monkeys.. *Mol Biol Evol,* **23**(8), 1465–79.

7 Figure Legends

- Figure 1: PTHR14754 Phylogenetic Tree
- Figure 2: PTHR14754 Synteny
- Figure 3: PTHR19860:SF2 Phylogenetic Tree
- Figure 4: PTHR11431:SF14 Phylogenetic Tree
- Figure 5: PTHR11708:SF146 Phylogenetic Tree
- Figure 6: PTHR12629 Phylogenetic Tree
- Figure 7: PTHR15503 Phylogenetic Tree
- Figure 8: PTHR19818:SF18 Phylogenetic Tree