

Research article

Open Access

Genome-wide analysis of chimpanzee genes with premature termination codons

Anna Wetterbom*¹, Ulf Gyllensten¹, Lucia Cavelier^{†1} and Tomas F Bergström^{†2}

Address: ¹Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, SE-751 85, Sweden and ²Department of Animal breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, SE-750 07, Sweden

Email: Anna Wetterbom* - anna.wetterbom@genpat.uu.se; Ulf Gyllensten - ulf.gyllensten@genpat.uu.se; Lucia Cavelier - lucia.cavelier@genpat.uu.se; Tomas F Bergström - tomas.bergstrom@hgen.slu.se

* Corresponding author †Equal contributors

Published: 29 January 2009

Received: 10 November 2008

BMC Genomics 2009, **10**:56 doi:10.1186/1471-2164-10-56

Accepted: 29 January 2009

This article is available from: <http://www.biomedcentral.com/1471-2164/10/56>

© 2009 Wetterbom et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Premature termination codons (PTCs) cause mRNA degradation or a truncated protein and thereby contribute to the transcriptome and proteome divergence between species. Here we present the first genome-wide study of PTCs in the chimpanzee. By comparing the human and chimpanzee genome sequences we identify and characterize genes with PTCs, in order to understand the contribution of these mutations to the transcriptome diversity between the species.

Results: We have studied a total of 13,487 human-chimpanzee gene pairs and found that ~8% were affected by PTCs in the chimpanzee. A majority (764/1,109) of PTCs were caused by insertions or deletions and the remaining part was caused by substitutions. The distribution of PTC genes varied between chromosomes, with Y having the highest proportion. Furthermore, the density of PTC genes varied on a megabasepair scale within chromosomes and we found the density to be correlated both with indel divergence and proximity to the telomere. Within genes, PTCs were more common close to the 5' and 3' ends of the amino acid sequence. Gene Ontology classification revealed that olfactory receptor genes were over represented among the PTC genes.

Conclusion: Our results showed that the density of PTC genes fluctuated across the genome depending on the local genomic context. PTCs were preferentially located in the terminal parts of the transcript, which generally have a lower frequency of functional domains, indicating that selection was operating against PTCs at sites central to protein function. The enrichment of GO terms associated with olfaction suggests that PTCs may have influenced the difference in the repertoire of olfactory genes between humans and chimpanzees. In summary, 8% of the chimpanzee genes were affected by PTCs and this type of variation is likely to have an important effect on the transcript and proteomic divergence between humans and chimpanzees.

Background

The genetic basis of the observed phenotypic divergence between humans and chimpanzees has since long

intrigued scientists. Since humans and chimpanzees diverged 5–7 million years ago [1,2] their genomes have acquired a multitude of lineage specific mutations,

including nucleotide substitutions, insertions and deletions (indels), duplications and inversions. The sequence divergence caused by substitutions and indels has been estimated to 1.2–1.5% [3-5] and 3–7% [3,4,6], respectively. Several hypotheses have been proposed to reconcile the observed overall low sequence divergence with the large phenotypic differences between humans and chimpanzees. These hypotheses include sequence divergence of protein coding genes [7], gain and loss of genes [8-11], differential gene expression [12-14] and divergent patterns of alternative splicing [15]. In this study we have focused on the role of nonsense mutations and frame shift indels that cause premature termination codons (PTCs) in chimpanzee protein coding genes. The dramatic effect of this type of mutations has been described in a number of human genes [16-18].

PTCs have the capability of changing the pattern of alternative splicing and expressed protein isoforms, since mRNAs affected by PTCs will either be degraded by nonsense-mediated decay (NMD) or translated into truncated protein sequences [19,20]. The coupling of PTCs and NMD has previously been suggested as a novel mechanism to regulate gene expression [21-24].

Although the occurrence of genes with premature stop codons in human genes has been described on a genome wide scale [25,26] there is no parallel study in the chimpanzee. Two previous studies [4,5] have examined PTCs in chimpanzee genes using the high quality genome sequence of chromosome 21 (previously referred to as chromosome 22 in the chimpanzee [27]). In the initial publication of the chimpanzee genome sequence the identification of PTCs was hampered by the low quality of the sequence [3]. By using the current (6×) chimpanzee assembly and stringent quality criteria we have overcome this problem and here we report the first genome wide study of PTCs in the chimpanzee. Chimpanzee specific PTCs were identified in 1,109 genes, which were further studied with respect to their genome location and context, biological function and the position of the PTC within genes.

Results and discussion

Detection of premature termination codons in chimpanzee protein-coding genes

In an attempt to identify chimpanzee genes with PTCs we have analyzed pairs of human-chimpanzee genes. Annotations of all human protein coding genes were collected from the Ensemble database [28], resulting in a total of 21,021 genes and 45,455 associated transcripts. Chimpanzee genes and transcripts were inferred from human annotations and human exon coordinates were translated to the chimpanzee genome using the liftOver tool (available from the UCSC Genome Browser [28]). Translation

between the genomes depends on a pairwise alignment (also available from the UCSC Genome Browser [28]) and we required all chimpanzee exons in a transcript to have the same orientation and be located on the same chromosome after the translation. Failure to translate coordinates accurately may be due to e.g. lack of sequence in a region or problems in aligning regions with large indels or lots of repeats. Both human and chimpanzee exon sequences were concatenated into complete coding sequences, translated into amino acid sequences and then screened for PTCs in the chimpanzee. Several filtering steps were employed to ensure that the detected PTCs were not a result of incorrect gene predictions or low quality of the chimpanzee assembly (see Figure 1 and the Methods section for details). Gene predictions were required to have correct start and stop codons in human and a correct start codon in chimpanzee. To account for the lower quality of the chimpanzee assembly we required the region surrounding a PTC to have a quality score > 40, corresponding to less than 1 sequencing error in 10,000 bp.

After filtering, a total of 13,487 genes and 26,779 associated transcripts remained. The genes were divided into two datasets: (i) PTC genes where one or several of the associated chimpanzee transcripts were affected by PTCs, and (ii) non-PTC genes where none of the annotated transcripts were affected by PTCs in the chimpanzee. In the present study all analyses will be based on genes instead of transcripts, thereby avoiding the bias introduced by different genes having different numbers of transcripts.

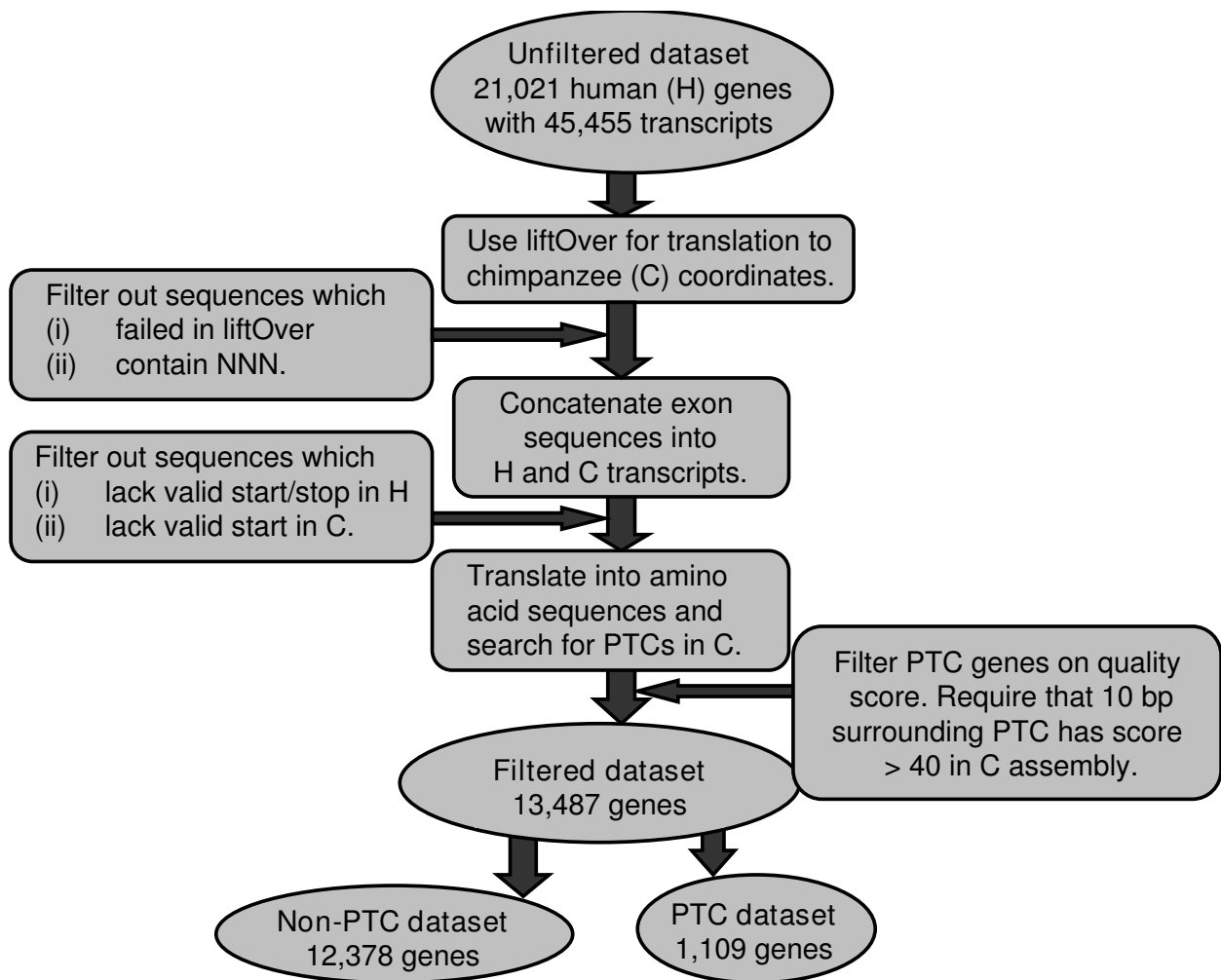
Approximately 8% (1,109 of 13,487) of the chimpanzee genes were affected by PTCs, indicating that the genes were either pseudo genes in the chimpanzee (in cases where all transcripts had PTCs) or that the genes had a different transcriptional pattern as compared to the corresponding human genes (in cases where the genes had both functional transcripts and transcripts with PTCs).

Furthermore, we determined the type of mutational event leading to the PTC. Indels were found to be the causative mutation in approximately 70% of the PTC genes ($n = 764$) and in the remaining 30% ($n = 345$) of the genes the PTC was caused by a substitution. The exact location of the premature termination codons and the type of mutational event is detailed in the Additional file 1 (Table S1).

Characteristics of genes with PTCs in the chimpanzee

The density of PTC genes varies both between and within chromosomes

An estimated 8% of all chimpanzee genes were affected by PTCs. This figure is higher than the 5% we reported in a previous study of chromosome 21 [4] and the discrepancy is likely to be due to differences in density of PTC genes between chromosomes (Figure 2). In the present study,

**Figure 1**

Outline of the filtering process. To obtain reliable predictions of PTCs in chimpanzee genes we applied a series of different filtering steps to remove genes where the PTCs were caused by unreliable gene annotations, lack of sequence or low quality of the chimpanzee assembly. A majority of the removed genes were filtered out because they failed to be correctly translated between the genomes (i.e. the liftOver step).

the proportion of PTC genes varied from 5 to 16% for the autosomes, and a significant difference between chromosomes was observed (Wilcoxon test, $p < 6 \cdot 10^{-8}$). The Y chromosome had a strikingly high percentage (> 40%) of PTC genes. The Y-chromosome is known to have an increased proportion of inactivating mutations in the chimpanzee lineage and a large number of pseudo genes [29,30], thereby explaining the increased fraction of PTC genes observed in this study. The proportion of PTC genes was next estimated for 1 Mbp-windows across all chromosomes and a significant variation between different chromosomal regions was observed (Wilcoxon test, $p < 3 \cdot 10^{-16}$). This suggests that the large-scale chromosomal structure influences the density of PTC genes, although the cause for this effect is unclear.

The proportion of PTC genes was correlated to both indel divergence and proximity to the telomere

Having noticed that the proportion of PTC genes varied both between chromosomes and between 1 Mb windows from the same chromosome, we sought to understand the underlying reasons. The density of PTC genes in a particular region may be affected by several genomic properties such as GC-content, presence of repeats, CpG-islands, segmental duplications, substitutions, indel divergence or the distance to the telomere or centromere. To study the relationship between the density of PTC genes and these genomic properties, we divided the chromosomes into non-overlapping 1 Mbp windows and applied a regression model, evaluating the relationship between the fraction of genes in the two datasets (PTC and non-PTC

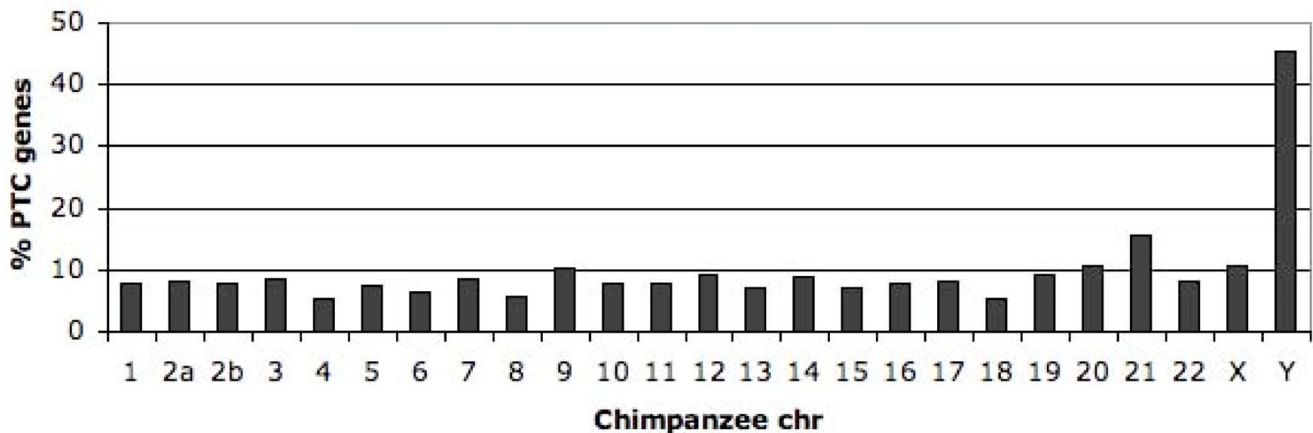


Figure 2
Chromosomal distribution of PTC genes. The x-axis denotes the chimpanzee chromosomes and the y-axis represents the % of genes that were affected by PTCs on a specific chromosome.

genes) and each genomic feature. Occurrence of PTC genes was found to be significantly associated with both indel divergence and distance to the telomere. The other genomic properties showed no correlation with the occurrence of PTC genes, or a similar correlation to non-PTC genes. The results indicate that PTC genes were in most respects similar to non-PTC genes.

The correlation with indel divergence (calculated as the % of bp located in indels) was higher for the PTC genes (Table 1) as compared to the non-PTC genes. The observed association was significant but rather weak, possibly as a result of the genomic alignment obtained from UCSC [28]. The indel divergence estimated from the alignment was only 0.64%, which was considerably lower than the 3–7% previously reported [3,4,6]. This is explained by the alignment parameters having been optimized not to allow for gaps longer than 100 bp. Although the majority of indels are shorter than 100 bp [3], longer indels contribute more to the indel divergence since they include more base pairs. Thus, given that indels > 100 bps were not taken into account, the correlation seen between PTC genes and indel divergence was most likely underestimated. A correlation between PTC genes and indel divergence was further supported by the fact that approximately 70% of the PTCs were actually caused by indel events.

In addition to indel divergence, the proportion of PTC genes in a region increased with proximity to the telomere (Table 1). The closer to the telomere, the more likely the gene was to be affected by a PTC. Telomeres are known to have an increased indel divergence [3] and therefore we combined these two variables in a multiple regression model (Table 1). The combined model has a higher R-value than the previous regression analyses, suggesting that the two variables should be considered together.

PTCs were preferentially located in distal parts of the gene and in regions with less functional domains

To examine the relative location of PTCs within affected genes, each gene was partitioned into windows covering 5% of the length, and the proportion of genes with PTCs in a specific region was estimated. The results showed a higher frequency of PTCs in the distal parts of the genes than in the central region. When a PTC occurs early in the transcript, the nucleotide sequence will most likely not be translated into a protein but instead be subjected to NMD [19,20]. On the other hand, when the PTC occurs late in the transcript, translation it is likely to take place and result in a truncated protein sequence. To further investigate the consequences of PTCs we searched for a relationship between the location of PTCs and functional domains of the gene, using the Pfam database [31] to map functional domains onto the PTC genes. All PTC genes

Table 1: Summary of the regression analyses: predictor(s) of the density of PTC genes.

Predictor(s) of PTC gene density	Adjusted R	p-value
Distance to telomere	0.137	4*10 ⁻¹⁵
Indel divergence (% bp in indels)	0.087	2*10 ⁻⁶
Combined model (telomeric distance and indel divergence)	0.165	3*10 ⁻¹⁶

were matched against the Pfam database [31] and the relative location of Pfam hits was recorded. A negative correlation ($r = -0.75$, $p\text{-value} < 0.0002$) was found between the location of PTCs and the location of Pfam matches (Figure 3), suggesting a stronger selection against PTC in functional regions of genes. The results support the assumption that PTCs may be tolerable as long as the termination codon does not disrupt the structure of centrally located functional domains in the protein.

Olfactory receptor genes were overrepresented in the PTC dataset
 The biological functionality of genes in the two datasets was assessed using Gene Ontology [32] classifications. All three ontologies; biological process, cellular component and molecular function were used for the classification and the FatiGO tool [33] was used to test if any GO classes were overrepresented in the PTC dataset compared to the non-PTC dataset. We found that PTC genes were significantly overrepresented in three GO classes: olfactory receptor activity, sensory perception of smell and G-protein coupled receptor protein signaling pathway (see Table 2 for details). These GO terms are to a large extent overlapping, and 46 of the PTC genes were found in all three GO classes (listed in Additional file 1, Table S2).

All PTC genes mapping to the GO terms olfactory receptor activity and sensory perception of smell belong to the olfactory receptor (OR) superfamily of genes. There are almost one thousand OR genes known in human, with ~50% of them being pseudogenes [34,35]. The OR genes belong to the G-protein coupled receptor (GPCR) hyperfamily, explaining the presence of G-protein coupled receptor protein signaling pathway in the GO classification. It is generally thought that the number of functional OR genes have been reduced in the primate lineage as compared to other mammals, and different selective mechanisms have been proposed [35-38]. A recent study [39] suggested that humans and chimpanzees have simi-

lar numbers of OR genes and also comparable fractions of pseudogenes. The repertoire of OR genes expressed has been suggested to differ between species, allowing for species-specific adaptations of odor perception [39].

Among the 46 PTC genes that were overrepresented in all three GO categories (Table 2 and Table S2), all but two have only a single annotated transcript. This implies that these genes are pseudogenes in the chimpanzee, since all known transcripts have been silenced. For the two genes with several annotated transcripts, the effect may be different. Most likely these genes were not completely switched off but rather the range of available transcripts was altered.

The effect of PTCs on alternative splicing

A surprisingly high proportion of the genes in the study were found to have PTCs, but not all affected genes represented pseudogenes. Approximately half of the genes in the study have multiple annotated transcripts and the same holds true for the PTC dataset. The majority of these genes have at least one transcript unaffected by the PTC, suggesting that most PTC mutations do not switch off the gene, but rather a specific transcript. Furthermore, the Ensembl prediction that ~50% of the genes have multiple splice variants is likely to be an underestimate since recent studies have shown that nearly all human multiexon genes have multiple splice variants [40,41]. Thus it is reasonable to assume that virtually none of the multiexon genes with PTCs in chimpanzee were entirely silenced by the mutation.

Depending on the location of the PTC, the affected mRNA may either be translated into a truncated protein product or become degraded by NMD [19,20]. The coupling of PTCs and NMD has previously been proposed as a novel mechanism for regulating gene expression [21-24]. By selectively silencing transcripts, the coupled action of PTCs and NMD may modulate gene expression and alter

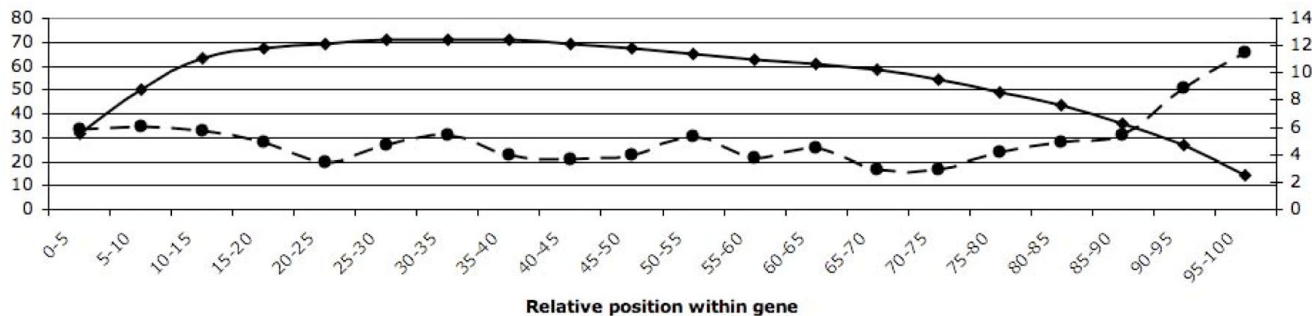


Figure 3
Location of PTCs and Pfam matches within PTC genes. The x-axis represents the relative position within the gene (0–5% of the length and so on), the left y-axis shows the % of all PTC genes with a Pfam match at a specific position (solid line and diamonds) and the right y-axis shows the % of all PTC genes with a PTC in the same interval (dashed line and dots).

Table 2: Gene ontology classification of PTC genes.

Ontology	Overrepresented GO term in the PTC dataset	Number of PTC genes in the GO category*	Adjusted p-value
Molecular function	Olfactory receptor activity (GO:0004984)	46	1×10^{-6}
Biological process	Sensory perception of smell (GO:0007608)	46	3×10^{-7}
Biological process	G-protein coupled receptor protein signaling pathway (GO:0007186)	82	2×10^{-5}

* Note that the GO terms overlap. The two groups olfactory receptor activity and sensory perception of smell are identical and these 46 genes are in turn included in the group G-protein coupled receptor activity.

the repertoire of expressed isoforms, thereby providing for species-specific patterns of alternative splicing.

Since there are no *de novo* annotations of chimpanzee genes and transcripts, the present study is based solely on human annotations [42]. It is possible that in some cases where we predict a PTC in a chimpanzee gene, the chimpanzee may instead have different gene structure and/or different splice variants, thereby adding to the transcriptional complexity. To study this in more detail it would be necessary to perform large-scale sequencing of the chimpanzee transcriptome.

Conclusion

We have performed a genome wide study of premature termination codons (PTCs) in the chimpanzee to estimate the frequency of such events and further characterize the affected genes. We found that ~8% (1,109/13,487) of the genes had at least one transcript affected by a PTC. Indels rather than substitutions were the main cause of PTCs. We observed both inter- and intra-chromosomal fluctuation in the density of PTC genes and this variation was related both to local variations in indel divergence and proximity to the telomere. Within genes, PTCs occurred towards the 5' and 3' ends of the genes, thereby preserving functional domains in the central part of the genes. This indicates that selection against PTCs was stronger in the central and more conserved functional parts of proteins.

Gene Ontology classification revealed that PTC genes were overrepresented in the groups associated with olfaction. This was especially intriguing since the repertoire of olfactory receptors in primates has been extensively studied and it has been shown that many olfactory receptor genes have become pseudo genes in the primate lineage [35-37,39]. Premature termination codons are most likely a major factor in the pseudogenisation process.

Approximately half of the genes in this study had multiple annotated transcripts and in most cases the PTC did not affect all of the transcripts. Instead of becoming pseudo genes, the affected genes seemed to have certain transcripts silenced by the PTCs. Transcripts affected by PTCs will either be degraded by NMD or produce truncated protein isoforms. Either way, the PTCs result in a different a

transcript repertoire between humans and chimpanzees and PTCs may therefore contribute to the factors determining the phenotypic differences between the species. The phenotypic differences between humans and chimpanzees has been attributed to divergence in protein coding genes [7], gene expression [12-14] and alternative splicing [15] as well as gain and loss of genetic material [8-11]. Indels and substitutions causing PTCs potentially affect several of these mechanisms, in particular the pattern of gene expression and alternative splicing. Moreover, the paired action of PTCs and nonsense-mediated decay has been suggested as a novel mechanism to regulate gene expression [21-24] and the observed frequent occurrence of PTC mutations further support the hypothesis that divergent gene expression and alternative splicing affect the phenotypic divergence.

Methods

Selection and filtering of transcripts

Genomic sequence for human (hg18/build 36) and chimpanzee (panTro2) were downloaded from the Ensembl database and annotations of all human protein coding genes and transcripts were likewise obtained from Ensembl (release 50) [42]. Chimpanzee genes and transcripts were inferred from human gene annotations and the coordinates were translated between the genomes using a command line version of liftOver (available from the UCSC Genome Browser [28]). After translation with liftOver, all chimpanzee exons from the same transcript were required to be both on the same chromosome and on the same strand, or else the transcript was removed from further analysis. Exon sequences were concatenated into complete coding sequences for both human and chimpanzee, translated to amino acid sequences and scanned for PTCs in the chimpanzee. To ensure the correctness of the predicted PTCs we applied a number of additional filtering steps (outlined in Figure 1).

(i) All transcripts with Ns in the chimpanzee sequence were removed to exclude regions with low sequence coverage.

(ii) All transcripts with PTCs in human were removed and likewise for transcripts lacking a valid start (M or L) or stop codon in human. The Ensembl set of genes is based

on automatic annotation and a minority of transcripts violates the above requirements. This group of transcripts tends to have fewer supporting mRNA/EST sequences and was therefore considered less reliable.

(iii) Transcripts lacking a valid start codon in the chimpanzee were removed since such transcripts were not likely to be coding.

(iv) To account for the varying quality of chimpanzee genome assembly, all transcripts with predicted PTCs were filtered on the sequence quality score. The quality score-track from the UCSC Genome Browser [28] was downloaded and we used a threshold of 40, which corresponds to less than 1 sequencing error per 10,000 bp. The substitution or indel event causing the PTC was determined for all PTC transcripts and we required that a 10 bp region (5 bp upstream and 5 bp downstream) surrounding the event had a quality score > 40. To determine the type of mutational event leading to the PTC, all affected transcripts were first scanned for substitutions. If a non-sense mutation coincided with the predicted PTC, this substitution was regarded as the causative event. Otherwise, the PTC must be caused by an indel event and the first indel in the transcript was inferred to be the cause of the subsequent PTC.

The aim of the filtering process was to make sure that the PTCs were not caused by incorrect gene predictions, lacking sequence or bad quality of the chimpanzee assembly. Furthermore, the filtering removed transcripts with speculative coding capacity. All chimpanzee PTCs that remained after the filtering procedure were considered to be true.

Description of the two datasets

The filtering process resulted in two datasets, one group of genes with predicted PTCs in the chimpanzee (the PTC dataset) and one group with no predicted PTCs in the chimpanzee (the non-PTC dataset). The two datasets could be analyzed either at the gene level or at the transcript level and we have chosen the first approach. It is likely that multiple transcripts from a gene will cover almost the same genomic regions and therefore the analyses was done on the gene level to avoid the bias introduced if we sample the same genomic region several times. Approximately half of the genes have more than one annotated transcript and we randomly chose one transcript per gene for the analyses. However, for genes in the PTC-dataset the chosen transcript was required to have a PTC.

Characterizing genes with PTCs in chimpanzee

Annotation of genomic features

To determine if chimpanzee PTC genes share some genomic properties we studied a number of genomic fea-

tures such as GC-content, repeat content, CpG-islands, segmental duplications, sequence divergence and proximity to the centromere and telomere. Since the studied PTCs occur in the chimpanzee we used annotations based on the chimpanzee genome. GC-content was calculated directly from the chimpanzee genome sequence. Coordinates for CpG-islands and repeats in the chimpanzee genome were downloaded from the UCSC Genome browser (CpG-island track and RepeatMasker track) [28]. Annotations of chimpanzee segmental duplications were obtained from <http://chimpparalogy.gs.washington.edu>. In this case the segmental duplications had been mapped to an old version of the human genome (hg16, build 34) and liftOver was used to translate the coordinates to the most recent chimpanzee genome. Sequence divergence caused by substitutions and indels was estimated from the pairwise alignment of the human and chimpanzee genomes, also downloaded from UCSC. Substitutions were counted directly from the alignment whereas the indel divergence was estimated in two ways: (i) as the number of indel events or (ii) the number of bp involved in indels (i.e. the number of bp in indels/the number of aligned bp).

Regression analysis and choice of window size

In order to investigate regional variation in the density of PTC genes we first estimated the proportion of PTC genes, as well as non-PTC genes, in windows across all chromosomes. To determine the genomic location of genes we used the start coordinates, and genes were counted once in the window where they started. The number of PTC genes, and non-PTC genes, genes in a window was normalized with the total number of genes in the same window. In addition, we analyzed a number of genomic features over the same windows to assess a possible correlation with genes in the two datasets. GC-content, substitutions and indel divergence (calculated as the proportion of bp located in indels) were estimated on the nucleotide levels and thus the occurrence of these features was simply counted within each window. CpG-islands, repeats, segmental duplications and indel events on the other hand span over several nucleotides and in this case the feature was counted once, in the window where the start was located. RepeatMasker [43] classifies repeats into ten different classes and we performed the analyses both with separate repeat classes and with a merged set of all repeats. There were no annotations of the location of chimpanzee centromeres available so the positions were estimated from the pairwise alignment files previously used to determine the sequence divergence. The centromere was inferred to the 1 Mbp-window with zero aligned nucleotides and if several consecutive windows with zero aligned bp occurred the centromere was placed in the middle.

After estimating the proportion of genes in the two datasets, as well as a number of genomic features, we applied a linear regression model. The relationship between each genomic feature and the proportion of genes in the two datasets was evaluated separately. Variables with a higher correlation to PTC genes than to non-PTC genes were kept and included in a multiple linear regression model. All statistical calculations were done in R [44].

Several different window sizes were evaluated and we settled for 1 Mbp intervals, which has also previously proved to be a reasonable choice for this type of analyses [45]. It may be hypothesized that shorter windows would improve the resolution in the analyses but this was not the case in the present study. Shortening the window length obviously increased the number of windows and resulted in more windows with zero genes in them. We also tried to increase the window size but this did not improve the analysis either. Probably, because larger windows tend to 'average out' the differences between different genomic regions.

Location of PTCs within the affected gene and analysis of a possible association with functional protein domains

The exact position of PTCs within the genes was determined and the locations were then normalized with the gene length to be able to compare positions across all genes. The normalized positions of PTCs were merged into relative intervals (0–5%, 5–10% etc of the gene length). Next, the translated sequences of all PTC genes were searched against the Pfam database [31] to examine a possible association between the location of PTCs and the location of functional domains in the protein. The positions of all Pfam matches within a sequence were recorded and the matches were filtered using an e-value of 0.001 as cut off (other cut offs were tried and the results were similar, data not shown). The locations of Pfam matches were calculated in relative intervals (0–5%, 5–10% and so on), similar to the location of PTCs in genes, and matches spanning over several intervals were recorded once in each interval. Finally, we used R [44] to calculate the correlation between PTC positions and functional domains.

Gene Ontology classification

The FatiGO tool [33] was used to test for over/under representation of specific GO classes in the PTC dataset, as compared to the non-PTC dataset. All three GO categories [32] (biological process, cellular component and molecular function) were tested.

Abbreviations

Indel: insertion or deletion of nucleotides in a sequence; Mbp: megabasepairs; NMD: nonsense mediated decay; PTC: premature termination codon.

Authors' contributions

AW designed the study, performed the analysis and statistics and drafted the paper. UG, LC and TB helped in the design of the study and in writing of the paper. All authors read and approved the final manuscript.

Additional material

Additional file 1

Genes affected by PTCs. The file contains Table S1 which includes the complete PTC gene dataset and Table S2 with the 46 genes found in all three GO categories that were overrepresented in the PTC dataset.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-56-S1.xls>]

Acknowledgements

This work was supported by grants from the Graduate Research School in Genomics and Bioinformatics, the Swedish Research Council in Medicine, the Swedish Research Council in Natural Sciences, the Marcus Borgström Foundation and the Magnus Bergwall Foundation.

References

1. Glazko GV, Nei M: **Estimation of divergence times for major lineages of primate species.** *Molecular biology and evolution* 2003, **20(3)**:424-434.
2. Kumar S, Filipowski A, Swarna V, Walker A, Hedges SB: **Placing confidence limits on the molecular age of the human-chimpanzee divergence.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102(52)**:18842-18847.
3. The Chimpanzee Sequencing and Analysis Consortium: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437(7055)**:69-87.
4. Wetterbom A, Sevov M, Cavellier L, Bergstrom TF: **Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution.** *J Mol Evol* 2006, **63(5)**:682-690.
5. Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, Kuroki Y, Noguchi H, BenKahla A, Lehrach H, Sudbrak R, Kube M, Taenzer S, Galgoczy P, Platzer M, Scharfe M, Nordsiek G, Blöcker H, Hellmann I, Khaitovich P, Pääbo S, Reinhardt R, Zheng H-J, Zhang XL, Zhu GF, Wang BF, Fu G, Ren SX, Zhao GP, Chen Z, Lee YS, Cheong JE, Choi SH, Wu KM, Liu TT, Hsiao KJ, Tsai SF, Kim CG, Oota S, Kitano T, Kohara Y, Saitou N, Park HS, Wang SY, Yaspo ML, Sakaki Y: **DNA sequence and comparative analysis of chimpanzee chromosome 22.** *Nature* 2004, **429(6990)**:382-388.
6. Britten RJ: **Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99(21)**:13633-13635.
7. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG: **Natural selection on protein-coding genes in the human genome.** *Nature* 2005, **437(7062)**:1153-1157.
8. Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, Karimpour-Fard A, Glueck D, McGavran L, Berry R, Pollack J, Sikela JM: **Lineage-specific gene duplication and loss in human and great ape evolution.** *PLoS biology* 2004, **2(7)**:E207.
9. Ohno S: **Evolution by Gene Duplication.** Springer-Verlag; 1970.
10. Olson MV: **When less is more: gene loss as an engine of evolutionary change.** *American journal of human genetics* 1999, **64(1)**:18-23.
11. Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, Haussler D: **Comparative Genomics Search for Losses of Long-Estab-**

- lished Genes on the Human Lineage. *PLoS computational biology* 2007, **3**(12):e247.
12. King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees.** *Science* 1975, **188**(4184):107-116.
 13. Preuss TM, Caceres M, Oldham MC, Geschwind DH: **Human brain evolution: insights from microarrays.** *Nature reviews* 2004, **5**(11):850-860.
 14. Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S: **Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees.** *Science* 2005, **309**(5742):1850-1854.
 15. Calarco JA, Xing Y, Caceres M, Calarco JP, Xiao X, Pan Q, Lee C, Preuss TM, Blencowe BJ: **Global analysis of alternative splicing differences between humans and chimpanzees.** *Genes & development* 2007, **21**(22):2963-2975.
 16. Winter H, Langbein L, Krawczak M, Cooper DN, Jave-Suarez LF, Rogers MA, Praetzel S, Heidt PJ, Schweizer J: **Human type I hair keratin pseudogene phihHaA has functional orthologs in the chimpanzee and gorilla: evidence for recent inactivation of the human gene after the Pan-Homo divergence.** *Human genetics* 2001, **108**(1):37-42.
 17. Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA: **Myosin gene mutation correlates with anatomical changes in the human lineage.** *Nature* 2004, **428**(6981):415-418.
 18. Dolphin CT, Beckett DJ, Janmohamed A, Cullingford TE, Smith RL, Shephard EA, Phillips IR: **The flavin-containing monooxygenase 2 gene (FMO2) of humans, but not of other primates, encodes a truncated, nonfunctional protein.** *The Journal of biological chemistry* 1998, **273**(46):30599-30607.
 19. Chang YF, Imam JS, Wilkinson MF: **The nonsense-mediated decay RNA surveillance pathway.** *Annual review of biochemistry* 2007, **76**:51-74.
 20. Maquat LE: **Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics.** *Nat Rev Mol Cell Biol* 2004, **5**(2):89-99.
 21. Lewis BP, Green RE, Brenner SE: **Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(1):189-192.
 22. Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC: **Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutates genomic noise.** *Nature genetics* 2004, **36**(10):1073-1078.
 23. Hillman RT, Green RE, Brenner SE: **An unappreciated role for RNA surveillance.** *Genome biology* 2004, **5**(2):R8.
 24. Saltzman AL, Kim YK, Pan Q, Fagnani MM, Maquat LE, Blencowe BJ: **Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay.** *Mol Cell Biol* 2008, **28**(13):4320-4330.
 25. Hahn Y, Lee B: **Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences.** *Bioinformatics (Oxford, England)* 2005, **21**(Suppl 1):i186-194.
 26. Hahn Y, Lee B: **Human-specific nonsense mutations identified by genome sequence comparisons.** *Human genetics* 2006, **119**(1-2):169-178.
 27. McConkey EH: **Orthologous numbering of great ape and human chromosomes is essential for comparative genomics.** *Cytogenetic and genome research* 2004, **105**(1):157-158.
 28. **The UCSC Genome Browser** [<http://genome.ucsc.edu/>]
 29. Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, Rozen S, Wilson RK, Page DC: **Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee.** *Nature* 2005, **437**(7055):100-103.
 30. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou S-F, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang S-P, Waterston RH, Wilson RK, Rozen S, Page David C: **The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.** *Nature* 2003, **423**(6942):825-837.
 31. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database.** *Nucleic acids research* 2008:D281-288.
 32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Rinwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**(1):25-29.
 33. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics (Oxford, England)* 2004, **20**(4):578-580.
 34. Glusman G, Yanai I, Rubin I, Lancet D: **The complete human olfactory subgenome.** *Genome research* 2001, **11**(5):685-702.
 35. Niimura Y, Nei M: **Extensive gains and losses of olfactory receptor genes in mammalian evolution.** *PLoS ONE* 2007, **2**(1):e708.
 36. Gilad Y, Man O, Glusman G: **A comparison of the human and chimpanzee olfactory receptor gene repertoires.** *Genome research* 2005, **15**(2):224-230.
 37. Gimelbrant AA, Skaletsky H, Chess A: **Selective pressures on the olfactory receptor repertoire since the human-chimpanzee divergence.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(24):9019-9022.
 38. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, Sninsky JJ, Adams MD, Cargill M: **A scan for positively selected genes in the genomes of humans and chimpanzees.** *PLoS biology* 2005, **3**(6):e170.
 39. Go Y, Niimura Y: **Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees.** *Molecular biology and evolution* 2008, **25**(9):1897-1907.
 40. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nature genetics* 2008.
 41. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456**(7221):470-476.
 42. **The Ensembl Database** [<http://www.ensembl.org>]
 43. Smit A, Hubley R, Green P: **RepeatMasker Open-3.0.** 1996 [<http://www.repeatmasker.org>].
 44. **The R-project** [<http://www.r-project.org/>]
 45. Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD: **A macaque's-eye view of human insertions and deletions: differences in mechanisms.** *PLoS computational biology* 2007, **3**(9):1772-1782.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

