Triangulation of the human, chimpanzee and Neanderthal genome sequences identifies potentially compensated mutations

Guojie Zhang, Pei Zhang, Michael Krawczak, Edward V. Ball, Matthew Mort, Hildegard Kehrer-Sawatzki, David N. Cooper

Human Mutation

Triangulation of the human, chimpanzee and Neanderthal genome sequences identifies potentially compensated mutations

Journal:	*Human Mutation*
Manuscript ID:	humu-2010-0442.R1
Wiley - Manuscript type:	Rapid Communication
Date Submitted by the Author:	06-Oct-2010
Complete List of Authors:	Zhang, Guojie; Beijing Genomics Institute at Shenzhen, Bioinformatics Department Zhang, Pei; Beijing Genomics Institute at Shenzhen, Bioinformatics Department Krawczak, Michael; Christian-Albrechts-Universität zu Kiel, Institut für Medizinische Informatik und Statistik Ball, Edward; Cardiff University, Institute of Medical Genetics, College of Medicine Mort, Matthew; Cardiff University, Institute of Medical Genetics, College of Medicine Kehrer-Sawatzki, Hildegard; University of Ulm Cooper, David; Cardiff University, Institute of Medical Genetics, College of Medicine
Key Words:	Neanderthal genome , HGMD, potentially compensated mutations, complex disease susceptibility

Humu-2010-0442

Rapid Communication

Supporting Information for this preprint is available from the

Human Mutation editorial office upon request (humu@wiley.com)

Triangulation of the human, chimpanzee and Neanderthal genome sequences identifies potentially compensated mutations

Guojie Zhang[1*], Zhang Pei[1], Michael Krawczak[2], Edward V. Ball[3], Matthew Mort[3], Hildegard Kehrer-Sawatzki[4], David N. Cooper[3*]

[1]Bioinformatics Department, Beijing Genomics Institute at Shenzhen, Shenzhen 518083, China.

[2]Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität zu Kiel, Arnold-Heller-Straße 3, Haus 31, 24113 Kiel, Germany.

[3]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.

[4]Institut für Humangenetik, Universität Ulm, Albert-Einstein-Allee 11, 89081 Ulm, Germany.

*Corresponding authors:

Guojie Zhang. Email: zhanggj@genomics.org.cn

Tel: +86-0755-25273794 Fax: +86-0755-25273114

David N. Cooper. Email: cooperdn@cardiff.ac.uk

Tel:+44-2920-744062 Fax: +44-2920-746551

Abstract

Triangulation of the human, chimpanzee and Neanderthal genome sequences with respect to 44,348 disease-causing or disease-associated missense mutations and 1,712 putative regulatory mutations listed in the *Human Gene Mutation Database* was employed to identify genetic variants that are apparently pathogenic in humans but which may represent a 'compensated' wild-type state in at least one of the other two species. Of 122 such 'potentially compensated mutations' (PCMs) identified, 88 were deemed 'ancestral' on the basis that the reported wild-type Neanderthal nucleotide was identical to that of the chimpanzee. Another 33 PCMs were deemed to be 'derived' in that the Neanderthal wild-type nucleotide matched the human but not the chimpanzee wild-type. For the remaining PCM, all three wild-type states were found to differ. Whereas a derived PCM would require compensation only in chimpanzee, ancestral PCMs are useful as a means to identify sites of possible adaptive differences between modern humans on the one hand, and Neanderthals and chimpanzees on the other. Ancestral PCMs considered to be disease-causing in humans were identified in two Neanderthal genes (*DUOX2, MAMLD1*). Since the underlying mutations are known to give rise to recessive conditions in human, it is possible that they may also have been of pathological significance in Neanderthals.

KEY WORDS: Human; chimpanzee; Neanderthal; genome sequence; potentially compensated mutations; complex disease susceptibility

Introduction

The comparison of the human genome sequence with that of our closest living relative, the chimpanzee, has given us a broad overview of the spectrum of genetic changes that accompanied human evolution over the last 5-6 Myrs since the divergence of the two species [Kehrer-Sawatzki and Cooper, 2007; Marques-Bonet et al., 2009]. It is clear that both gross karyotypic rearrangements and submicroscopic variation (involving deletions, duplications and inversions) have made a significant contribution to the structural divergence of the two genomes, which is at least three-fold greater than the sequence divergence due to nucleotide substitution. This notwithstanding, detailed evaluation of the latter has served to identify many genes that exhibit signatures of selective sweeps (i.e. periods of intense positive selection) and which may therefore have been involved in the development of human lineage-specific traits [Bakewell *et al*., 2007; Enard *et al*., 2010]. Since signals of selective sweeps are likely to coincide with genetic variants that exert a significant effect on phenotypic variation in modern humans, they could, at least in principle, also impact upon susceptibility to common disease [Di Rienzo and Hudson, 2005; Nielsen *et al*., 2007]. In support of this assertion, there is a tendency for genes that have been positively selected during mammalian evolution to be disproportionately associated with human inherited disease [Clark *et al*., 2003; Vamathevan *et al*., 2008; Corona *et al*., 2010].

Positively selected mutations may include 'compensated' mutations, i.e. variants that were (actually or potentially) deleterious for a certain period of time, but which persisted long enough in a given population or species to have become positively selected upon the

introduction of a 'compensatory' mutation [Di Rienzo and Hudson, 2005; Corona *et al*., 2010]. Suriano et al. (2007) provided a good example of the interplay of compensated and compensating mutations in the context of a human disease. The human and chimpanzee OTC amino acid sequences differ at only two positions, 125 and 135, where Thr was found to represent both ancestral states. Replacements Thr135Ala and Thr125Met have occurred respectively in the human and chimpanzee lineages since their divergence from their common ancestor. However, whilst the wild-type combination of Met125 and Thr135 in chimpanzee gives rise to an apparently normal phenotype, recurrence of Met125 against the background of the human-specific Ala135 residue results in a clinical phenotype (human neonatal hyperammonemia). Suriano et al. (2007) demonstrated *in vitro* that human OTC bearing Met125 is inactive whereas the chimpanzee version with Met at the same position possesses an enzymatic activity comparable to wild-type human OTC. The presence of Thr135 in chimpanzee therefore rescues the otherwise deleterious effect of Met125 through intra-locus compensation.

The chimpanzee genome has been found to harbour several examples of such potentially compensated mutations (PCMs), formally defined as human disease-causing or disease-associated missense mutations for which the substituting amino acid is identical to the wild-type amino acid residue at the orthologous position in chimpanzee [Mikkelsen *et al*., 2005; Azevedo *et al*., 2006]. The absence of strongly deleterious consequences of a specific PCM in chimpanzee would be explicable either by virtue of the very different (simian) environment or by dint of hitherto unidentified variants ('compensatory mutations') in the chimpanzee genome

that could have served to epistatically buffer the PCM [Azevedo *et al*., 2006]. It should be noted, however, that a PCM may only have become disadvantageous in the human lineage as a consequence of the acquisition of other lineage-specific genetic variants or due to changes in the human environment and/or lifestyle [Di Rienzo and Hudson, 2005; Corona *et al*., 2010]. In this case, it would not have been necessary for the chimpanzee PCM to be compensated for at any time, which is why the qualifier 'potentially' is important when PCMs are defined on the basis of current genetic and clinical data.

Although the chimpanzee is our closest living relative, our closest known relatives were the now extinct Neanderthals (*Homo neanderthalensis*) from whom modern humans diverged between 300,000 and 700,000 years ago [Noonan *et al*., 2006]. The Neanderthals finally disappeared from the fossil record ~28,000 years ago [Mellars, 2004; Noonan, 2010], due to either climate change [Tzedakis *et al*., 2007] or competitive exclusion by anatomically modern humans [Banks *et al*., 2008]. Recently, a draft sequence (with ~1.3-fold coverage) of the Neanderthal genome, comprising >4 billion nucleotides derived from three apparently unrelated individuals, became available [Green *et al*., 2010]. Sequence differences between Neanderthals and modern humans were identified that pinpointed genomic regions potentially affected by positive selection [Green *et al*., 2010]. Furthermore, comparison of the Neanderthal and human genome sequences revealed that Neanderthal genetic material is present at a low level (1-4%) in those modern humans whose ancestors had probably migrated across Europe or Asia, but not in the genomes of individuals from two extant African populations studied [Green *et al*., 2010].

In this study, we have compared 3,202,190 nucleotide positions at which the Neanderthal

genome differs from either the human or the chimpanzee genome, or both, with a collection of 46,060 disease-causing or disease-associated mutations listed in the *Human Gene Mutation Database* [HGMD; Stenson *et al*., 2009]. From this comparison, we identified 122 HGMD mutations that apparently correspond to the wild-type allele in the Neanderthal or chimpanzee genomes (Figure 1). These mutations were further analysed in an attempt to identify PCMs in the Neanderthal genome that might in turn indicate sites of adaptive difference between modern humans and Neanderthals.

Methods

HGMD dataset

A total of 46,060 disease-causing or disease-associated mutations were obtained from the *Human Gene Mutation Database* [Stenson *et al*., 2009; http://www.hgmd.org] as of 13 May 2010. These data comprised 44,348 missense mutations from within the coding regions of 2,628 genes, and 1,712 single base-pair substitutions from within the regulatory regions (5' and 3' untranslated/flanking regions) of 807 genes. Some 42,595 of the mutations were disease-causing (DM; 41,960 missense and 635 regulatory) whereas 3,465 represented disease-associated or functional polymorphisms (2,388 missense and 1,077 regulatory) (Table 1). The latter were further ascribed to three distinct sub-categories: (i) disease-associated polymorphisms (DP), comprising variants reported to be in statistically significant association with a particular human disease state ($p<0.05$) but lacking experimental evidence of functionality e.g. from expression studies, (ii) disease-associated polymorphisms with experimental evidence of functionality

(DFP) such as, for example, altered *in vitro* gene expression or protein function (iii) functional polymorphisms (FP) that have been shown *in vitro* or *in vivo* to affect the structure, function or expression of the gene or gene product, but for which no statistically significant disease association has yet been reported (see http://www.hgmd.cf.ac.uk/docs/poly.html for further information).

Search for potentially compensated mutations (PCMs)

Sequence data on a total of 3,202,190 nucleotide positions at which the Neanderthal genome differs from either the human or chimpanzee genome (or both) were downloaded from the European Bioinformatics Institute website (ftp://ftp.ebi.ac.uk/). These positions were originally identified through mismatches between the human and chimpanzee genomes, followed by an alignment of Neanderthal sequence reads against both other species [Green *et al*., 2010]. At a total of 2,813,802 positions (87.9%), human and Neanderthal exhibited the same nucleotide, so that the human-chimpanzee mismatch must have arisen before the divergence of modern humans and Neanderthals (termed a 'derived' or 'D' state in the Neanderthal). A total of 38,228 positions (1.2%) displayed the same nucleotide in both Neanderthal and chimpanzee, suggesting that the respective substitutions were human-specific ('ancestral' or 'A' state in the Neanderthal). The remaining ~350,160 positions, which display different nucleotides in modern humans, Neanderthals and chimpanzees, were termed 'undefined' ('N' state). Since most of the N state positions overlapped with low quality Neanderthal sequencing reads, they were not considered further in this analysis. Among the remaining 2,852,030 positions, constituting the

'Neanderthal nucleotide substitution dataset', we identified 122 sites for which the apparent wild-type nucleotide in either Neanderthal or chimpanzee was logged in HGMD as disease-causing or disease-associated in humans (Figure 1).

Gene Ontology (GO) enrichment analysis

A Gene Ontology (GO) enrichment analysis of PCM-containing genes was performed using the Cytoscape software (http://www.cytoscape.org; version 2.5.2) with Bingo plugin (http://www.psb.ugent.be/cbd/papers/BiNGO/, version 2.3). The statistical significance of particular GO terms was assessed using a hypergeometric distribution, adjusted for multiple testing by consideration of the Benjamini-Hochberg False Discovery Rate [Benjamini and Hochberg, 1995].

Calculation of F_{ST}

The fixation index, F_{ST}, measures the proportion of genetic diversity in a sub-divided population that is due to allele frequency differences between sub-populations. Pair-wise F_{ST} values may also be used as a measure of genetic distance between populations. Here, allele frequencies of polymorphic ancestral PCMs in selected populations were obtained from HapMap (http://hapmap.ncbi.nlm.nih.gov/) and pair-wise F_{ST} values were estimated for each polymorphism using the small sample estimate proposed by Weir and Hill [2002]. The significance of individual F_{ST} values was then assessed by reference to the empirical distribution of F_{ST} among all SNPs in HapMap.

Protein structure modeling and ligand-protein binding site prediction

Protein structure modeling was carried out for three different versions of the MAMLD1 protein, jointly defined by a PCM (V505A) and a second, apparently non-pathological human-chimpanzee mismatch (I510M) in the MAMLD1 amino acid sequence: chimpanzee wild-type (p.A505 & p.M510), disease-causing in human, apparently wild-type in Neanderthal (p.A505 & p.I510), and human wild-type (p.V505 & p.I510). Since experimentally determined structures were not available, all three protein structures and the predicted ligand-protein binding sites had to be determined *in silico* using the I-TASSER server [Roy *et al*., 2010]. The C-score is a confidence score that quantifies the presumed quality of a protein model predicted by I-TASSER. C-scores are typically in the range of -5 to 2, where a higher C-score indicates a greater degree of confidence in a given model. The models derived in the present analysis had C-scores in the range of -1.46 to 0.99.

Results

Identification of PCMs in the Neanderthal and/or chimpanzee genome

A total of 44,348 missense mutations from 2,628 genes, logged in HGMD as being either causative of, or associated with, a human inherited disease state, were cross-compared to the corresponding nucleotide positions in the Neanderthal and chimpanzee genomes. HGMD-derived mutations at sites where the apparent wild-type nucleotide in Neanderthal and/or chimpanzee corresponded to the substituting nucleotide in humans (‘potentially

compensated mutations'; PCMs) were selected for further study. Although PCMs are most likely to be missense mutations, it nevertheless remains theoretically possible that PCMs also occur in other gene regions. Indeed, Kondrashov et al. [2002] argued that regulatory mutations may also be compensated for by the mediating effects of DNA- or RNA-binding proteins. For this reason, an additional 1,712 single base-pair substitutions from within the regulatory regions (5' and 3' untranslated/flanking regions) of 807 genes associated with human disease were retrieved from HGMD and were also included in this analysis.

Of the 122 PCMs identified (Table 2), 62 were missense mutations (Supp. Table S1) including five that were deemed to be disease-causing in humans (Table 3). Another 60 PCMs were disease-associated regulatory polymorphisms (Supp. Table S2). A total of 88 PCMs (72.1%) were ancestral, i.e. the Neanderthal and chimpanzee nucleotides were identical ('A' state), whereas 33 (27.0%) were derived, meaning that the Neanderthal nucleotide matched the human wild-type and that the PCM was confined to chimpanzee ('D' state) (Figure 1). Since being an A-state PCM implies that compensation, if any, must have occurred in both Neanderthal and chimpanzee, the 88 A-state PCMs (including two DMs) may serve to identify sites of adaptive difference between modern humans and Neanderthals. A D-state PCM, however, would only have required compensation in chimpanzee. For one of the 122 PCMs, namely a disease-associated functional polymorphism in the *F7* gene promoter (G/T at nucleotide -401; Supp. Table S2), the Neanderthal wild-type nucleotide matched the mutant state (-401T) as logged in HGMD, but differed from both the chimpanzee and human wild-type sequence [and hence constituted a third category ('N' state) in Table 2]. The -401T allele, which

is rare in humans, is known to be associated with a reduced basal rate of transcription of the human *F7* gene *in vitro* and with reduced plasma concentrations of total FVII antigen and activated FVII *in vivo* [van 't Hooft *et al*., 1999].

The number of A-state PCMs exceeded the number of D-state PCMs in every mutational category (Table 2) despite the fact that, among the missense and promoter mutations included in the Neanderthal nucleotide substitution dataset, the total numbers were 7,384 for state A and 47,391 for state D (data not shown). This discrepancy suggests that the vast majority of PCMs identified via HGMD (which it should be noted are actually polymorphisms rather than rare or private pathological mutations) have only acquired their phenotypic relevance fairly recently in the lineage leading to modern humans.

Disease-causing PCMs in Neanderthal and/or chimpanzee?

A total of five disease-causing human missense mutations were identified as PCMs in that they corresponded precisely to the alleles that were wild-type in Neanderthal and/or chimpanzee (Table 3). Two of these [*DUOX2* (MIM# 606759), *MAMLD1* (MIM# 300120)] were A-state in that the substituting residues in human matched the wild-type residues in both Neanderthal and chimpanzee. The *DUOX2* mutation (H678R) is associated with autosomal recessive hypothyroidism whereas the *MAMLD1* mutation (p.V505A) is associated with X-linked recessive hypospadias.

We next sought mutations in the *DUOX2* and *MAMLD1* genes that might have served to compensate for the deleterious effect of the respective PCM in chimpanzee and/or Neanderthal.

Compensatory mutations appear to occur disproportionately in the vicinity of the sites of the original deleterious (i.e. compensated) mutations [Davis *et al*., 2009]. However, only in *MAMLD1* did pair-wise alignment reveal a potentially compensating chimpanzee amino acid residue (p.M510) in the vicinity of the human disease-causing mutation (p.V505A). Residue 510 is a methionine not only in chimpanzee but also in all other primates investigated, with the sole exceptions of human and Neanderthal where isoleucine is present at this location (Figure 2). Interestingly, protein structure modeling predicted that the combination of p.V505 and p.I510 (i.e. the human wild-type) would introduce a novel nucleic acid-protein binding site (Figure 2) that would not have been present in either chimpanzee or Neanderthal (and which would have been abolished by the hypospadias-causing p.V505A mutation). These findings suggest that the nucleotide substitutions giving rise to a valine at p.505 and an isoleucine at p.510 may have exerted a cooperative effect which allowed both of them to become fixed during human evolution.

One of the three disease-causing D-state PCMs (Table 3), the *SLC5A1* (MIM# 182380) H615Q substitution, is thought to be responsible for recessive glucose/galactose malabsorption in human, and could represent an example of a clinically significant mutation that is present in both modern humans and chimpanzees. The second D-state PCM, a mutation in *IL12RB1* (MIM# 601604), is also associated with a recessive condition in humans (susceptibility to mycobacterial infection). Intriguingly, the third D-state PCM, a mutation of the *EXT1* (MIM# 608177) gene, occurs in the only amino acid residue that differs between the wild-type chimpanzee and human EXT1 proteins. Thus, this mutation cannot be a classic compensated

mutation (i.e. accompanied by a compensatory mutation in the immediate vicinity), but it could still conceivably be compensated for by non-allelic changes (between chimpanzee and human), for example in a protein binding partner. Although, in all three cases, the disease-causing mutation in human matches the wild-type residue in chimpanzees, the possible relevance to chimpanzee pathophysiology remains unclear.

Pathophysiology of PCMs in modern humans and Neanderthals

Upon Gene Ontology (GO) enrichment analysis, genes that contained PCMs (both missense and regulatory) in Neanderthal were found to be significantly enriched in functions related to the immune system and stimulus response (Table 4). This finding is consistent with previous reports of immune response genes having been subject to positive selection during primate evolution [Nielsen *et al*., 2005; Vamathevan *et al*., 2008; Enard *et al*., 2010]. It is clearly not possible, however, on the basis of this GO analysis, to extrapolate from human to Neanderthals.

Human variants with different population frequencies at sites of PCMs

Of the 86 ancestral polymorphic PCMs identified in our study (i.e. categories DP, FP and DFP in Table 2), 19 were characterized by at least one nominally significant allele frequency difference between extant African and non-African human populations (International HapMap Consortium, 2007; Table 5). This represents a more than two-fold excess over random expectation ($2\times0.05\times86=8.6$). We used Wright's fixation index F_{ST} to quantify the degree of genetic divergence between the different populations at these loci. Alleles that have been the

target of localized positive selection tend to exhibit unusually high F_{ST} values [Thornton and Jensen, 2007; Holsinger and Weir, 2009]. Therefore, we compared the F_{ST} values of the ancestral polymorphic PCMs to the empirical F_{ST} distribution derived from all HapMap SNPs [International HapMap Consortium, 2007] to assess the significance of individual F_{ST} values. Previous analysis has shown that the average pair-wise F_{ST} for Africans and Asians is around 0.19, compared to 0.15 for Africans and Europeans [International HapMap Consortium, 2007]. The 19 PCMs mentioned above exhibited at least one pair-wise F_{ST} value between Africans and Asians or Europeans of 0.4 or higher, consistent with the differential action of recent positive selection in these populations.

The highest F_{ST} values were observed for a disease-associated/functional polymorphism (K121Q; AAG>CAG) in the *ENPP1* (MIM# 173335) gene (Table 5; Supp. Table S2), that is associated with insulin resistance and obesity in modern humans. This variant, which exhibited particularly high pair-wise F_{ST} values in our comparison of African and non-African populations (Table 5), occurs within a chromosomal region that has been subject to a selective sweep in early humans [Green *et al.*, 2010]. The A allele of the polymorphism is human-specific and must have originated in the modern human lineage. Consistent with a strong environmental influence upon its maintenance, the increased risk of type-2 diabetes conferred by the alternative C (Q121) allele may be abolished by lifestyle intervention [Moore *et al.*, 2009]. The reported success of a clinical intervention in this context provides ample retrospective justification for the identification and analysis of genetic susceptibility variants for human disease, even if only statistically associated.

Discussion

Previous studies have shown that humans and Neanderthals shared some key evolutionary sequence changes (e.g. in the *FOXP2* [MIM# 605317; Krause *et al*., 2007], *CMAH* [MIM# 603209; Chou *et al*., 2002] and *AGAP1* [MIM# 608651; Hünemeier *et al*., 2010] genes) which are absent from the chimpanzee genome. In addition, the two hominins have been shown to have had polymorphisms in the *ABO* blood group [MIM# 110300; Lalueza-Fox *et al*., 2008], *MCPH1* [MIM# 607117; Lari *et al*., 2010] and *TAS2R38* taste receptor [MIM# 607751; Lalueza-Fox *et al*., 2009] genes in common, testifying to their ancient origin. On the other hand, a Neanderthal-specific variant in the melanocortin 1 receptor (*MC1R*; MIM# 155555) gene has also been reported [Lalueza-Fox *et al*., 2007]. Having sequenced the Neanderthal genome, Green *et al*. [2010] were able to identify a further 78 amino acid differences between the two hominins, with the derived allele having become fixed in modern humans whilst Neanderthals appear to have carried the ancestral (chimpanzee-like) allele.

The sequencing of the Neanderthal genome represents an outstanding scientific achievement and provides us with an invaluable resource for comparative genomic studies involving humans. However, when the resultant sequence data are employed in studies such as ours, there are various *caveats* that need to be considered. Firstly, the possibility of cross-contamination with modern human DNA is omnipresent [Wall and Kim, 2007]. Green *et al*. [2010] estimated that the extent of DNA contamination by modern human males is about 0.6% in the combined Neanderthal sequence data (with an upper 95% bound of 1.5%). Secondly, artefactual C>T and

G>A substitutions, resulting from the time-dependent deamination of 5-methylcytosine *post mortem*, may be quite frequent although it is unclear precisely how frequent [Briggs *et al*., 2007; Briggs *et al*., 2010]. By contrast, other types of artefactual nucleotide substitution were found to occur at significantly lower frequencies [Green *et al*., 2010]. Thus, for C>T and G>A substitutions at CpG dinucleotides, it is very difficult to distinguish between *bona fide* germline mutations that occurred recurrently in both lineages and artefactual mutations that steadily accumulated in the DNA samples *post mortem*. One way in which *post mortem* CG>TG, CA mutations could have inadvertently altered our conclusions would have been if *bona fide* D-state sites had been mis-scored as A-state sites (i.e. false positives), thereby artificially inflating the number of A-state PCMs identified. To assess the potential for such a source of error, we compared the frequency of CpG-located C>T and G>A substitutions among A-state PCMs (10/88; 0.11) with that in the HGMD missense and regulatory mutation dataset (8812/46060; 0.19). Owing to the nearly two-fold excess seen in HGMD, we may conclude that the Neanderthal genome sequence data reported by Green *et al*. [2010] are unlikely to have been compromised by any excess CpG mutations originating *post mortem*. The above notwithstanding, it should be appreciated that the attribution of an A-state has the potential to be somewhat misleading for mutations occurring within CpG sites because identical C>T or G>A transitions may have occurred quite independently in the Neanderthal and chimpanzee lineages as a consequence of high frequency methylation-mediated deamination of 5-methylcytosine *in vivo*.

We know comparatively little about health and disease in Neanderthals, who appear to have

suffered from various systemic infections [Ember and Ember, 2004] and for whom both a natural kyphosis of the lumbar spine [Weber and Pusch, 2008] and osteochondromas/ exostoses [Trinkhaus, 2008] have been inferred from skeletal remains..Neanderthals and modern humans differed quite markedly in terms of craniofacial features; although Neanderthal brain size was similar at birth to that of modern humans, brain growth rates during early infancy may have been higher [Ponce de León and Zollikofer, 2001; Ponce de León *et al*., 2008]. With the above in mind, the A761V amino acid sequence polymorphism (associated with cranial volume) identified as a PCM in the *MCPH1* (MIM# 607117) gene is intriguing, as is the V537I polymorphism (associated with osteochondromas/exostoses) identified as a PCM in the *EXT1* gene (Supp. Table S1). It may also be pertinent to note that, of the 60 regulatory PCMs in Neanderthal or chimpanzee (listed in Supp. Table S2), four are associated with body fat/obesity, five are associated with lipid metabolism, five are associated with insulin resistance/diabetes, three are associated with asthma, and two are associated with resistance to infection.

Green *et al*. [2010] identified a number of genetic regions that were involved in gene flow from Neanderthals to non-African modern humans. One of these regions harbours the abovementioned *SLC5A1* H>Q mutation, responsible for the disease phenotype 'glucose/galactose malabsorption'. However, since there is no difference in allele frequency between Africans and non-Africans (a finding which could also be explicable in terms of balancing selection), this lesion would appear not to have originated via gene-flow from Neanderthals.

The idea that two individually deleterious mutations might be capable of restoring normal

fitness when they occur in combination can be traced back to Kimura [1985] who suggested that 'compensatory neutral mutations' might play an important role in evolution. More recently, Kondrashov *et al*. [2002] compared pathological missense mutations in 32 human proteins to the amino acid substitutions that occurred during the course of evolution of these same proteins, and estimated that ~10% of all amino acid sequence differences between a human protein and a non-human (mammalian) orthologue could represent what they termed 'compensated pathogenic deviations' (CPDs), essentially equivalent to the PCMs discussed here. Since CPDs would be pathogenic in humans, Kondrashov *et al*. [2002] surmised that the normal functioning of a CPD-containing protein in the non-human species must be due to other (compensatory) amino acid sequence deviations from the human sequence. In other words, as many as 10% of all the amino acid substitutions that become fixed in an evolving protein may be dependent upon compensatory substitutions to be benign. It is assumed that the compensatory mutations serve to neutralize the potential detrimental effect of the compensated mutation on the structure and stability of the protein [Ferrer-Costa *et al*., 2007].

With the sequencing of the genomes of various model organisms, numerous examples have emerged of compensated mutations i.e. human pathological missense mutations where the substituting amino acid has been found to be identical to the wild-type amino acid residue at the orthologous position in mouse [Gao and Zhang 2003], macaque [Gibbs *et al*., 2007] or chimpanzee [Mikkelsen *et al*., 2005; Azevedo *et al*., 2006]. This apparent paradox is potentially explicable in terms of the 'compensatory mutation hypothesis' which holds that the absence of strongly deleterious effects in other mammalian species is due to the buffering effect of epistatic

interactions between the mutation which causes disease in human (the 'compensated mutation') and other mutational changes that compensate in some way for the functional alteration in the mammalian protein (the 'compensatory mutations'). In principle, these compensatory changes could be either allelic to the compensated mutation or non-allelic. It is assumed that the compensatory mutations, present in the non-human protein, serve to neutralize the potential detrimental effect of the compensated mutation on the structure of the protein [Ferrer-Costa *et al*., 2007]. In evolution, compensatory mutations are unlikely to occur singly; indeed, Poon *et al*. [2005] have suggested that an average of 11.8 compensatory mutations may interact epistatically with a given deleterious mutation so as to restore wild-type levels of fitness. CPDs tend to be less severe in terms of the physicochemical difference between the substituted and substituting amino acids than normally pathological mutations [Ferrer-Costa *et al*., 2007; Barešić *et al*., 2010]. Barešić *et al*. [2010] have also shown that amino acid residues surrounding the compensated residue in the folded protein are mutated more often than residues surrounding an uncompensated mutation, consistent with the view that compensation relies upon structurally local mutations.

Consistent with the idea that common disease susceptibility may be the result of long-term human adaptation to a steady ancient environment, a number of the alleles that increase the risk of common diseases are indeed ancient [Di Rienzo & Hudson 2005; Corona *et al*., 2010]. In this framework, the newly acquired allele confers protection against the disease. The human risk alleles could correspond to either persistent (i.e. ancient) or recurring mutations that represent the recapitulation of ancestral states that may once have been protective, but which

now result in adverse consequences for human health. If compensated mutations are more likely to be fixed by positive selection than by drift [Corona *et al*., 2010], then these variants could indeed represent adaptive differences.

Here we have performed triangulation of human, chimpanzee and Neanderthal genome sequence data and cross-compared the chromosomal coordinates of the recognized sequence differences with known human disease-related mutations. A total of 122 mutations were identified (62 of which were missense mutations including five deemed to be disease-causing in humans) that were potentially compensated in chimpanzee or Neanderthal. Of these 122 human mutations, termed PCMs, 88 were 'ancestral' in that the PCM matched the wild-type in both non-human species whereas 33 were 'derived' in that the Neanderthal wild-type matched the human wild-type (i.e. the PCM was confined to chimpanzee). The 88 ancestral PCMs could potentially be indicative of genomic regions that harbour adaptive differences between modern humans and Neanderthals. 'Ancestral' PCMs that are disease-causing in humans were identified in two human genes (*DUOX2, MAMLD1*). These pathological lesions could thus provide examples of compensated mutations in Neanderthal but, since all these mutations give rise to recessively inherited conditions in humans, it is also possible that they could have been associated with disease causation/susceptibility in Neanderthals. Intriguingly, we identified 19 PCMs in both Neanderthals and chimpanzees that are polymorphic in human and that exhibit nominally significant frequency differences between the African and at least one non-African population. These PCMs represent major candidates for recent population-specific selection, consistent with different alleles having exhibited differential functional importance in different

environments.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248-249.

Azevedo L, Suriano G, van Asch B, Harding RM, Amorim A. 2006. Epistatic interactions: how strong in disease and evolution? Trends Genet 22:581-585.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci USA 104:7489-7494.

Banks WE, d'Errico F, Peterson AT, Kageyama M, Sima A, Sánchez-Goñi MF. 2008. Neanderthal extinction by competitive exclusion. PLoS One 3:e3972.

Barešić A, Hopcroft LE, Rogers HH, Hurst JM, Martin AC. 2009. Compensated pathogenic deviations: analysis of structural effects. J Mol Biol 396:19-30.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Series B (Methodological) 57:289-300.

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci USA 104:14616-14621.

Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. (2010) Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. Nucleic Acids Res 38:e87.

Chou HH, Hayakawa T, Diaz S, Krings M, Indriati E, Leakey M, Paabo S, Satta Y, Takahata N, Varki A. 2002. Inactivation of CMP-*N*-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. Proc Natl Acad Sci USA 99:11736-11741.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960-1963.

Corona E, Dudley JT, Butte AJ. 2010. Extreme evolutionary disparities seen in positive selection across seven complex diseases. PLoS ONE 5:e12236.

Davis BH, Poon AF, Whitlock MC. 2009. Compensatory mutations are repeatable and clustered within proteins. Proc Biol Sci 276:1823-1827

Di Rienzo A, Hudson RR. 2005. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21:596-601.

Ember CR, Ember M (Eds). 2004. Encyclopedia of Medical Anthropology: Health and Illness in the World's Cultures. Springer, New York.

Enard D, Depaulis F, Roest Crollius H. 2010. Human and non-human primate genomes share hotspots of positive selection. PLoS Genet 6:e1000840.

Ferrer-Costa C, Orozco M, de la Cruz X. 2007. Characterization of compensated mutations in terms of structural and physico-chemical properties. J Mol Biol 365:249-256.

Gao L, Zhang J. 2003. Why are some human disease-associated mutations fixed in mice? Trends Genet 19:678-681.

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW,

Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csürös M, Glasscock J, Harris RA, Havlak P, Jackson AR, Jiang H, Liu Y, Messina DN, Shen Y, Song HX, Wylie T, Zhang L, Birney E, Han K, Konkel MK, Lee J, Smit AF, Ullmer B, Wang H, Xing J, Burhans R, Cheng Z, Karro JE, Ma J, Raney B, She X, Cox MJ, Demuth JP, Dumas LJ, Han SG, Hopkins J, Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, Addo-Quaye C, Degenhardt J, Denby A, Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson HA, Marklein A, Nielsen R, Vallender EJ, Clark AG, Ferguson B, Hernandez RD, Hirani K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu LL, Ren Y, Smith DG, Wheeler DA, Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F, Kent WJ, Lesk A, Nelson DL, O'brien WE, Prüfer K, Stenson PD, Wallace JC, Ke H, Liu XM, Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith KE, Zwieg AS. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316: 222-234.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Pääbo S. 2010. A draft sequence of the Neanderthal genome. Science 328:710-722.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat Rev Genet 10:639-650.

Hünemeier T, Ruiz-Linares A, Silveira A, Paixão-Côrtes VR, Salzano FM, Bortolini MC. 2010. Population data support the adaptive nature of HACNS1 sapiens/neandertal-chimpanzee differences in a limb expression domain. Am J Phys Anthropol In press.

International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel

B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren

BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Kehrer-Sawatzki H, Cooper DN. 2007. Understanding the recent evolution of the human genome: insights from human-chimpanzee genome comparisons. Hum Mutat 28:99-130.

Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution. J Genet 64: 7-19.

Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci USA 99:14878-14883.

Krause J, Lalueza-Fox C, Orlando L, Enard W, Green RE, Burbano HA, Hublin JJ, Hänni C, Fortea J, de la Rasilla M, Bertranpetit J, Rosas A, Pääbo S. 2007. The derived *FOXP2* variant of modern humans was shared with Neanderthals. Curr Biol 17:1908-1912.

Lalueza-Fox C, Römpler H, Caramelli D, Stäubert C, Catalano G, Hughes D, Rohland N, Pilli E, Longo L, Condemi S, de la Rasilla M, Fortea J, Rosas A, Stoneking M, Schöneberg T, Bertranpetit J, Hofreiter M. 2007. A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. Science 318:1453-1455.

Lalueza-Fox C, Gigli E, de la Rasilla M, Fortea J, Rosas A, Bertranpetit J, Krause J. 2008. Genetic characterization of the ABO blood group in Neanderthals. BMC Evol Biol 8:342.

Lalueza-Fox C, Gigli E, de la Rasilla M, Fortea J, Rosas A. 2009. Bitter taste perception in Neanderthals through the analysis of the *TAS2R38* gene. Biol Lett 5:809-811.

Lari M, Rizzi E, Milani L, Corti G, Balsamo C, Vai S, Catalano G, Pilli E, Longo L, Condemi S, Giunti P, Hänni C, De Bellis G, Orlando L, Barbujani G, Caramelli D. 2010. The microcephalin ancestral allele in a Neanderthal individual. PLoS One 5:e10648.

Marques-Bonet T, Ryder OA, Eichler EE. 2009. Sequencing primate genomes: what have we learned*?* Annu Rev Genomics Hum Genet 10:355-386.

Mellars P. 2004. Neanderthals and the modern human colonization of Europe. Nature 432:461-465.

Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, JaVe DB, Yang S-P, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK, Archidiacono N, Bork P, Butler J, Chang JL, Cheng Z, Chinwalla AT, deJong P, Delehaunty KD, Fronick CC, Fulton LL, Gilad Y, Glusman G, Gnerre S, Graves TA, Hayakawa T, Hayden KE, Huang X, Ji H, Kent WJ, King M-C, Kulbokas EJ, Lee MK, Liu G, Lopez-Otin C, Makova KD, Man O, Mardis ER, Mauceli E, Miner TL, Nash WE, Nelson JO, Pääbo S, Patterson NJ, Poh CS, Pollard KS, Prüfer K, Puente XS, Reich D, Rocchi M, Rosenbloom K, Ruvolo M, Richter DJ, SchaVner SF, Smit AFA, Smith SM, Suyama M, Taylor J, Torrents D, Tuzun E, Varki A, Velasco G, Ventura M, Wallis JW, Wend MC, Wilson RK, Lander ES, Waterston RJ. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

Moore AF, Jablonski KA, Mason CC, McAteer JB, Arakaki RF, Goldstein BJ, Kahn SE, Kitabchi AE, Hanson RL, Knowler WC, Florez JC; Diabetes Prevention Program Research Group. 2009. The association of *ENPP1* K121Q with diabetes incidence is abolished by lifestyle modification in the diabetes prevention program. J Clin Endocrinol Metab 94:449-455.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, J Sninsky J, Adams MD, Cargill M. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3:e170.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing

selection in the human genome. Nat Rev Genet 8:857–868.

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, Rubin EM. 2006. Sequencing and analysis of Neanderthal genomic DNA. Science 314:1113-1118.

Noonan JP. 2010. Neanderthal genomics and the evolution of modern humans. Genome Res 20:547-553.

Ponce de León MS, Zollikofer CP. 2001. Neanderthal cranial ontogeny and its implications for late hominid diversity. Nature 412:534-538

Ponce de León MS, Golovanova L, Doronichev V, Romanova G, Akazawa T, Kondo O, Ishida H, Zollikofer CP. 2008. Neanderthal brain size at birth provides insights into the evolution of human life history. Proc Natl Acad Sci USA 105:13764-13768.

Poon A, Davis BH, Chao L. 2005. The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. Genetics 170:1323-1332.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein

structure and function prediction. Nature Protocols 5:725-738.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The Human Gene Mutation Database: 2008 update. Genome Med 1:13.

Suriano G, Azevedo L, Novais M, Boscolo B, Seruca R, Amorim A, Ghibaudi EM. 2007. *In vitro* demonstration of intra-locus compensation using the ornithine transcarbamylase protein as model. Hum Mol Genet 16:2209-2214.

Thornton KR, Jensen JD. 2007. Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175:737-750.

Trinkhaus E, Maley B, Buzhilova AP. 2008. Paleopathology of the Kiik-Koba 1 Neandertal. Am J Phys Anthropol 137:106-112.

Tzedakis PC, Hughen KA, Cacho I, Harvati K. 2007. Placing late Neanderthals in a climatic context. Nature 449:206-208.

Vamathevan JJ, Hasan S, Emes RD, Amrine-Madsen H, Rajagopalan D, Topp SD, Kumar V, Word M, Simmons MD, Foord SM, Sanseau P, Yang Z, Holbrook JD. 2008. The role of positive selection in determining the molecular cause of species differences in disease. BMC Evol Biol

8:273.

van 't Hooft FM, Silveira A, Tornvall P, Iliadou A, Ehrenborg E, Eriksson P, Hamsten A. 1999. Two common functional polymorphisms in the promoter region of the coagulation factor VII gene determining plasma factor VII activity and mass concentration. Blood 93:3432-3441.

Wall JD, Kim SK. 2007. Inconsistencies in Neanderthal genomic DNA sequences. PLoS Genet 3:e175.

Weber J, Pusch CM. 2008. The lumbar spine in Neanderthals shows natural kyphosis. Eur Spine J 17 Suppl 2:S327-330.

Weir BS, Hill WG. 2002. Estimating F-statistics. Annu Rev Genet 36:721-750.

Figure legends

Figure 1: Potential compensated mutations (PCMs) identified by genomic triangulation. (A) 88 PCMs were ancestral, i.e. the Neanderthal and chimpanzee nucleotides were identical (ancestral state). (B) 33 PCMs were derived, denoting that the Neanderthal nucleotide matched the human wild-type nucleotide and that the PCM was confined to chimpanzee (derived state). An ancestral PCM should be compensated for in both Neanderthal and chimpanzee. By

contrast, a derived PCM would only have required compensation in chimpanzee.

Figure 2: A potential example of a compensating amino acid residue (p.M510) in the chimpanzee MAMLD1 protein in the vicinity of the PCM at amino acid residue p.A505. (A) Multiple alignment of the orthologous regions of the MAMLD1 protein for a variety of primate species. In human wild-type MAMLD1, the combination of p.V505 and p.I510 is predicted by protein modelling to introduce a novel human-specific ligand-protein binding site (which would be abolished by the hypospadias-causing p.V505A mutation). (B) Predicted protein-ligand complex. MAMLD1 residues that interact with the ligand are coloured yellow. The p.V505 and p.I510 residues are shown in red with green highlighting.

Table 1. Missense and regulatory mutations from HGMD used in this study, categorised by mutation type and putative role in disease aetiology

Mutation/ polymorphism type	Type and putative role in disease aetiology				
	DM	DP	DFP	FP	Total
Missense	41,960	942	295	1,151	44,348
Regulatory	635	340	391	346	1,712
Total	42,595	1,282	686	1,497	46,060

DM: disease-causing mutation; DP: disease-associated polymorphism lacking functional evidence; DFP: disease-associated polymorphism with functional evidence; FP: polymorphism with functional evidence, but lacking a reported disease association as yet.

Table 2. HGMD-derived mutations identified as potentially compensated mutations (PCMs) in the Neanderthal and/or chimpanzee genome

Mutation/ polymorphism type	Mutation type and basis of disease aetiology						Neanderthal substitutions
	PCM state	DM	DP	FP	DFP	Total	
Missense	A	2	25	6	14	47	3,703
	D	3	9	2	1	15	21,655
	Total	5	34	8	15	62	25,358
Regulatory	A	0	20	10	11	41	3,681
	D	0	9	6	3	18	25,736
	Total	0	29	16	15	60	29,471

A: ancestral (i.e. Neanderthal nucleotide identical to both the chimpanzee wild-type nucleotide and the human disease-causing/disease-associated mutation, but not the human wild-type nucleotide); D: derived (i.e. Neanderthal nucleotide identical to the human wild-type but not the chimpanzee wild-type nucleotide, which is identical to the human disease-causing/disease-associated mutation); N: Neanderthal wild-type identical to the human disease-causing/disease-associated mutation, but not to the chimpanzee or the human wild-type nucleotide. For further details, see legend to Table 1.

Table 3. Human disease-causing mutations (DMs) identified as PCMs in chimpanzee and/or Neanderthal

Gene	Mutation	HGVS nomenclature (cDNA)HGVS nomencla (protein)	PCM state	Disease phenotype
SLC5A1	C=>G:CG	NM_000343.1:c.1845C>G	NP_000334.1:p.H615Q	D	Glucose/galactose malabsorption
EXT1	G=>A:GA	NM_000127.2:c.1609G>A	NP_000118.2:p.V537I	D	Multiple osteochondromas
IL12RB1	A=>G:AG	NM_005535.1:c.641A>G	NP_005526.1:p.Q214R	D	Mycobacterial infection, susceptibility to
DUOX2	A=>G:GG	NM_014080.4:c.2033A>G	NP_054799.4:p.H678R	A	Hypothyroidism
MAMLD1	T=>C:CC	NM_005491.2:c.1514T>C	NP_005482.2:p.V505A	A	Hypospadias

Mutations are given in the format human wild-type => human mutation: Neanderthal nucleotide, chimpanzee nucleotide. HGVS (cDNA) and HGVS (protein) represent mutation descriptions according to HGVS guidelines (www.hgvs.org/mutnomen) at the cDNA and protein levels, respectively. HGVS (cDNA) nomenclature provides cDNA numbering with +1 corresponding to the A of the ATG translational initiation codon in the corresponding reference sequence. HGVS (protein) nomenclature provides numbering relative to the reference protein sequence with the translational initiation codon = 1. NM & NP denote NCBI RefSeq & RefProt accession numbers respectively [http://www.ncbi.nlm.nih.gov/RefSeq/key.html]. For further details, see legends to Tables 1 and 2.

Table 4. GO term enrichment for genes with potential compensatory mutations (in relation to Neanderthal) occurring specifically in modern humans

GO ID	GO term	GO class	P
GO:0002520	immune system development	BP	0.0034
GO:0048583	regulation of response to stimulus	BP	0.0040
GO:0002376	immune system process	BP	0.0040
GO:0002682	regulation of immune system process	BP	0.0220
GO:0005624	membrane fraction	CC	0.0220
GO:0050896	response to stimulus	BP	0.0220
GO:0000267	cell fraction	CC	0.0314
GO:0030225	macrophage differentiation	BP	0.0342
GO:0048534	hemopoietic or lymphoid organ development	BP	0.0342
GO:0002521	leukocyte differentiation	BP	0.0367
GO:0006955	immune response	BP	0.0367
GO:0010743	regulation of foam cell differentiation	BP	0.0416
GO:0030097	hemopoiesis	BP	0.0469

BP: Biological process. CC: Cellular component. P: p-value from a hypergeometric distribution, adjusted for test multiplicity by consideration of the false discovery rate.

Table 5. Ancestral polymorphic PCMs with significantly different genotype frequencies in different HapMap populations (http://hapmap.ncbi.nlm.nih.gov/)

Gene	rs number	HGMD Accession Number	Nucleotide		Asian		European		African		Pair-wise F_{ST} (p value)	
			WT	PCM	f_{WT}	n	f_{WT}	n	f_{WT}	n	Asian-African	European-African
ENPP1	rs1044498	CM993455	A	C	0.92	340	0.88	402	0.15	692	0.729 (0.0046)	0.690 (0.0033)
WRAP53	rs2287499	CM077855	C	G	0.72	176	0.84	120	0.08	120	0.575 (0.0192)	0.733 (0.0020)
TP53BP1	rs560191	CM067475	C	G	0.52	178	0.69	120	0.00	120	0.472 (0.0414)	0.687 (0.0034)
*GHRHR**	rs2302019	CR066667	C	T	0.69	340	0.55	402	0.06	692	0.646 (0.0105)	0.486 (0.0239)
*LPL**	rs1800590	CR971950	T	G	1.00	176	0.98	120	0.48	114	0.576 (0.0190)	0.483 (0.0245)
TP53BP1	rs2602141	CM067476	A	C	0.59	340	0.67	400	0.09	692	0.480 (0.0391)	0.558 (0.0129)
*THPO**	rs6141	CR014438	G	A	0.53	340	0.42	402	0.01	586	0.580 (0.0184)	0.443 (0.0337)
LTF	rs1126478	CM096382	A	G	0.37	340	0.68	402	0.06	692	0.279 (0.1397)	0.619 (0.0072)
*UGT1A1**	rs4124874	CR025220	T	G	0.68	340	0.56	402	0.12	690	0.517 (0.0300)	0.371 (0.0566)
ABCB1	rs2032582	CM033585	T	G	0.57	340	0.45	402	0.05	464	0.507 (0.0323)	0.362 (0.0603)
*SLC6A4**	rs1042173	CR084012	T	G	0.17	340	0.53	402	0.84	692	0.622 (0.0130)	0.217 (0.1599)
VNN1	rs4897612	CR075274	T	G	0.63	340	0.68	400	0.17	692	0.382 (0.0753)	0.430 (0.0370)
*ARG1**	rs2781666	CR073540	G	T	0.68	340	0.74	290	0.22	292	0.349 (0.0922)	0.427 (0.0380)
SFTPD	rs2243639	CM067461	A	G	0.26	340	0.43	402	0.02	584	0.261 (0.1557)	0.431 (0.0368)
*ENPP1**	rs7754561	CR052970	A	G	0.41	340	0.72	402	0.15	692	0.173 (0.2568)	0.507 (0.0201)
TLR1	rs4833095	CM094340	A	G	0.32	336	0.70	402	0.12	690	0.110 (0.3634)	0.532 (0.0162)
C17orf53	rs227584	CM093418	A	C	0.26	340	0.70	400	0.12	690	0.068 (0.4612)	0.535 (0.0158)
BCR	rs140504	CM057927	A	G	0.49	340	0.12	402	0.03	692	0.494 (0.0355)	0.055 (0.4664)
AGT	rs699	CM920010	T	C	0.19	340	0.58	402	0.11	692	0.020 (0.6231)	0.415 (0.0416)

WT: wild-type nucleotide in (non-African) human populations; PCM: potentially compensated mutation in the Neanderthal and chimpanzee genome; f_{WT}: frequency of the human wild-type allele in the respective population; n: sample size (number of chromosomes) in HapMap; F_{ST}: small sample estimate of Wright's fixation index; p value: as obtained by reference to the overall distribution of SNP-based pair-wise F_{ST} values in HapMap. Disease-associated/functional polymorphisms in gene regulatory regions are denoted by *.

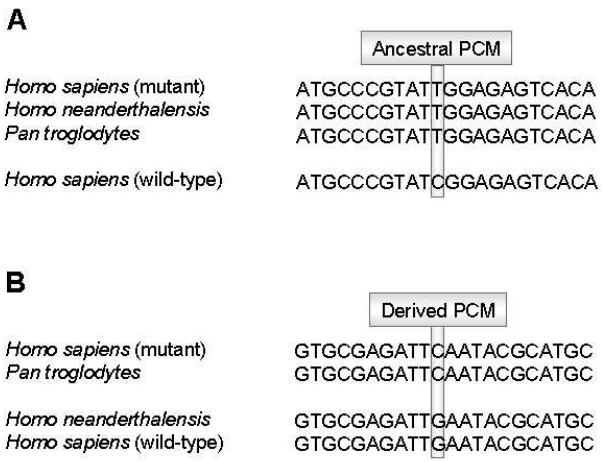
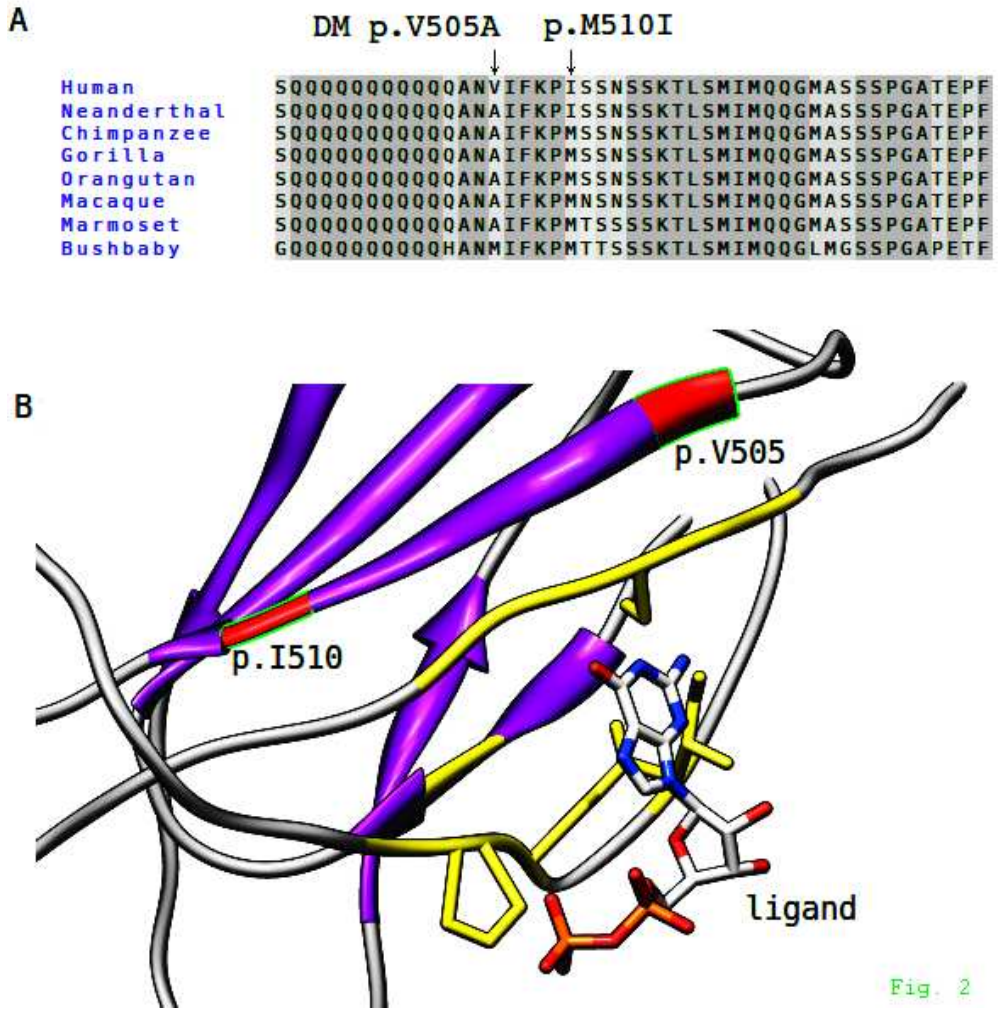
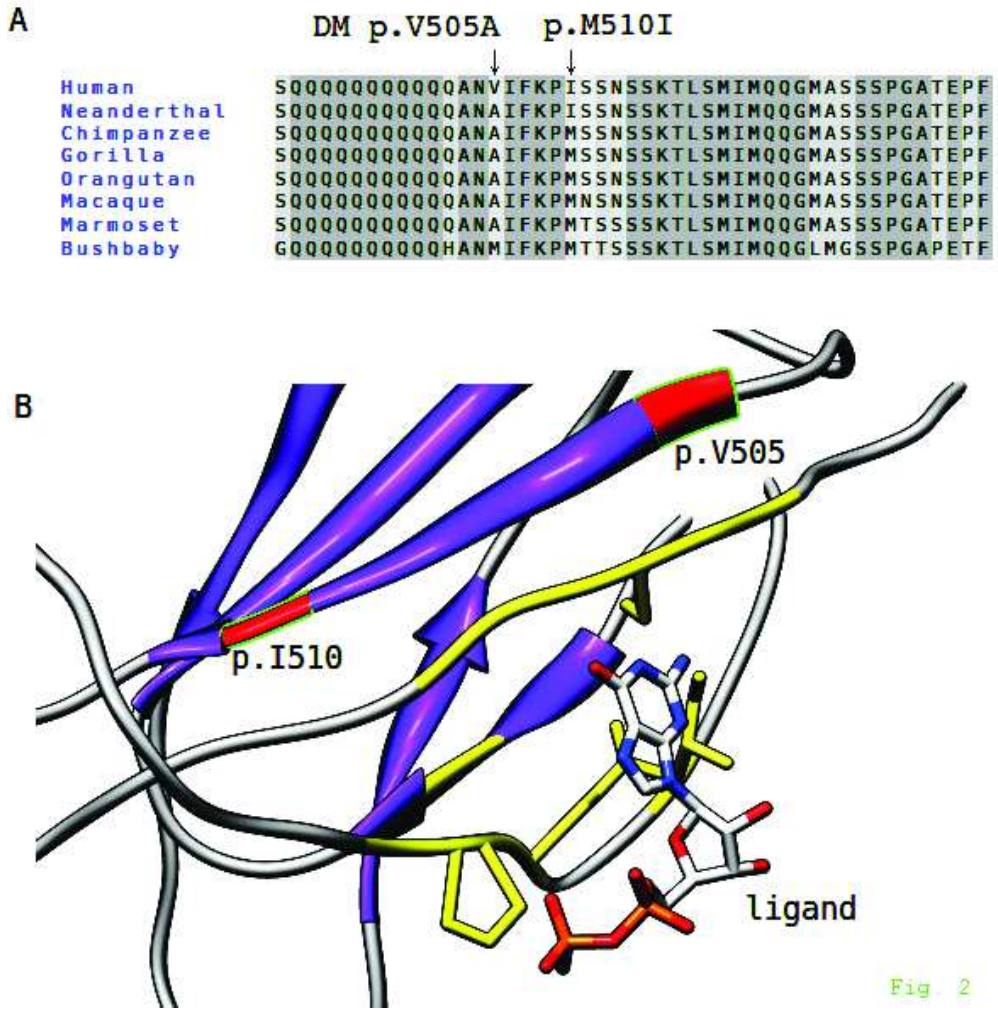


Figure 1

81x60mm (300 x 300 DPI)



49x50mm (300 x 300 DPI)



49x50mm (300 x 300 DPI)

Supp. Table S1. Potential compensated missense mutations in the Neanderthal and/or chimpanzee genome

HGMD Acc. No.	Chromosomal location	Mutation	State	Tag	Human_disease_state	pmid	Gene	CpG	dbsnp id.	HGVS (cDNA) nomenclature	HGVS (protein) nomenclature
CM067656	chr1:156491643:+	C=>G:CG	D	DP	Guillain-Barre syndrome, reduced risk, association with?	16820217	*CD1A*	N	rs2269715	NM_001763.2:c.204C>G	NP_001754.2:p.C68W
CM073141	chr1:67457975:+	T=>C:CC	A	DP	Psoriasis, increased risk, association with	17236132	*IL23R*	N	rs7530511	NM_144701.2:c.929T>C	NP_653302.2:p.L310P
CM084632	chr1:205366693:+	G=>A:GA	D	DFP	Haemolytic uraemic syndrome, atypical, assoc. with	18424762	*C4BPA*	Y	rs45574833	NM_000715.3:c.719G>A	NP_000706.1:p.R240H
CM100611	chr1:12005513:+	A=>G:GG	A	DFP	Breast cancer, reduced risk, association with	20103646	*MIIP*	N	rs2295283	NM_021933.2:c.499A>G	NP_068752.2:p.K167E
CM920010	chr1:228912417:-	T=>C:CC	A	DP	Hypertension, association with	1394429	*AGT*	N	rs699	NM_000029.2:c.803T>C	NP_000020.1:p.M268T
CM980072	chr1:21767322:+	T=>C:CC	A	DFP	Hypophosphatasia, association with	9781036	*ALPL*	N	rs3200254	NM_000478.3:c.787T>C	NP_000469.3:p.Y263H
CM066604	chr2:230758959:-	C=>T:TC	D	DP	Tuberculosis, reduced susceptibility, association with	16803959	*SP110*	Y	rs3948464	NM_004509.2:c.1274C>T	NP_004500.2:p.S425L
CM073086	chr2:85634047:-	G=>A:GA	D	DP	Higher body mass index, association with	17029979	*GGCX*	Y	rs699664	NM_000821.3:c.974G>A	NP_000812.2:p.R325Q
CM087379	chr2:100957736:+	A=>G:GG	A	FP	Higher testosterone levels, association with	18990770	*NPAS2*	N	rs2305160	NM_002518.3:c.1180A>G	NP_002509.2:p.T394A
CM065290	chr3:187925712:+	T=>C:CC	A	DP	Nephropathy, reduced risk, association with	17065357	*KNG1*	N	rs1656922	NM_001102416.1:c.533T>C	NP_001095886.1:p.M178T
CM066581	chr3:126109714:-	A=>C:CC	A	DP	Ulcerative colitis, association with	17058067	*MUC13*	N	rs1127233	NM_033049.2:c.1506A>C	NP_149038.2:p.R502S
CM096382	chr3:46476217:-	A=>G:GG	A	DFP	Periodontitis, aggressive, association with	18973542	*LTF*	N	rs1126478	NM_002343.2:c.140A>G	NP_002334.2:p.K47R
CM941277	**chr3:172214994:-**	**C=>T:CT**	**D**	**DP**	**Diabetes, NIDDM, association with**	**8063045**	***SLC2A2***	**N**	**rs5400**	**NM_000340.1:c.329C>T**	**NP_000331.1:p.T110I**
CM031390	**chr4:141708518:-**	**G=>A:AA**	**A**	**DP**	**Waist-to-hip ratio, association with**	**12756473**	***UCP1***	**Y**	**rs45539933**	**NM_021833.3:c.190G>A**	**NP_068605.1:p.A64T**
CM094340	chr4:38476105:-	A=>G:GG	A	DFP	Leprosy, association with	19456232	*TLR1*	N	rs4833095	NM_003263.3:c.743A>G	NP_003254.2:p.N248S
CM060415	chr6:150156438:+	A=>G:GA	D	FP	Reduced stability, association with	9343375	*PCMT1*	N	rs4816	NM_005389.1:c.358A>G	NP_005380.1:p.I120V
CM072043	chr6:160462998:+	C=>T:CT	D	FP	Reduced metformin uptake, association with	17476361	*SLC22A1*	N	rs34447885	NM_003057.2:c.41C>T	NP_003048.1:p.S14F

HGMD Acc. No.	Chromosomal location	Mutation	State	Tag	Human_disease_state	pmid	Gene	CpG	dbsnp id.	HGVS (cDNA) nomenclature	HGVS (protein) nomenclature
CM074911	chr6:39433056:-	T=>C:CC	A	DP	Coronary heart disease, association with	18073581	*KIF6*	N	rs20455	NM_145027.4:c.2155T>C	NP_659464.3:p.W719R
CM993455	**chr6:132214061:+**	**A=>C:CC**	**A**	**DFP**	**Insulin resistance/obesity, association with**	**10480624**	***ENPP1***	**N**	**rs1044498**	**NM_006208.1:c.361A>C**	**NP_006199.1:p.K121Q**
CM031368	chr7:141319814:-	G=>C:CC	A	DP	Phenylthiocarbamide taste sensitivity, association	12595690	*TAS2R38*	N	rs713598	NM_176817.2:c.145G>C	NP_789787.2:p.A49P
CM031370	chr7:141319073:-	A=>G:GG	A	DP	Phenylthiocarbamide taste sensitivity, association with	12595690	*TAS2R38*	N	rs10246939	NM_176817.2:c.886A>G	NP_789787.2:p.I296V
CM033585	chr7:86998554:-	T=>G:GG	A	DP	Inflammatory bowel disease, association with	14610718	*ABCB1*	N	rs2032582	NM_000927.3:c.2677T>G	NP_000918.2:p.S893A
CM930596	**chr7:94775382:-**	**A=>G:GG**	**A**	**DFP**	**Longevity, association with**	**15050299**	***PON1***	**N**	**rs662**	**NM_000446.3:c.575A>G**	**NP_000437.3:p.Q192R**
CM024569	chr8:18124476:+	T=>G:GG	A	FP	Increased activity, association with	12172214	*NAT1*	N	rs4986783	NM_000662.4:c.640T>G	NP_000653.3:p.S214A
CM081694	chr8:6466450:+	C=>T:TT	A	DP	Cranial volume, association with	18204051	*MCPH1*	Y	rs1057090	NM_024596.2:c.2282C>T	NP_078872.2:p.A761V
CM081761	chr8:143758933:+	C=>T:TT	A	DFP	Gastric cancer, diffuse-type, association with	18488030	*PSCA*	Y	rs2294008	NM_005672.3:c.2C>T	NP_005663.1:p.T1M
CM099178	chr8:118899878:-	G=>A:GA	D	DM	Multiple osteochondromas	19810120	*EXT1*	Y		NM_000127.2:c.1609G>A	NP_000118.2:p.V537I
CM950017	chr8:37942955:-	T=>C:CC	A	DFP	Hyperinsulinaemia, association with	7487991	*ADRB3*	N	rs4994	NM_000025.1:c.190T>C	NP_000016.1:p.W64R
CM940804	chr9:34639442:+	A=>G:GG	A	DFP	Galactosaemia, Duarte variant	8198125	*GALT*	N	rs2070074	NM_000155.2:c.940A>G	NP_000146.2:p.N314D
CM990005	**chr9:106626574:-**	**A=>G:GG**	**A**	**FP**	**Higher plasma HDL cholesterol, association with**	**10431237**	***ABCA1***	**N**	**rs2066714**	**NM_005502.2:c.2649A>G**	**NP_005493.2:p.I883M**
CM067461	chr10:81691702:-	A=>G:GG	A	DP	Lung cancer, susceptibility to, association with	16741161	*SFTPD*	N	rs2243639	NM_003019.4:c.538A>G	NP_003010.4:p.T180A
CM074765	chr10:67710331:-	G=>A:GA	D	DP	Alzheimer disease, increased risk, association with?	17209133	*CTNNA3*	N	rs4548513	NM_013266.1:c.1787G>A	NP_037398.1:p.S596N
CM025891	chr11:74585230:+	C=>T:TT	A	FP	Decreased enzyme activity, association with	12130747	*SLCO2B1*	N	rs2306168	NM_007256.2:c.1457C>T	NP_009187.1:p.S486F
CM080415	chr11:113308238:+	A=>C:CC	A	FP	Altered receptor function, association with	18184810	*HTR3B*	N	rs1176744	NM_006028.3:c.386A>C	NP_006019.1:p.Y129S
CM033453	chr12:107542027:-	G=>A:AA	A	DFP	Incident coronary heart disease, decreased risk in African Americans,	17420019	*SELPLG*	N	rs2228315	NM_003006.3:c.186G>A	NP_002997.1:p.M62I

HGMD Acc. No.	Chromosomal location	Mutation	State	Tag	Human_disease_state	pmid	Gene	CpG	dbsnp id.	HGVS (cDNA) nomenclature	HGVS (protein) nomenclature
CM085048	chr12:78539038:-	T=>G:GG	A	DP	association with Schizophrenia in females, association with	18281137	*PAWR*	N		NM_002583.2:c.597T>G	NP_002574.2:p.I199M
CM950862	chr12:5473868:+	G=>A:GA	D	DP	Schizophrenia, severe, increased risk, association with	7733919	*NTF3*	N	rs1805149	NM_002527.4:c.227G>A	NP_002518.1:p.G76E
CM994637	chr12:6327323:-	A=>G:GG	A	DFP	Hypertension, reduced risk, association with	10523338	*SCNN1A*	N	rs2228576	NM_001038.4:c.1987A>G	NP_001029.1:p.T663A
CM022034	chr13:32526193:+	G=>C:GC	D	DP	Age-related phenotypes, association with	11792841	*KL*	N	rs9527025	NM_004795.2:c.1109G>C	NP_004786.2:p.C370S
CM033777	chr14:24170122:-	T=>C:CC	A	DP	Apoptosis, unable to induce, association with	12594335	*GZMB*	N	rs2236338	NM_004131.3:c.739T>C	NP_004122.1:p.Y247H
CM070246	chr14:60993992:+	G=>A:AA	A	DFP	Cerebral infarction, association with	17206144	*PRKCH*	N	rs2230500	NM_006255.3:c.1120G>A	NP_006246.2:p.V374I
CM067475	chr15:41555066:-	C=>G:GG	A	DP	Lung cancer, susceptibility to, association with	16741161	*TP53BP1*	N	rs560191	NM_005657.1:c.1059C>G	NP_005648.1:p.D353E
CM067476	chr15:41511938:-	A=>C:CC	A	DP	Lung cancer, susceptibility to, association with	16741161	*TP53BP1*	N	rs2602141	NM_005657.1:c.3406A>C	NP_005648.1:p.K1136Q
CM085365	chr15:43185730:-	A=>G:GG	A	DM	Hypothyroidism	18765513	*DUOX2*	N		NM_014080.4:c.2033A>G	NP_054799.4:p.H678R
CM983400	chr16:27263704:+	A=>G:GG	A	DFP	Asthma, atopic, association with	9620765	*IL4R*	N	rs1805010	NM_000418.2:c.223A>G	NP_000409.1:p.I75V
CM030773	chr17:19753133:-	A=>G:GG	A	DP	Cardiac disease, susceptibility to, association	12646697	*AKAP10*	N	rs203462	NM_007202.2:c.1936A>G	NP_009133.2:p.I646V
CM032397	chr17:41432502:+	A=>G:AG	D	DP	Progressive supranuclear palsy, association with	12913211	*STH*	N	rs62063857	NM_001007532.1:c.20A>G	NP_001007533.1:p.Q7R
CM057933	chr17:4585312:-	C=>T:CT	D	DP	Atherosclerotic stenosis, increased severity, association with	15836657	*CXCL16*	N	rs2277680	NM_022059.2:c.599C>T	NP_071342.2:p.A200V
CM077855	chr17:7532893:+	C=>G:GG	A	DP	Breast cancer, ER negative, association with?	17683073	*WRAP53*	N	rs2287499	NM_018081.1:c.202C>G	NP_060551.1:p.R68G
CM087381	chr17:7987497:-	G=>C:CC	A	FP	Increased sex hormone-binding globulin levels, association with	18990770	*PER1*	N	rs2585405	NM_002616.1:c.2884G>C	NP_002607.1:p.A962P
CM093418	chr17:39581073:+	A=>C:CC	A	DP	Hip bone mineral density, association with?	19079262	*C17orf53*	N	rs227584	NM_024032.2:c.376A>C	NP_076937.2:p.T126P

HGMD Acc. No.	Chromosomal location	Mutation	State	Tag	Human_disease_state	pmid	Gene	CpG	dbsnp id.	HGVS (cDNA) nomenclature	HGVS (protein) nomenclature
CM004814	chr19:50546759:-	A=>C:CC	A	DFP	Basal cell carcinoma, reduced risk, association with	9950243	*ERCC2*	N	rs13181	NM_000400.2:c.2251A>C	NP_000391.1:p.K751Q
CM984025	chr19:18047618:-	A=>G:AG	D	DM	Mycobacterial infection	9603732	*IL12RB1*	N	rs11575934	NM_005535.1:c.641A>G	NP_005526.1:p.Q214R
CM014824	chr20:4653718:+	C=>T:TT	A	DP	Creutzfeldt-Jakob disease, association with	11702213	*PRND*	Y	rs2245220	NM_012409.2:c.521C>T	NP_036541.2:p.T174M
CM064121	chr20:44075813:+	G=>C:CC	A	DP	Leukemia, risk, association with	16574953	*MMP9*	N	rs2250889	NM_004994.2:c.1721G>C	NP_004985.2:p.R574P
CM025479	**chr21:44534334:+**	**C=>G:GG**	**A**	**DP**	**Alopecia universalis, association with**	**12542742**	***AIRE***	**N**	**rs1800520**	**NM_000383.2:c.834C>G**	**NP_000374.1:p.S278R**
CM057711	chr21:33536125:+	T=>G:GG	A	DP	Multiple sclerosis, susceptibility to, association with	15885318	*IFNAR2*	N	rs1051393	NM_207585.1:c.28T>G	NP_997468.1:p.F10V
CM057927	chr22:21957369:+	A=>G:GG	A	DP	Bipolar disorder, association with?	15866548	*BCR*	N	rs140504	NM_004327.3:c.2387A>G	NP_004318.3:p.N796S
CM096696	chr22:35792882:-	T=>C:CC	A	DP	Iron status and erythrocyte volume, association with	19820699	*TMPRSS6*	N	rs855791	NM_153609.2:c.2207T>C	NP_705837.1:p.V736A
CM910052	chr22:49410905:-	C=>G:GG	A	DP	Phenotype modifier, association with?	11941485	*ARSA*	N	rs743616	NM_000487.4:c.1172C>G	NP_000478.2:p.T391S
CM961339	chr22:30836050:+	C=>G:CG	D	DM	Glucose/galactose malabsorption	8563765	*SLC5A1*	N	rs33954001	NM_000343.1:c.1845C>G	NP_000334.1:p.H615Q
CM085353	chrX:149390017:+	T=>C:CC	A	DM	Hypospadias	18635673	*MAMLD1*	N		NM_005491.2:c.1514T>C	NP_005482.2:p.V505A

Mutation: human normal =>human mutation: Neanderthal genotype, chimpanzee genotype. HGMD: Human Gene Mutation Database. A: Ancestral, D: Derived. DM: disease-causing mutation; DFP: disease-associated functional polymorphism; DP: disease-associated polymorphism; FP: functional polymorphism. CpG column provides information as to whether the corresponding mutation occurred in a CpG dinucleotide in human and was either a C>T or G>A transition compatible with methylation-mediated deamination of 5-methylcytosine (Y denotes yes, N denotes no). HGVS (cDNA) nomenclature provides cDNA numbering with +1 corresponding to the A of the ATG translational initiation codon in the corresponding reference sequence. HGVS (protein) nomenclature provides numbering with respect to the corresponding reference protein sequence, with the translational initiation codon being codon 1. Entries marked in bold type were previously identified as 'compensated mutations' by Mikkelsen et al. (2005). These are putative disease-causing/disease-associated variants in human that correspond to the wild-type in chimpanzee.

Supp. Table S2. Potential compensated regulatory mutations in the Neanderthal and/or chimpanzee genome

HGMD Acc. No.	Chromosomal location	Mutation	State	Tag	Human_disease_state	pmid	Gene	CpG	dbsnp
CR060579	chr1:111020443:-	A=>G:GA	D	DP	Low insulin sensitivity, association with	16317062	*KCNA3*	N	rs2821557:A>G
CR043164	chr1:43575707:+	C=>A:AA	A	DP	Platelet count, association with?	15307100	*MPL*	N	rs839993:C>A
CR057791	chr1:111571946:+	G=>T:TG	D	FP	Increased promoter activity, association with	16251966	*CHI3L2*	N	rs755467:G>T
CR025943	chr1:228917021:-	C=>T:CT	D	DP	Increased angiotensinogen levels, association with?	12404103	*AGT*	N	rs5046:C>T
CR025220	chr2:234330398:+	T=>G:GG	A	DFP	Hyperbilirubinaemia, association with	11906189	*UGT1A1*	N	rs4124874:T>G
CR033690	chr2:227372145:-	G=>A:AA	A	DP	Diabetes, NIDDM, association with	14633864	*IRS1*	N	rs13306465:G>A
CR086331	chr2:234291987:+	C=>T:CT	D	FP	Reduced expression, association with	18433817	*UGT1A4*	N	rs3732219:C>T
CR066664	chr3:129680794:-	C=>T:CT	D	DP	Coronary artery disease, association with	16934006	*GATA2*	Y	rs2713579:C>T
CR032439	chr3:12328198:+	C=>G:GG	A	DFP	Increased height/lipid metabolism, association with	12588773	*PPARG*	N	rs10865710:C>G
CR014438	chr3:185572960:-	G=>A:AA	A	DP	Myocardial infarction, association with	11257273	*THPO*	Y	rs6141:G>A
CR025435	chr4:111053559:+	A=>G:GG	A	DFP	Malignant melanoma, association with	11844511	*EGF*	N	rs4444903:A>G
CR004797	chr4:26101320:-	G=>T:GT	D	DP	Higher percent body fat, association with	10682840	*CCKAR*	N	rs1800908:G>T
CR045948	chr4:69995928:+	G=>A:AA	A	FP	Promoter activity, association with	15001974	*UGT2B7*	Y	rs7438135:G>A
CR071281	chr4:156348632:+	C=>T:TT	A	DP	Obesity, association with	17235527	*NPY2R*	Y	rs6857715:C>T
CR035513	chr5:131436741:+	A=>C:CC	A	DP	Reduced severity in atopic dermatitis, association with	13679820	*CSF2*	N	rs4124874:T>G
CR057231	chr5:71047268:+	T=>C:CC	A	DP	Obesity, association with?	15823203	*CARTPT*	N	rs4703647:T>C
CR071289	chr5:1499389:-	T=>C:CC	A	DP	Attention-deficit hyperactivity disorder, association with	17044101	*SLC6A3*	N	rs2652511:T>C
CR052970	chr6:132254387:+	A=>G:GG	A	DP	Obesity, association with	16025115	*ENPP1*	N	rs7754561:A>G
CR082018	chr6:78227843:-	G=>A:GA	D	DFP	Aggressive behaviour, association with	18283276	*HTR1B*	Y	rs13212041:G>A
CR077383	chr6:154401054:+	A=>G:GG	A	FP	Increased promoter activity, association with	16843022	*OPRM1*	N	rs17174629:A>G
CR075274	chr6:133077018:-	T=>G:GG	A	DP	HDL cholesterol concentration, association with	17873875	*VNN1*	N	rs4897612:T>G
CR075243	chr6:132314950:-	G=>C:CG	D	DFP	Systemic sclerosis, association with	17881752	*CTGF*	N	rs6918698:G>C
CR073540	chr6:131935252:+	G=>T:TT	A	DP	Myocardial infarction, association with	17369504	*ARG1*	N	rs2781666:G>T
CR012231	chr6:32260420:-	T=>C:CC	A	DFP	Diabetic retinopathy, association with	11375354	*AGER*	N	rs1800625:T>C
CR092300	chr7:111902894:+	C=>T:TT	A	DFP	Severity in cystic fibrosis, association with	19242412	*IFRD1*	N	rs7817:C>T
CR068449	chr7:128381961:+	C=>T:TT	A	DP	Systemic lupus erythematosus, association with?	16642019	*IRF5*	Y	rs2280714:C>T
CR066667	chr7:30969948:+	C=>T:TT	A	DP	Breast cancer, decreased risk, association with	16606630	*GHRHR*	Y	rs2302019:C>T
CR971950	chr8:19840951:+	T=>G:GG	A	FP	Lower plasma triglyceride level, association with	9017514	*LPL*	N	rs1800590:T>G

HGMD Acc. No.	Chromosomal location	Mutation	State	Tag	Human_disease_state	pmid	Gene	CpG	dbsnp
CR034594	chr9:124172343:+	A=>G:GG	A	FP	Inhibition of prostaglandin H2 formation, association with?	12545150	*PTGS1*	N	rs10306114:A>G
CR102176	chr9:100952292:+	A=>G:GG	A	DFP	Breast cancer, association with	20332227	*TGFBR1*	N	rs334348:A>G
CR091269	chr9:116608587:-	T=>C:CC	A	DFP	Crohn's disease, susceptibility to, association with	19124533	*TNFSF15*	N	rs6478109:T>C
CR054255	chr9:127043845:-	A=>G:GG	A	DP	Bipolar disorder, association with?	16168956	*HSPA5*	N	rs391957:A>G
CR045560	chr9:106730659:-	G=>C:CC	A	FP	Reduced plasma HDL cholesterol, association with	15262183	*ABCA1*	N	rs2246293:G>C
CR072313	chr10:94452862:+	C=>T:CT	D	DP	Diabetes, type 2, association with?	17293876	*HHEX*	N	rs1111875:C>T
CR942079	chr10:104587142:-	T=>C:CC	A	DP	Polycystic ovaries, association with	7849715	*CYP17A1*	N	rs743572:T>C
CR102882	chr10:64279946:-	G=>A:GA	D	DFP	Systemic lupus erythematosus, association with	20194224	*EGR2*	N	rs1412554:G>A
CR094845	chr11:74539529:+	G=>A:AA	A	FP	Increased mRNA expression, association with	19620935	*SLCO2B1*	N	rs2712807:G>A
CR096333	chr11:85546288:-	A=>G:GG	A	DP	Alzheimer disease, association with?	19734902	*PICALM*	N	rs3851179:A>G
CR035965	chr11:45863406:+	A=>G:GG	A	DFP	Alzheimer disease, association with	12740599	*MAPK8IP1*	N	rs1554338:A>G
CR025510	chr11:102331749:-	G=>A:GA	D	FP	Increased transcriptional activity, association with	12392760	*MMP13*	N	rs2252070:G>A
CR082031	chr12:55796928:-	G=>A:AA	A	DP	Schistosomiasis infection, association with	18273035	*STAT6*	N	rs324013:G>A
CR031478	chr12:10203556:-	C=>T:CT	D	DP	Alzheimer disease, reduced risk, association with	12807963	*OLR1*	N	rs1050283:C>T
CR080758	chr13:45577313:-	A=>G:AG	D	FP	Increased promoter activity, association with	17855631	*CPB2*	N	rs11574980:A>G
CR994765	chr13:112807756:+	G=>T:TC	N	DFP	Reduced plasma F7 levels, association with	10233895	*F7*	N	Rs510335:G>T
CR045986	chr14:51803734:+	T=>C:CC	A	DFP	Asthma, association with	15496624	*PTGDR*	N	rs8004654:T>C
CR993820	chr15:72828970:+	C=>A:AA	A	DFP	Increased activity in smokers, association with	10233211	*CYP1A2*	N	rs762551:C>A
CR066661	chr15:49336891:-	C=>T:TT	A	DP	Alzheimer's, in APOE4 carriers, increased risk, association with	16882736	*CYP19A1*	Y	rs1008805:C>T
CR002154	chr15:56511231:+	G=>A:GA	D	DP	Dyslipidaemia and insulin resistance, association	10894818	*LIPC*	N	rs2070895:G>A
CR994768	chr17:41327398:+	G=>C:CG	D	DP	Supranuclear palsy, progressive, association with	10580705	*MAPT*	N	rs62056778:G>C
CR084012	chr17:25549137:-	T=>G:GG	A	FP	Increased expression, association with	18445138	*SLC6A4*	N	rs1042173:T>G
CR052976	chr17:43163827:+	T=>C:CC	A	DP	Asthma, aspirin-induced, association with	15806396	*TBX21*	N	rs4794067:T>C
CR078280	chr17:35323475:-	G=>A:GA	D	DP	Asthma, increased risk, association with?	17611496	*GSDMB*	Y	rs7216389:G>A
CR090198	chr17:38531642:-	A=>G:GG	A	FP	Promoter activity, association with	18782836	*BRCA1*	N	rs799906:A>G
CR010588	chr19:60077416:+	T=>C:CC	A	DP	IgA nephropathy, association with	11281451	*FCAR*	N	rs3816051:T>C
CR051707	chr19:7718733:-	T=>C:CC	A	DFP	Dengue disease, prot. against, association with	15838506	*CD209*	N	rs4804803:T>C
CR050427	chr19:46188969:+	T=>C:CC	A	FP	CYP2B6 expression, association with	15722458	*CYP2B6*	N	rs34223104:T>C
CR075263	chr20:17370063:+	T=>C:CC	A	DP	Diabetes, type 2, association with	17618154	*PCSK2*	N	rs2021785:T>C
CR054260	chr21:38590628:+	T=>G:TG	D	FP	Promoter activity, association with	16086313	*KCNJ15*	N	rs2236606:T>G

HGMD Acc. No.	Chromosomal location	Mutation	State	Tag	Human_disease_state	pmid	Gene	CpG	dbsnp
CR063398	chrX:135554616:+	A=>G:GG	A	FP	Increased soluble CD40L levels, association with	16627810	*CD40LG*	N	rs3092952:A>G
CR077381	chrX:113724838:+	G=>C:GC	D	FP	Reduced promoter activity, association with	17376412	*HTR2C*	N	rs518147:G>C

Mutation: human normal =>human mutation: Neanderthal genotype, chimpanzee genotype. HGMD: Human Gene Mutation Database. A: Ancestral, D: Derived. DM: disease-causing mutation; DFP: disease-associated functional polymorphism; DP: disease-associated polymorphism; FP: functional polymorphism. CpG column provides information as to whether the corresponding (human) mutation occurred in a CpG dinucleotide and was either a C>T or G>A transition compatible with methylation-mediated deamination of 5-methylcytosine (Y denotes yes, N denotes no). The dbSNP column refers to the dbSNP identifier and base change (dbSNP:wildtype>mutant).