# Supporting Online Material for

## Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture

Hernán A. Burbano,* Emily Hodges, Richard E. Green, Adrian W. Briggs, Johannes Krause, Matthias Meyer, Jeffrey M. Good, Tomislav Maricic, Philip L. F. Johnson, Zhenyu Xuan, Michelle Rooks, Arindam Bhattacharjee, Leonardo Brizuela, Frank W. Albert, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Michael Lachmann, Gregory J. Hannon, Svante Pääbo

*To whom correspondence should be addressed. E-mail: hernan_burbano@eva.mpg.de

Published 7 May 2010, *Science* **328**, 723 (2010)
DOI: 10.1126/science.1188046

**This PDF file includes:**

Supplementary Information

This document contains (in order):
Material and methods
Figures S1-S4
Tables S1-S5

Materials and Methods

Neandertal DNA extraction and library preparation

We extracted DNA as previously described (S*1*) under clean room conditions from the Sidron 1253 bone, which was excavated in 2006 by Javier Fortea and his team (S*2*) in El Sidrón cave, Asturias, Spain. In order to reduce the amount of ancient DNA-associated misincorporations caused principally by cytosine deamination to uracil (S*3, S4*), extracts were treated using a protocol that includes uracil-DNA glycosylase and endonuclease VIII. These enzymes remove uracils from ancient DNA and repairs most of the resulting abasic sites, but leaves undamaged parts of the DNA fragments intact (S*5*). We prepared in total five 454 libraries as previously described (S*6*), from 60 µl DNA extract each,by the ligation of sequencing adaptors carrying a project specific key sequence (TGAC) (S*4*). The 300 µl extract used for the library preparation come from a total of 867 mg of Sidron 1253 bone.

Identification of non-synonymous substitutions (NSS) that occurred on the human lineage

In order to ascertain the NSS that occurred in the human lineage, we use the human-chimpanzee homology annotation from NCBI HomoloGene (Build 58). HomoloGene is a system for automated detection of homologs (including orthologs and paralogs) among the annotated genes of several completely sequenced eukaryotic genomes (S*7*). Between two given species each HomoloGene entry can have a gene homology mapping one to one, one to many, many to one, or many to many. We selected 15,857 HomoloGene entries where the gene homology mapping between human and chimpanzee was one to one. We note that this strategy precludes human and chimpanzee genes that have close paralogs. We then aligned the protein sequences corresponding to each homologous gene pair using the RefSeq protein ID reported by HomoloGene. The sequences were aligned with the multiple sequence alignment software Muscle (S*8*) using default parameters. We employ the protein pair wise alignments to generate a list of human/chimpanzee NSS.

To infer on which lineage, human or chimpanzee, each NSS occurred, we aligned independently all orangutan proteins present in Ensembl (assembly version PPYG2, database version 49) against both all human and all chimpanzee RefSeq proteins (release 7) using the FASTA package program ssearch3, which implements the Smith-Watermann algorithm (S*9*). For each human and chimpanzee protein we selected the orangutan hit with the highest bit score. Human lineage NSS were inferred by parsimony, taking all positions where chimpanzee and orangutan agreed but the human amino acid was different. Using this methodology we located 13,841 NSS to the human evolutionary lineage.

Array design

To capture Neandertal libraries we used Agilent custom one million feature capture arrays. We designed overlapping microarray probes of 60 bases targeting the NSS and flanking regions of 50 bases on each side of them. Probes were tiled every three bases across the target regions. Probes containing repetitive elements – detected by 15-mer frequency counts - were discarded (S*10*). The sequence used to design the probes was the reference human sequence NCBI Build 36.1 (hg18). Note that nucleotide differences and small indels are unlikely to influence capture when tiling arrays are used (S*11, S12*). In addition to probes targeting the NSS, we included on the array probes to estimate the level of modern human contamination, including the human mitochondria sequence, with 10 bp tiling, and 46 chromosome X control positions (see below) with 3bp tiling.

Capture and sequencing of Neandertal libraries

Before array capture three of the five Neandertal 454 libraries were combined into a single library. The resulting 454 libraries were PCR amplified in order to obtain a total of 20µg of library DNA for each array capture experiment. For this purpose all libraries were amplified by splitting them up into 10 PCR reactions each in order to avoid PCR saturation effects. Each 100 µl reaction contained 4 Units of *Taq* Gold polymerase (Applied Biosystems, Foster City, CA, USA), 1X Amplitaq Gold buffer (Applied Biosystems, Foster City, CA, USA), 2.5 mM $MgCl_2$ (Applied Biosystems, Foster City, CA, USA), 1 mg/ml BSA, 250 µM of each dNTP (New England Biosystems, Ispwich, MA, USA) and 1µM of each 454-emulsion PCR primer (Sigma-Aldrich, St. Louis, MO, USA). Annealing temperature was 60°C and a total of 10 cycles of PCR was performed. After amplification, PCR products were spin column purified using the Quiagen Minelute system (Qiagen, Hilden, Germany) and all amplification products originating from the same library were pooled resulting in total of 100µl. A second round of PCR was performed by splitting up the first amplification products into 20 separate PCR reactions. The libraries were amplified in a 100µl reaction containing 50µl Phusion™ High-Fidelity Master Mix (Finnzymes, Espoo, Finland), 1µM of each 454-emulsion PCR primer and 2.5µl library template from the previous amplification. Annealing temperature was 60°C and a total of 7 cycles of PCR were performed. PCR products were spin column purified, products originating from the same library were pooled and quantified with a Nanodrop (Thermo Fischer Scientific, Waltham, MA, USA). The three independent amplified libraries each containing 20µg of amplified library were used as template for three parallel experiments as described in (S*13*) (step 29 –step 58). In each case, two successive captures were performed. After the first round of capture the 490µl eluate from each array were amplified in 12 independent PCRs in a 100µl reaction containing 50µl Phusion™ High-Fidelity Master Mix, 1µM of each 454-emulsion PCR primer and 40 µl library template and a total of 20 cycles. PCR products were spin column purified and quantified on a Nanodrop™. A second round of array capture was performed using the same conditions as in the first round.

To sequence the libraries using the Illumina platform, we converted the libraries to the Illumina format after the second round of array capture by amplifying them with primers that are complementary to the 454 adaptors in their 3'-ends and carry Illumina adaptor sequences in their 5'-ends as described (S*14*). The converted Illumina/454 library product

was spin column purified and quantified on a Agilent DNA1000 chip (Agilent, Santa Clara, CA, USA). The captured libraries were sequenced in seven lanes on the Illumina Genome Analyzer II platform (Illunina, San Diego, CA, USA) together with a control lane of PhiX 174 variant, which carries a compatible adapter sequences with the 454 standard key (TCAG). The sequencing run was performed according to manufacturer's intructions for a paired-end sequencing run with 2x51 cycles and FC-104-100x chemistry. Instead of the standard Genomic R1 and R2 sequencing primers, project-specific primers were used for the forward and reverse sequencing reads. These primers anneal to the 454 adaptor sequences and allow for sequencing a project-specific key at the beginning of each read. To facilitate the estimation of base caller parameters of the Illumina base caller (Bustard 1.3.2), the control lane used a sequencing primer covering the standard key in the first read (see also Processing and mapping of Neandertal reads).

Capture and sequencing of HGDP libraries

We selected 50 individuals from 50 different populations of the Human Genome Diversity Panel (HGDP) (CEPH, Paris, France) (Table S1). From these individuals we produce genomic libraries that were barcoded, pooled and captured on a single array (S*15*). The HGDP libraries were sequenced in five lanes on the Illumina Genome Analyzer II platform together with a control lane of a PhiX 174 variant, which carries compatible adapter sequences. The sequencing run was performed according to the manufacturer's instructions (for the FC-103-300x chemistry) as a multiplexed paired-en sequencing run of 2x101 cycles plus 7nt index/barcode read. Instead of the DNA polymerase provided with the FC-103-300x sequencing kits, the DNA polymerase provided with the FC-104-40xx kits was used.

Processing and mapping of Neandertal reads

Sequencing runs were analyzed from raw images using Illumina Genome Analyzer pipeline 1.3.2. To overcome analytical challenges introduced by identical key sequences at the beginning of the first read, we used the first five (instead of two) sequencing cycles for cluster identification. Base calling parameters estimates for the Illumina base caller (Bustard 1.3.2) were based only on the control lane. To avoid problems with the parameters estimation introduced by the key at the beginning of the second read, the run was based called as a single run with 102 cycles resetting phasing at cycle 52. The obtained sequences for the PhiX control lane were used to train cycle-specific models for the alternative base caller *Ibis* (S*16*). The models were then applied to base call the complete run.

After base calling, raw sequences of forward and reverse reads were filtered for three bases of the project specific key ('GAC') at the beginning of both reads. The two reads of each cluster were then merged (including adapter removal) requiring at least an 11 nucleotide overlap. For bases within the overlapping part of the sequence, the consensus sequence was called by considering the base with the higher quality score or, in case of agreement, summing up the quality scores. For further analysis, only successfully merged sequences were considered.

To assemble the mitochondrial (mt) DNA we used an iterative mapping assembly (MIA) program as described previously (S*6, S17*). We used as a reference the mtDNA sequence of Sidrón 1253 (S*6*) obtaining the same sequence after our iterative mapping assembly.

To learn the Neandertal state at the NSS we employed MIA in a non-iterative way. For each NSS we used as a reference 110 bases (hg18) flanking regions on each side of the NSS. To avoid mapping problems due to sequence similarity, we aligned the references to the human genome using BLAT (S*18*). If the second best hit had a percentage identity to the reference greater or equal than 80%, the substitutions were not included in the further analysis.

Sequence redundancy, a particular problem in this case because each hybridization step is followed by library re-amplification, was identified and removed by clustering reads based on start and end genomic coordinates. Within each cluster the read with the highest sum of quality scores was selected as cluster representative.

Because reads are aligned with MIA only against the references, it is not possible to know if a read align uniquely to regions of hg18 outside the references. Therefore, we also mapped the reads to the hg18 using PatMaN (S*19*), allowing for two mismatches. We removed reads that were located uniquely to a region of hg18 outside the references by PatMaN. Reads that could not be mapped by PatMaN because they had no hit with two or fewer mismatches but that mapped by MIA to a reference sequences were kept. Finally, if a read was aligned by MIA to more than one position in the references, the aligned to the reference with the highest alignment quality score was used. We used the same method to learn the Neandertal state of the chromosome X control positions.

For the 13,250 positions for which we collected data from the Neandertal, we did not analyze 2,298, or 17.4%. This is because we refrained from using 825 (6.2%) positions that carried both ancestral and derived alleles (to avoid errors in the large number of derived alleles causing apparent ancestral alleles); 393 (3.0%) positions where we did not call data from at least 25 HGDP individuals; and 254 (1.9%) positions where we found only a third state, both ancestral and a third state, or a triallelic state. In addition, second best hit filtering of the targeted regions removed 826 (6.2%) positions.

Processing, mapping and genotype calling of HGDP reads

The run was analyzed based on SCS2.4/RTA intensity files. Control lane reads were base called using Bustard 1.4.0 and then used to create cycle-specific models for the base caller *Ibis* (S*16*). These models were then applied for the base calling of all lanes. Reads from the lanes with HGDP libraries were assigned to a HGDP individual according to the individual-specific barcode/index sequence (allowing for one sequencing error and the loss of the first base (S*15*)). If possible, forward and reverse reads were merged as it was done with the Neandertal reads and they were further processed as single-end reads. Succesfully merged reads were further processed as single-end reads, whereas reads that were not merged were processed as paired-end reads. All reads were aligned against hg18 using bwa (S*20*). We removed duplicates and merge single- and pair-end reads in a single BAM file per individual (S*21*). We called genotypes using the pileup command of SAMtools (S*21*), which includes the same genotype calling algorithm used by MAQ (S*22*). For each

individual we used a position if its genotype call had a phred-like quality score equal or greater than 20. A position was considered fixed in modern humans if it was found homozygous and derived, and data was available for at least 25 HGDP individuals (50 chromosomes).

Authenticity estimate – mtDNA

It is known from previous work that El Sidrón 1253 mtDNA differs from modern humans' mtDNA at 130 positions (S*6*). We estimated the level of modern human mtDNA of the Sidrón 1253 libraries before capture and Illumina sequencing using primer extension and capture (PEC) (S*6*). We targeted six mtDNA informative positions between human and Neandertals, where the Neandertal was found to be different from a panel of 311 aligned modern human mtDNA genomes (S*17*). The captured library molecules were then sequenced on the 454 FLX platform. A total of 748 mtDNA fragments that carry informative positions were captured and none of them carried sequences that suggest that they are of modern human origin. From this, we derived an estimate of mtDNA contamination of 0% (C.I. 0-0.5 %).

We employed the same strategy to estimate the level of mtDNA contamination after capture and sequencing (S*6*). From a total of 254,296 mtDNA fragments, 747 were scored as modern human contaminants, giving an mtDNA contamination estimate of 0.29% (C.I. 0.27-0.31 %).

Authenticity estimate – autosomes

Our method of estimating contamination in autosomal sites uses the fact that at every position the Neandertal individual that we sequenced is either homozygous ancestral, homozygous derived, or heterozygous. Therefore, without sequencing error or contamination, we expect to see at every position either only ancestral alleles, only derived, or a draw with equal chance for ancestral and derived. Introducing contamination and sequencing error will skew these ratios.
We look at sites at which:

-At least 25 humans were sequenced, and all of them show only the derived allele. We assume that the majority of these sites will be derived in the contaminating human(s).
-Mapping the reads to the human lineage yields only a single good hit.
-If other than the ancestral or derived allele, a third allele is seen, this allele is only observed once.
-The sequences contains no gaps.

Table S5 shows the distribution of derived and ancestral allele counts for these sites. Even by manually observing this table, one can see that the contamination rate must be very low. Only two reads in the whole table are obvious candidates of originating from a human – when mostly ancestral alleles are observed, an only one derived. In all other cases, when many ancestral alleles are observed, either no derived alleles are observed, or more than one derived allele is observed. And these two contamination candidates could also be the result of sequencing error. Our method is very similar to this eyeballing estimate of

contamination, except that it tried to give a precise estimate of how many sites are heterozygous in Neandertal, and what is the contamination level, taking into account the base-specific sequencing error rate.

For these sites, define $p_{aa}$, $p_{dd}$, $p_{ad}$ the fraction of sites at which our Neandertal individual is homozygous ancestral, homozygous derived, or heterozygous. Let us call the level of contamination $c$, and the coverage $m$.

The probability to see $n_a$ ancestral reads, and $n_d$ derived is:

$$\begin{aligned}
&P(n_a, n_d \mid p_{aa}, p_{ad}, p_{dd}, c, m) = \\
&= p_{aa} P(n_a, n_d \mid aa, c, m) + p_{ad} P(n_a, n_d \mid ad, c, m) + p_{dd} P(n_a, n_d \mid dd, c, m) \\
&= p_{aa} Pois(n_a, \lambda = m(1-c)) \cdot Pois(n_d, \lambda = mc) + \\
&p_{ad} Pois(n_a, \lambda = m\frac{1-c}{2}) Pois(n_d, \lambda = m(c + \frac{1-c}{2})) + \\
&p_{dd} Pois(n_a, \lambda = 0) \cdot Pois(n_d, \lambda = m)
\end{aligned}$$

When we compare likelihoods, $m$ will not play a role. We can use $p_{dd} = (1 - p_{aa} - p_{ad})$ to get a 3-dimensional likelihood function. We can reduce the space further by writing $p_a$ for the overall chance to see the ancestral allele in our whole dataset. Since

$$p_a = (1-c)(p_{aa} + \frac{1}{2} p_{ad})$$

We get $p_{aa} = p_a/(1-c) - p_{ad}/2$. Since $p_a$ can be directly observed, and because of the large number of samples involved in its estimation, we will use the observed $p_a$ as our estimate for the real $p_a$.

It is easy to incorporate sequencing error into the above framework. If we write $E_{ad}$ and $E_{da}$ for the error rates from ancestral to derived and back, and $E_{aa}$ and $E_{dd}$ for their complements, we can write:

$$\begin{aligned}
P(n_a, n_d \mid p_{ad}, c) = &p_{aa} Pois(n_a, cE_{da} + (1-c)E_{aa}) \cdot Pois(n_d, cE_{dd} + (1-c)E_{ad}) + \\
&p_{ad} Pois(n_a, cE_{da} + \frac{1-c}{2}(E_{da} + E_{aa})) Pois(n_d, (cE_{dd} + \frac{1-c}{2}(E_{dd} + E_{ad}))) + \\
&p_{dd} Pois(n_a, E_{da}) \cdot Pois(n_d, E_{dd})
\end{aligned}$$

where now we have a single parameter $p_{ad}$. Since we are interested in likelihood ratios between different levels of contamination, and since the dependence on $m$ introduces just a multiplicative term that does not depend on the contamination level or heterozygosity, the dependence on $m$ drops out of the likelihood ratios, and we can just assume $m$=1 for simplicity. The value of $p_{ad}$ is the heterozygosity of Neandertal at sites where humans are fixed derived.

To calculate the likelihood we multiply the likelihoods of each of the sites, using the base-specific error rate. The base specific error rates were estimated from triallelic sites, where Neandertal had a single observtion of a non ancestral or non derived state
The likelihood surface can be seen in Figure S3, where is shown that the bounds on contamination do not depend very much on the actual value of $p_{ad}$, and thus our 2% upper bound on contamination is quite robust.

Authenticity estimate - chromosome X

Estimating contamination directly in the human nuclear DNA sequence is difficult since no fixed modern human/Neandertal differences are currently know. To find putative modern human/Neandertal nuclear DNA differences that could be used as contamination control sites, we used data from the Neandertal shot-gun project, which has sequenced a different Neandertal individual (Vindija 33.16) than the one used in this study (Sidron 1253). From the genome project data, we selected 145 transversions on chromosome X where Vindija 33.16 was found different from the human reference, and where the positions were not polymorphic in present-day humans according to dbSNP. These positions were then genotyped in ~1000 individuals from the Human Genome Diversity Panel (HGDP-CEPH) using ABI SNPlex and Sequenom genotyping systems. We found that 46 transversions were fixed derived in humans. Probes targeting these positions were included in the capture array tiled in the same way as probes targeting the NSS.

A very similar approach as the one decribed for the autosomes will yield an estimate of contamination using the X chromosome, except that since our Neandertal is a male (S*23*), only one copy of each allele is present, and thus without contamination we expect to see either all ancestral, or all derived.

We can therefore write:

$$P(n_a, n_d \mid p_a, c) = p_a Pois(n_a, cE_{da} + (1-c)E_{aa}) \cdot Pois(n_d, cE_{dd} + (1-c)E_{ad}) + p_d Pois(n_a, E_{da}) \cdot Pois(n_d, E_{dd})$$

The resulting likelihood surface is shown in Figure S4. In this case the inferred contamination level is 4% with a confidence interval of 1% to 12%. Since our coverage on the X chromosome is half of that on the autosomes, and we have fewer sites on the X than on autosomes, the confidence interval is larger.

NSS and evolutionary conservation scores

Evolutionary conservation at nucleotide positions reflects the effect of past purifying selection, and therefore the biological constraint on a site. We evaluate the evolutionary conservation of the different groups of non-synonymous substitutions using as a measure the substitution deficit, as reported by the genomic evolutionary rate profiling (GERP) (S*24*). The substitution deficit is measure as "rejected substitutions", which quantify the constraint at individual positions. A rejected substitution is the number of substitutions that

could have occurred at a site but were "rejected" by selection. The conservation scores we used were calculated using the 28-way vertebrate alignment form the UCSC (S*25*).

NSS and overlap with published genome wide scans for positive selection

To test for overlap, we used three different genome wide scans for positive selection that were analyzed in a recent meta-analysis (S*26*). We looked for overlap between these regions on several classes of NSS learned in this study. The first class contains NSS where the Neandertal carries the ancestral state and that are fixed derived in humans (88). The second class has cases where the Neandertal was observed to be derived (10,015). We used the Fisher exact test to ask if one of these two categories were enriched for positions overlapping regions previously proposed to be positively selected in humans. We found no enrichment for any of the three genome wide scans (Table S4).

GO analysis

We asked if the 83 genes, where the Neandertal carries the ancestral state and humans are fixed derived, cluster in a functional category according to the Gene Ontology (S*27*). We used a test based on the hypergeometric distribution implemented in the software FUNC (*28*). In two tests we used as control groups all genes present on the array, and all genes carrying fixed substitutions in humans where Neandertal carries the derived allele. The only category that was found to approach significance (p-values: 0.05 and 0.04, respectively) after correction for multiple testing was potassium ion binding (GO:0030955). The four from the 83 genes in this category were *KCNS1*, *KCNV2*, *SLC12A1* and *KCNH8*.

Supplementary references


S1. N. Rohland, M. Hofreiter, Comparison and optimization of ancient DNA extraction. *Biotechniques*. **42**, 343-352 (2007).

S2. A. Rosas *et al.*, Paleobiology and comparative morphology of a late Neandertal sample from El Sidron, Asturias, Spain. *Proc Natl Acad Sci U S A*. **103**, 19266-19271 (2006).

S3. M. Hofreiter, V. Jaenicke, D. Serre, A. Haeseler Av, S. Paabo, DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res*. **29**, 4793-4799 (2001).

S4. A. W. Briggs *et al.*, Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A*. **104**, 14616-14621 (2007).

S5. A. W. Briggs *et al.*, Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res*. (2009), doi:2010.1093/nar/gkp1163.

S6. A. W. Briggs *et al.*, Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*. **325**, 318-321 (2009).

S7. E. W. Sayers *et al.*, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. **38**, D5-16 (2010).

S8. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32**, 1792-1797 (2004).

S9. W. R. Pearson, Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol*. **132**, 185-219 (2000).

S10. E. Hodges *et al.*, Genome-wide in situ exon capture for selective resequencing. *Nat Genet*. **39**, 1522-1527 (2007).

S11. A. Gnirke *et al.*, Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*. **27**, 182-189 (2009).

S12. S. B. Ng *et al.*, Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. **461**, 272-276 (2009).

S13. E. Hodges *et al.*, Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc*. **4**, 960-974 (2009).

S14. J. Krause *et al.*, A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. *Curr Biol*. **20**, 231-236 (2010).

S15. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*. (2010), doi:10.1101/pdb.prot5448.

S16. M. Kircher, U. Stenzel, J. Kelso, Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol*. **10**, R83 (2009).

S17. R. E. Green *et al.*, A complete neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*. **134**, 416-426 (2008).

S18. W. J. Kent, BLAT--the BLAST-like alignment tool. *Genome Res*. **12**, 656-664 (2002).

S19. K. Prufer *et al.*, PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*. **24**, 1530-1531 (2008).

S20. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**, 1754-1760 (2009).

S21. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078-2079 (2009).
S22. H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. **18**, 1851-1858 (2008).
S23. J. Krause *et al.*, The derived FOXP2 variant of modern humans was shared with Neandertals. *Curr Biol*. **17**, 1908-1912 (2007).
S24. G. M. Cooper *et al.*, Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. **15**, 901-913 (2005).
S25. W. Miller *et al.*, 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res*. **17**, 1797-1808 (2007).
S26. J. M. Akey, Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res*. **19**, 711-722 (2009).
S27. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. **25**, 25-29 (2000).
S28. K. Prufer *et al.*, FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics*. **8**, 41 (2007).

Supplementary figures

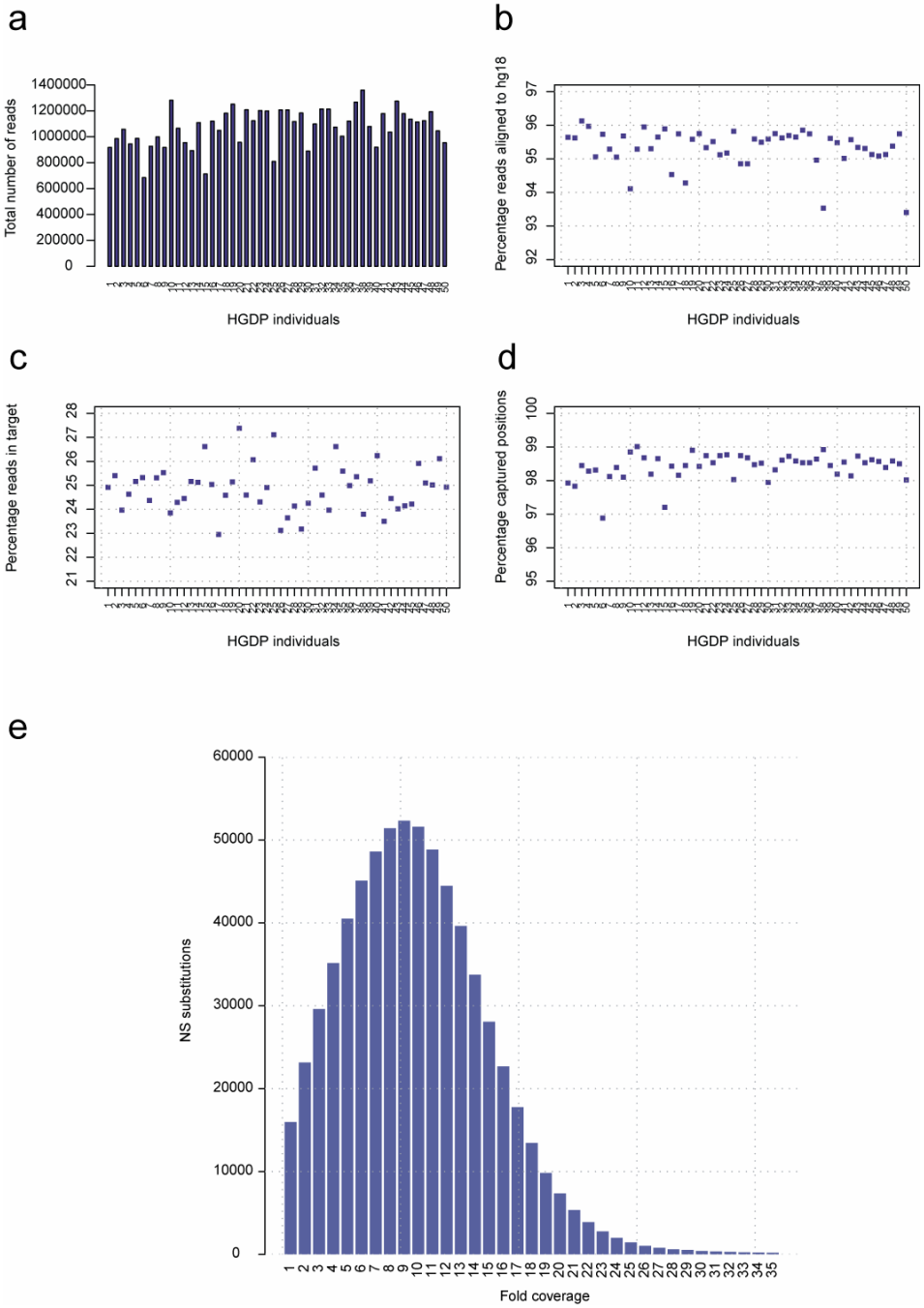


Figure S1. DNA capture of 50 HGDP individuals. (a) Total number of reads for each of the 50 HGDP individuals (b) Percentage of reads aligned to hg18 for each of the 50 HGDP individuals (c) Percentage of reads in target for each of the 50 HGDP individuals (d) Percentage of NSS captured for each of the 50 HGDP individuals. (e) Distribution average coverage across all HGDP individuals.

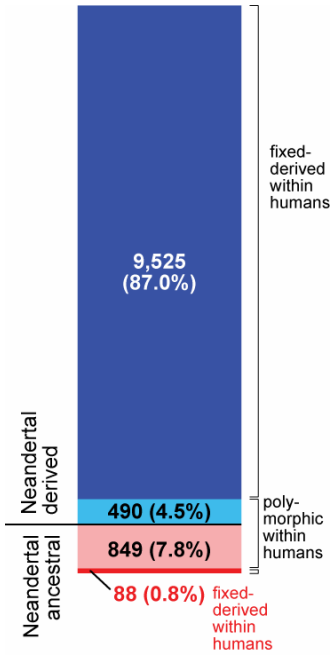


Figure S2. Classification of NSS using Neandertal sequence. In dark blue are the fixed NSS where Neandertal has the derived allele (matching the reference human). These substitutions are inferred to have occurred in hominin evolution prior to the Neandertal/human split. As expected, this accounts for the majority of hominin NSS. In dark red are the fixed NSS where the Neandertal hast the ancestral/chimpanzee base. These fixations are inferred to have occurred since the Neandertal/human split. In light blue and light red are the numbers of NSS where present-day humans within the HGDP panel are observed to be polymorphic. For the majority of positions where Neandertal is ancestral, present-day humans are polymorphic.

Because of our conservative definition of ancestral and derived in Neandertal, our rate of miscalling because of sequencing error and contamination is expected to be very low – less than 1% in sites observed just once, and lower in sites observed multiple times. Thus, of the 9,525 sites classified as "human fixed derived, Neandertal derived", a contamination rate of 3% might have caused us to classify on average 23 sites as derived even though we only ever observed human molecules – of these 23 on average 22 would belong the 724 sites that were observed with only one read (of these 23, around 20 would have still been correctly classified because Neandertal is derived in >90% of sites at which human is fixed).

Of the sites what were left unclassified, because both ancestral and derived reads were observed, some will be truly heterozygous in Neandertal, and some will have been misclassified because of sequencing error or because a molecule originating from a present-day human was observed. 236 sites are unclassified, most of which are probably derived in Neandertal.

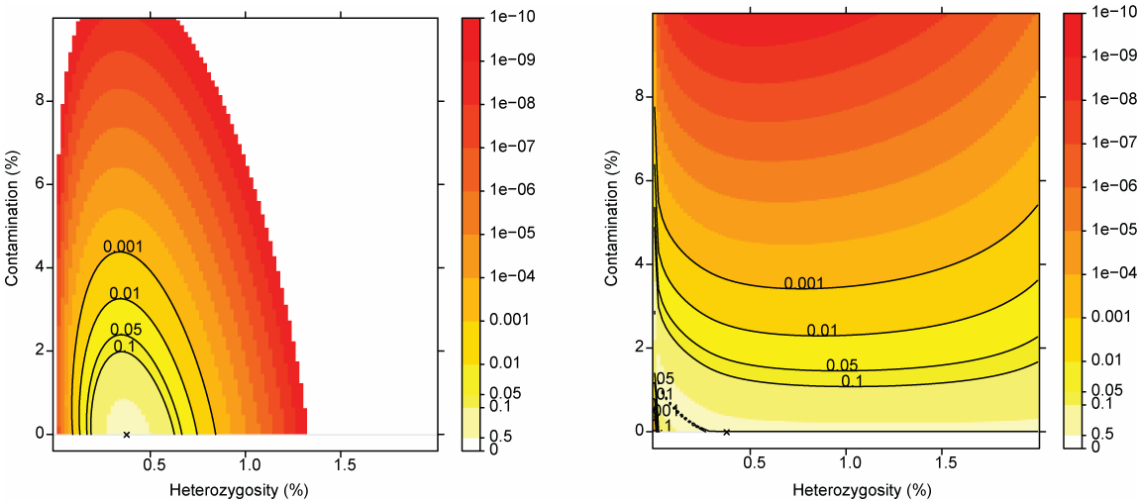


Figure S3. Left: likelihood surface for contamination by heterozygosity at human fixed derived sites. Likelihood ratio computed vs. the maximum likelihood, with colors corresponding to rejection cutoffs usint the $\chi^2$ distribution. Confidence intervals plotted using $\chi^2$ distribution with two degrees of freedom. Right: constrained likelihood surface, where heterozygosity is held constant - the x axis represents this constant. Likelihood ratios were computed vs. the maximum likelihood for each heterozygosity level. Confidence intervals plotted using the $\chi^2$ distribution with one degree of freedom.

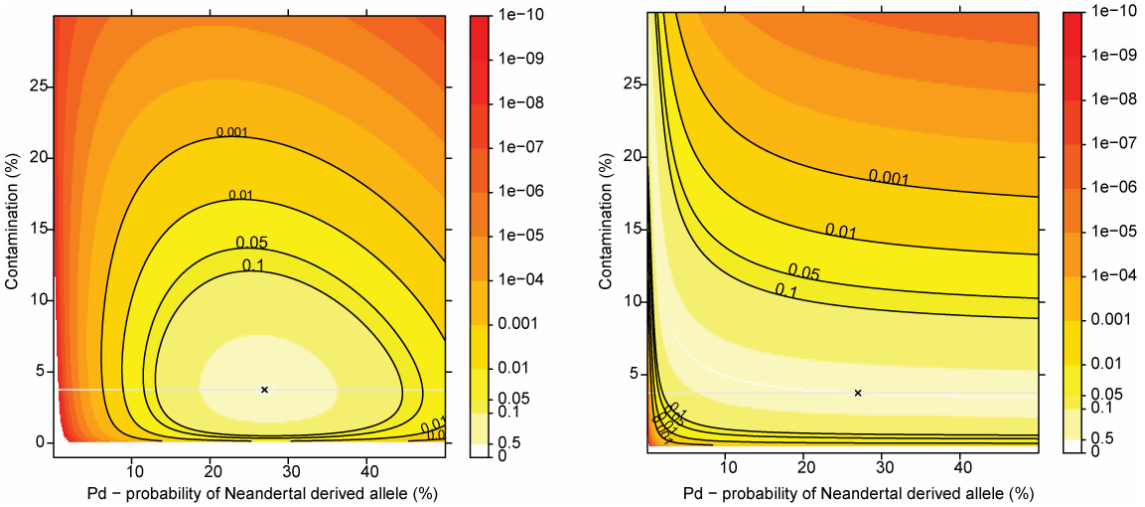


Figure S4. Left: likelihood surface base on positions on the X chromosome for contamination by fraction of Neandertal ancestral at human fixed derived sites. Likelihood ratio computed vs. the maximum likelihood, with colors corresponding to rejection cutoffs using the $\chi^2$ distribution. Confidence intervals plotted using $\chi^2$ distribution with two degrees of freedom. Right: constrained likelihood surface, where $P_d$, the probability for Neandertal derived is held constant. Likelihood ratios were compared vs. the maximum likelihood for each $P_d$. Confidence intervals plotted using the $\chi^2$ distribution with one degree of freedom.

Supplementary tables

Table S1. Human Genome Diversity Cell Line Panel individuals from whom genomic libraries were captured and sequenced.

| HGDP-CEPH ID | Sex | Population | Geographic origin | Continent |
|---|---|---|---|---|
| HGDP00980 | Female | Biaka Pygmies | Central African Republic | Africa |
| HGDP00471 | Female | Mbuti Pygmies | Democratic Republic of Congo | Africa |
| HGDP00909 | Female | Mandenka | Senegal | Africa |
| HGDP00920 | Female | Yoruba | Nigeria | Africa |
| HGDP01414 | Female | Bantu N.E. | Kenya | Africa |
| HGDP01254 | Female | Mozabite | Algeria (Mzab) | Africa |
| HGDP00607 | Female | Bedouin | Israel (Negev) | Asia |
| HGDP00557 | Female | Druze | Israel (Carmel) | Asia |
| HGDP00679 | Female | Palestinian | Israel (Central) | Asia |
| HGDP00151 | Female | Makrani | Pakistan | Asia |
| HGDP00192 | Female | Sindhi | Pakistan | Asia |
| HGDP00232 | Female | Pathan | Pakistan | Asia |
| HGDP00286 | Female | Kalash | Pakistan | Asia |
| HGDP00336 | Female | Burusho | Pakistan | Asia |
| HGDP00783 | Female | Han | China | Asia |
| HGDP01098 | Female | Tujia | China | Asia |
| HGDP01188 | Female | Yizu | China | Asia |
| HGDP01196 | Female | Miaozu | China | Asia |
| HGDP01209 | Female | Oroqen | China | Asia |
| HGDP01215 | Female | Daur | China | Asia |
| HGDP01223 | Female | Mongola | China | Asia |
| HGDP01234 | Female | Hezhen | China | Asia |
| HGDP01251 | Female | Xibo | China | Asia |
| HGDP01305 | Female | Uygur | China | Asia |
| HGDP01314 | Female | Dai | China | Asia |
| HGDP01323 | Female | Lahu | China | Asia |
| HGDP01334 | Female | She | China | Asia |
| HGDP01345 | Female | Naxi | China | Asia |
| HGDP01354 | Female | Tu | China | Asia |
| HGDP00955 | Female | Yakut | Siberia | Asia |
| HGDP00754 | Female | Japanese | Japan | Asia |
| HGDP00712 | Female | Cambodian | Cambodia | Asia |
| HGDP00544 | Female | Papuan | New Guinea | Oceania |
| HGDP00656 | Female | NAN Melanesian | Bougainville | Oceania |
| HGDP00513 | Female | French | France | Europe |
| HGDP01363 | Female | French Basque | France | Europe |
| HGDP00667 | Female | Sardinian | Italy | Europe |
| HGDP01156 | Female | North Italian | Italy (Bergamo) | Europe |
| HGDP01168 | Female | Tuscan | Italy | Europe |
| HGDP00794 | Female | Orcadian | Orkney Islands | Europe |
| HGDP01381 | Female | Adygei | Russia Caucasus | Europe |
| HGDP00881 | Female | Russian | Russia | Europe |

| | | | | |
|---|---|---|---|---|
| HGDP01038 | Female | Pima | Mexico | America |
| HGDP00854 | Female | Maya | Mexico | America |
| HGDP00702 | Female | Colombian | Colombia | America |
| HGDP00995 | Female | Karitiana | Brazil | America |
| HGDP00830 | Female | Surui | Brazil | America |
| HGDP00987 | Male | San | Namibia | Africa |
| HGDP00015 | Male | Brahui | Pakistan | Asia |
| HGDP00096 | Male | Balochi | Pakistan | Asia |
| HGDP00111 | Male | Hazara | Pakistan | Asia |

Table S2. Recently fixed NSS

| RefSeq ID | Gene Symbol | Gene name | Molecular function | Chr | Position | HSA AA | PTR AA |
|---|---|---|---|---|---|---|---|
| NM_033226 | ABCC12 | ATP-binding cassette, sub-family C (CFTR/MRP), member 12 | ATP-binding cassette (ABC) transporter | chr16 | 46679412 | I | T |
| NM_014237 | ADAM18 | ADAM metallopeptidase domain 18 | Metalloprotease | chr8 | 39683508 | R | K |
| NM_018963 | BRWD1 | bromodomain and WD repeat domain containing 1 | - | chr21 | 39494173 | V | M |
| NM_152440 | C12orf66 | hypothetical protein FLJ32549 | - | chr12 | 62873950 | V | L |
| NM_175741 | C15orf55 | nuclear protein in testis | - | chr15 | 32434060 | V | I |
| NM_025108 | C16orf59 | chromosome 16 open reading frame 59 | - | chr16 | 2450677 | G | E |
| NM_198532 | C19orf35 | chromosome 19 open reading frame 35 | Serine/threonine protein kinase receptor;Protein kinase | chr19 | 2229847 | P | L |
| NM_138358 | C19orf52 | chromosome 19 open reading frame 52 | - | chr19 | 10901126 | W | R |
| NM_023080 | C8orf33 | chromosome 8 open reading frame 33 | - | chr8 | 146248841 | G | R |
| NM_033138 | CALD1 | caldesmon 1 | Non-motor actin binding protein | chr7 | 134293530 | I | V |
| NM_144508 | CASC5 | cancer susceptibility candidate 5 | - | chr15 | 38700151 | R | H |
| NM_144508 | CASC5 | cancer susceptibility candidate 5 | - | chr15 | 38702931 | S | G |
| NM_001256 | CDC27 | cell division cycle 27 homolog (S. cerevisiae) | Other miscellaneous function protein | chr17 | 42587077 | P | S |
| NM_152562 | CDCA2 | cell division cycle associated 2 | - | chr8 | 25420017 | R | P |
| NM_001902 | CTH | cystathionase (cystathionine gamma-lyase) | Other lyase | chr1 | 70662610 | G | A |
| NM_182699 | DDX53 | DEAD (Asp-Glu-Ala-Asp) box polypeptide 53 | RNA helicase | chrX | 22928705 | T | M |
| NM_001930 | DHPS | deoxyhypusine synthase | Synthase | chr19 | 12651505 | E | D |
| NM_012242 | DKK1 | dickkopf homolog 1 (Xenopus laevis) | - | chr10 | 53744245 | M | L |
| NM_031304 | DOHH | deoxyhypusine hydroxylase/monooxygenase | - | chr19 | 3447595 | M | V |
| NM_001025248 | DUT | dUTP pyrophosphatase | Other phosphatase;Other hydrolase | chr15 | 46411265 | S | C |
| NM_152512 | ENTHD1 | ENTH domain containing 1 | Other membrane traffic protein | chr22 | 38491517 | R | T |
| NM_153332 | ERI1 | three prime histone mRNA exonuclease 1 | Exoribonuclease;Esterase | chr8 | 8906538 | F | C |
| NM_001441 | FAAH | fatty acid amide hydrolase | Other hydrolase | chr1 | 46650471 | A | G |
| NM_152698 | FAM123C | hypothetical protein FLJ38377 | - | chr2 | 131237414 | L | V |
| NM_018121 | FAM178A | chromosome 10 open reading frame 6 | - | chr10 | 102666423 | E | K |
| NM_005267 | GJA8 | gap junction protein, alpha 8, 50kDa (connexin 50) | Gap junction | chr1 | 145847862 | P | L |
| NM_153368 | GJD4 | connexin40.1 | - | chr10 | 35937468 | P | H |
| NM_003801 | GPAA1 | glycosylphosphatidylinositol anchor attachment protein 1 homolog (yeast) | - | chr8 | 145211312 | E | Q |
| NM_014668 | GREB1 | GREB1 protein | - | chr2 | 11675941 | R | C |
| NM_172002 | HSCB | HscB iron-sulfur cluster co-chaperone homolog (E. coli) | - | chr22 | 27471916 | I | M |
| NM_007044 | KATNA1 | katanin p60 (ATPase-containing) subunit A 1 | Non-motor microtubule binding protein;Other hydrolase | chr6 | 149960458 | T | A |
| NM_144633 | KCNH8 | potassium voltage-gated channel, subfamily H (eag-related), member 8 | Voltage-gated potassium channel | chr3 | 19550047 | T | M |
| NM_002251 | KCNS1 | potassium voltage-gated channel, delayed-rectifier, subfamily S, member 1 | - | chr20 | 43156982 | Q | R |
| NM_133497 | KCNV2 | potassium channel, subfamily V, member 2 | Voltage-gated potassium channel | chr9 | 2719704 | L | P |
| NM_018689 | KIAA1199 | KIAA1199 | - | chr15 | 78960362 | A | T |
| NM_020890 | KIAA1524 | KIAA1524 | - | chr3 | 109769738 | T | I |
| NM_052904 | KLHL32 | KIAA1900 | - | chr6 | 97639853 | E | D |
| NM_002286 | LAG3 | lymphocyte-activation gene 3 | Interleukin receptor | chr12 | 6754050 | A | T |

| RefSeq ID | Gene Symbol | Gene name | Molecular function | Chr | Position | HSA AA | PTR AA |
|---|---|---|---|---|---|---|---|
| NM_017980 | LIMS2 | LIM and senescent cell antigen-like domains 2 | Actin binding cytoskeletal protein;Structural protein | chr2 | 128113345 | T | A |
| NM_001039029 | LRTM2 | leucine-rich repeats and transmembrane domains 2 | Receptor;Extracellular matrix | chr12 | 1807600 | G | E |
| NM_000081 | LYST | lysosomal trafficking regulator | Select regulatory molecule | chr1 | 234036008 | S | N |
| NM_001040179 | MCHR2 | melanin-concentrating hormone receptor 2 | G-protein coupled receptor | chr6 | 100475588 | A | V |
| NM_020203 | MEPE | matrix, extracellular phosphoglycoprotein with ASARM motif (bone) | - | chr4 | 88986389 | N | T |
| NM_053050 | MRPL53 | mitochondrial ribosomal protein L53 | - | chr2 | 74553091 | T | A |
| NM_000662 | NAT1 | N-acetyltransferase 1 (arylamine N-acetyltransferase) | Acetyltransferase | chr8 | 18124280 | V | I |
| NM_000662 | NAT1 | N-acetyltransferase 1 (arylamine N-acetyltransferase) | Acetyltransferase | chr8 | 18124475 | S | A |
| NM_014071 | NCOA6 | nuclear receptor coactivator 6 | Other nucleic acid binding | chr20 | 32801189 | I | M |
| NM_024782 | NHEJ1 | nonhomologous end-joining factor 1 | - | chr2 | 219651057 | A | T |
| NM_033004 | NLRP1 | NLR family, pyrin domain containing 1 | - | chr17 | 5403917 | F | L |
| NM_017852 | NLRP2 | NLR family, pyrin domain containing 2 | - | chr19 | 60181000 | Q | L |
| NM_006489 | NOVA1 | neuro-oncological ventral antigen 1 | mRNA splicing factor | chr14 | 25987939 | V | I |
| NM_033334 | NR6A1 | nuclear receptor subfamily 6, group A, member 1 | Nuclear hormone receptor;Transcription factor;Nucleic acid binding | chr9 | 126340293 | Q | P |
| NM_022072 | NSUN3 | NOL1/NOP2/Sun domain family, member 3 | Nucleic acid binding;Methyltransferase | chr3 | 95285750 | S | F |
| NM_198474 | OLFML1 | olfactomedin-like 1 | Receptor | chr11 | 7463757 | T | A |
| NM_001005517 | OR5K4 | olfactory receptor, family 5, subfamily K, member 4 | - | chr3 | 99555735 | T | A |
| NM_001005517 | OR5K4 | olfactory receptor, family 5, subfamily K, member 4 | - | chr3 | 99556164 | G | R |
| NM_024791 | PDZD3 | PDZ domain containing 3 | - | chr11 | 118564408 | G | A |
| NM_138694 | PKHD1 | polycystic kidney and hepatic disease 1 (autosomal recessive) | - | chr6 | 51599824 | I | N |
| NM_000936 | PNLIP | pancreatic lipase | Lipase | chr10 | 118311044 | M | K |
| NM_199437 | PRDM10 | PR domain containing 10 | Zinc finger transcription factor;Other DNA-binding protein | chr11 | 129277502 | N | T |
| NM_144707 | PROM2 | prominin 2 | Membrane traffic protein | chr2 | 95309418 | D | E |
| NM_002830 | PTPN4 | protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte) | - | chr2 | 120451104 | T | S |
| NM_002841 | PTPRG | protein tyrosine phosphatase, receptor type, G | Other receptor;Protein phosphatase | chr3 | 62253890 | I | V |
| NM_080860 | RSPH1 | testis specific A2 homolog (mouse) | - | chr21 | 42770559 | Q | K |
| NM_001037540 | SCML1 | sex comb on midleg-like 1 (Drosophila) | Transcription factor;Chromatin/chromatin-binding protein | chrX | 17678235 | T | M |
| NM_030971 | SFXN3 | sideroflexin 3 | Cation transporter;Other transfer/carrier protein | chr10 | 102789253 | N | K |
| NM_022071 | SH2D4A | SH2 domain containing 4A | - | chr8 | 19266005 | E | K |
| NM_001037633 | SIL1 | SIL1 homolog, endoplasmic reticulum chaperone (S. cerevisiae) | Select regulatory molecule;Miscellaneous function | chr5 | 138484627 | Q | R |
| NM_000338 | SLC12A1 | solute carrier family 12 (sodium/potassium/chloride transporters), member 1 | Cation transporter;Other transporter | chr15 | 46287278 | S | N |
| NM_021097 | SLC8A1 | solute carrier family 8 (sodium/calcium exchanger), member 1 | Cation transporter;Other transporter | chr2 | 40510859 | T | I |
| NM_052910 | SLITRK1 | SLIT and NTRK-like family, member 1 | Receptor;Extracellular matrix | chr13 | 83352655 | S | A |
| NM_206996 | SPAG17 | sperm associated antigen 17 | - | chr1 | 118360154 | T | A |
| NM_206996 | SPAG17 | sperm associated antigen 17 | - | chr1 | 118435819 | Y | D |
| NM_003126 | SPTA1 | spectrin, alpha, erythrocytic 1 (elliptocytosis 2) | - | chr1 | 156914833 | N | D |
| NM_015136 | STAB1 | stabilin 1 | Extracellular matrix structural protein | chr3 | 52510805 | H | Y |
| NM_012449 | STEAP1 | six transmembrane epithelial antigen of the prostate 1 | - | chr7 | 89631970 | C | S |
| NM_014979 | SV2C | synaptic vesicle glycoprotein 2C | Other transfer/carrier protein | chr5 | 75627399 | P | H |
| NM_014258 | SYCP2 | synaptonemal complex protein 2 | - | chr20 | 57885976 | M | T |
| NM_001009991 | SYTL3 | synaptotagmin-like 3 | Membrane traffic regulatory protein | chr6 | 159004394 | T | I |

| RefSeq ID | Gene Symbol | Gene name | Molecular function | Chr | Position | HSA AA | PTR AA |
|---|---|---|---|---|---|---|---|
| NM_001010870 | TDRD6 | tudor domain containing 6 | Nuclease | chr6 | 46767869 | V | A |
| NM_018469 | TEX2 | testis expressed sequence 2 | - | chr17 | 59644188 | G | D |
| NM_005656 | TMPRSS2 | transmembrane protease, serine 2 | Serine protease | chr21 | 41788292 | V | A |
| NM_207377 | TOMM20L | TIMM9 | Other transporter | chr14 | 57932515 | N | D |
| NM_130466 | UBE3B | ubiquitin protein ligase E3B | Ubiquitin-protein ligase | chr12 | 108421901 | T | M |
| NM_025090 | USP36 | ubiquitin specific peptidase 36 | Cysteine protease | chr17 | 74311068 | R | W |
| NM_014594 | ZNF354C | zinc finger protein 354C | KRAB box transcription factor | chr5 | 178439192 | T | K |
| NM_014205 | ZNHIT2 | zinc finger, HIT type 2 | Zinc finger transcription factor | chr11 | 64641288 | Q | R |
| NM_014205 | ZNHIT2 | zinc finger, HIT type 2 | Zinc finger transcription factor | chr11 | 64641532 | C | R |

Table S3. Genes with 2 recently fixed NSS

| Gene symbol | Gene full name |
|---|---|
| SPAG17 | sperm associated antigen 17 |
| CASC5 | cancer susceptibility candidate 5 |
| ZNHIT2 | zinc finger, HIT type 2 |
| OR5K4 | olfactory receptor, family 5, subfamily K, member 4 |
| NAT1 | arylamine N-acetyltransferase |

Table S4. Overlap NSS and genome wide scans for positive selection

| Study | Data | Statistical Method | Sample | Neandertal Ancestral – Human fixed derived | Neandertal Ancestral | p-value |
|---|---|---|---|---|---|---|
| Carlson et al. (2005) | SNP | Site frequency spectrum | Perlegen | 1/88 | 21/10015 | 0.17 |
| Kelly et al. (2006) | SNP | Site frequency spectrum | Perlegen | 3/88 | 133/10015 | 0.12 |
| Williamson et al. (2007) | SNP | Site frequency spectrum | Perlegen | 1/88 | 166/10015 | 1 |

Table S5. Number of sites used in the contamination estimate. Each entry in the table shows the number of sites for which a certain number of sequences with the ancestral allele (column) and a certain number of sequences with the derived allele (row) were observed. Thus, the entry "2" in the row before last says that for two sites we observed 17 sequences that had the derived allele, and no sequences that had the ancestral allele. The table only lists sites for which humans were classified as fixed derived, since contamination is easiest to classify when Neandertal is homozygous ancestral, and humans are fixed derived. The fact that the first row in the table, which represents sites from which no derived allele was seen, is much larger than the second row, with sites for which one derived allele was seen, shows that the contamination rate is very low. Most of the entries that are in the middle of the table, i.e. not its first row or column, are probably the result of sequencing error. Only a small fraction seems to be heterozygosity in Neandertal. Thus, positions where both derived and ancestral alleles were observed were excluded from the analysis. The likelihood model asks what is the chance to observe a certain number of ancestral and a certain number of derived reads, given that we know the heterozygosity rate, the error rate and the contamination rate.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 248 | 17 | 12 | 18 | 11 | 10 | 5 | 5 | 2 | 3 | | 1 |
| 1 | 643 | 13 | 4 | | | 1 | | 1 | | | | |
| 2 | 1172 | 20 | 1 | 1 | | | | | | | | |
| 3 | 1400 | 21 | 1 | | | | | | | | | |
| 4 | 1453 | 22 | 4 | 2 | | 1 | | | | | | |
| 5 | 1320 | 34 | 4 | | | | | | | | | |
| 6 | 994 | 27 | 7 | | | | | | | | | |
| 7 | 731 | 22 | 4 | 2 | | | | | | | | |
| 8 | 478 | 16 | 3 | | | | | | | | | |
| 9 | 278 | 11 | 1 | | | | | | | | | |
| 10 | 143 | 4 | | | | | | | | | | |
| 11 | 82 | 4 | | | | | | | | | | |
| 12 | 40 | 3 | | | | | | | | | | |
| 13 | 18 | | | | | | | | | | | |
| 14 | 10 | 1 | | | | | | | | | | |
| 15 | 8 | 2 | | | | | | | | | | |
| 16 | 2 | | | | | | | | | | | |
| 17 | 2 | | | | | | | | | | | |
| 18 | 1 | | | | | | | | | | | |