On Characterizing Adaptive Events Unique to Modern Humans

Jessica L. Crisci
University of Massachusetts Medical School

Repository Citation
Crisci JL, Wong A, Good JM, Jensen JD. (2011). On Characterizing Adaptive Events Unique to Modern Humans. Morningside Graduate School of Biomedical Sciences Student Publications. https://doi.org/10.1093/gbe/evr075. Retrieved from https://escholarship.umassmed.edu/gsbs_sp/1749
Title

On Characterizing Adaptive Events Unique to Modern Humans

Authors

Jessica L. Crisci1*, Alex Wong2, Jeffrey M. Good3, Jeffrey D. Jensen1§

Author Affiliations

1Program in Bioinformatics and Integrative Biology, University of Massachusetts, Medical School, Worcester, Massachusetts, USA
2Department of Biology, Carleton University, Ottawa, Ontario, Canada
3Division of Biological Sciences, University of Montana, Missoula, Montana, USA
§ Current address: École Polytechnique Fédérale de Lausanne (EPFL), School of Life Sciences
*Author for Correspondence: Jessica Crisci, Program in Bioinformatics and Integrative Biology, University of Massachusetts, Medical School, Worcester, Massachusetts, USA, (Tel) 508-856-6983, (Fax) 508-856-0017, (email) jessica.crisci@umassmed.edu
Abstract

Ever since the first draft of the human genome was completed in 2001 there has been increased interest in identifying genetic changes that are uniquely human, which could account for our distinct morphological and cognitive capabilities with respect to other apes. Recently, draft sequences of two extinct hominin genomes, a Neanderthal and Denisovan, have been released. These two genomes provide a much greater resolution to identify human-specific genetic differences than the chimpanzee, our closest extant relative. The Neanderthal genome paper presented a list of regions putatively targeted by positive selection around the time of the human-Neanderthal split. We here seek to characterize the evolutionary history of these candidate regions - examining evidence for selective sweeps in modern human populations, as well as for accelerated adaptive evolution across apes. Results indicate that 3 of the top 20 candidate regions show evidence of selection in at least one modern human population ($p < 5 \times 10^{-5}$). Additionally, 4 genes within the top 20 regions show accelerated amino acid substitutions across multiple apes ($p < 0.01$), suggesting importance across deeper evolutionary time. These results highlight the importance of evaluating evolutionary processes across both recent and ancient evolutionary timescales, and intriguingly suggest a list of candidate genes that may have been uniquely important around the time of the human-Neanderthal split.

**Keywords**: adaptation, Neanderthal genomics, selective sweeps
Background

The identification of genomic regions that have been affected by positive selection in humans, but not in other primates, is a promising avenue for characterizing the genetic changes underlying phenotypic traits that are unique to humans. With the advent of whole-genome sequencing technology, a number of primate genomes have recently become available for such comparisons (e.g., chimpanzee, The Chimpanzee Sequencing and Analysis Consortium 2005; macaque, Rhesus Macaque Genome Sequencing and Analysis Consortium 2007; orangutan, Locke et al. 2011; and gorilla, Wellcome Trust Sanger Institute). Additionally, two extinct hominin genomes have recently been sequenced: the Neanderthal (Green et al. 2010) and a newly discovered archaic hominin from Denisova Cave in Siberia (Reich et al. 2010). Genomic information from these extinct hominin individuals provides a unique opportunity to identify genetic changes that occurred in the evolution of modern humans (see Figure 1).

Green et al. produced a list of putatively swept regions in humans by aligning the human, chimpanzee, and Neanderthal genomes. They looked for spans of the genome with sites polymorphic in five modern human populations, where Neanderthal carried the ancestral allele with respect to chimpanzee. The expected number of Neanderthal derived alleles was calculated and compared to the observed number - producing a measure, $S$, which was used to quantify the absence of Neanderthal derived sites within a given region (with larger $S$ corresponding to a higher confidence of a human-specific selective sweep). Because the expected number of Neanderthal derived alleles is conditioned on the genomic average of each configuration of observed human alleles at polymorphic sites, this approach has unique power to detect older selective sweeps along the human branch. Importantly, this allows detection at time scales for which standard frequency spectrum based tests lack power (Green et al. 2010, SOM).
Additionally, because the window size of variation affected by a sweep is related to $s/r$ (the strength of selection over the recombination rate; Kaplan et al. 1989) and the transition time for a beneficial mutation is $-\log(1/2N_e)/s$ generations, they were most likely to find regions that had been affected by strong selection (i.e., having fixed since the human-Neanderthal split, $\sim s > 0.001$).

In contrast, traditional genomic scans for positive selection rely on the hitchhiking pattern evident in linked neutral variation (Maynard Smith and Haigh 1974), and are limited to detecting adaptive fixations having occurred within $\sim 0.2 N_e$ generations (Kim and Stephan 2002).

Divergence-based methods, on the other hand, rely not on patterns in polymorphism but rather on detecting increased rates of amino acid substitution between lineages, and thus are appropriate to study recurrent selection across multiple species (i.e., on a much longer evolutionary time scale) - requiring multiple beneficial fixations in order to have power.

(Figure 1 around here)

Thus, the Green et al. approach is unique in that the timescale over which it may identify positive selection is in between purely divergence- or polymorphism-based approaches (Figure 1), and they provide a first glance at regions that may set humans apart from our closest evolutionary relatives. Using this method, they identified a total of 212 genomic regions, representing the top 5% of loci with signals of putative sweeps dating around the human-Neanderthal split.

As indicated by Figure 1, these candidate adaptive regions may be further characterized into four general categories of positive selection. They may be: 1) accelerated across apes, 2) accelerated in modern humans, 3) accelerated in the common ancestor of humans and Neanderthals, or 4) uniquely important around the time of the human-Neanderthal split. Our
objective is to characterize these regions across both broad and narrow evolutionary time in order
to reveal which regions may in fact have been uniquely important around the human-Neanderthal
split, and to discover the extent of overlap between their method and traditional site frequency
spectrum and \( d_N/d_S \) methods for detecting positive selection. We ask the question: given a list of
regions that in theory represent ancient sweeps along the human lineage, how many could have
been detected without the use of the Neanderthal genome?

In order to distinguish among the possible alternatives we utilize two additional classes of
methodology: 1) the codeml sites model and branch model (Yang 1998, Yang et al. 2000) from
the software package PAML, which identifies genes that show accelerated amino acid
substitution across multiple species (Yang 2007), and within a single branch, respectively, using
measures of \( d_N/d_S \) and 2) SweepFinder (Nielsen et al. 2005), which identifies genetic regions that
show evidence of a recent beneficial fixation within a single population using polymorphism
data. This direction is similar in principle to the recent work of Cai et al. (2009) who
demonstrated a relationship between high \( d_N \) and levels of polymorphism, which they interpret as
evidence of recurrent positive selection. While we are similarly comparing across multiple time-
scales, our starting dataset is composed of those genes recently suggested to be important around
the human-Neanderthal split (\( i.e., \) as opposed to high \( d_N \) across the tree), and thus results are not
directly comparable.

Our findings indicate that many of these regions would not have been detected as
candidates for positive selection using traditional frequency spectrum or divergence-based
approaches, and that the Neanderthal genome has indeed allowed for the identification of regions
experiencing positive selection over a unique time period of the human lineage. By focusing
exclusively on the putatively selected regions of the Green et al. study, we additionally parse this
gene set in to those most likely to have been important in differentiating human and Neanderthal.

Evidence for selection across apes

A common approach for detecting positive selection across multiple species is to
close the ratio of the rate of non-synonymous substitutions (mutations that lead to amino acid
changes; \(d_n\)) to the rate of synonymous substitutions (silent mutations; \(d_s\)), with \(d_n/d_s = 1\), <1 and
>1 being consistent with neutral, purifying and positive selection, respectively. In early
applications, \(d_n/d_s\) was averaged over all sites within a protein sequence and across the entire
evolutionary time scale of all lineages. This application has little power to detect positive
selection, because it is likely that most sites are functionally constrained (\(d_n/d_s << 1\)) and are
primarily shaped by purifying selection. For our analysis we utilize codeml, which has a sites
model allowing \(d_n/d_s\) (\(\omega\)) to vary at each site along a sequence (Yang et al. 2000). This method is
still conservative in that it averages \(d_n\) and \(d_s\) over lineages at each site, but it has improved
power to detect site specific positive selection in a functional protein sequence (Wong et al.
2004).

Tests of positive selection in the codeml sites model compare the fit of the data under a
neutral model, to that under a model of positive selection via a likelihood ratio test. For the
following analysis, three model comparisons were considered: M1a vs. M2a, M7 vs. M8, and
M8a vs. M8. M1a has two subsets of sites, one where \(\omega\) varies between 0 and 1 and one where \(\omega\)
is fixed at one; in M2a \(\omega\) can be less than 1, equal to 1, or greater than 1 (Wong et al. 2004). M7
assumes a beta-distribution for \(\omega\) between 0 and 1, and M8 adds an additional class of sites to
M7 with \(\omega > 1\) (Wong et al. 2004). In M8a this additional class is fixed at \(\omega = 1\) (Swanson et al.
Thus, M2a and M8 allow selection in each comparison, while M2, M7, and M8a fit the data to a neutral model. A maximum likelihood ratio is computed for each model, and the null and selection models are compared via a likelihood ratio comparison.

For our analysis we focused on the top 20 largest putative sweep regions from Green et al. (2010) and the 51 genes contained within them (Table 1). Orthologues were obtained in five primate species: macaque, chimpanzee, orangutan, human, and gorilla. Of the original 51 genes, 8 were noncoding RNA (MIR genes and MEG3), and thus not suitable for codeml analysis. Of the remaining 43 genes, 29 had annotated 1:1 orthologues in the above primate species in Ensembl. We did not use genes from species with more than one annotated orthologue. Multiple species alignments were constructed using the PRANK alignment algorithm (Loytynoja and Goldman 2005) and tested using the three codeml model comparisons described above. Results are summarized in Table 2. Two of the 29 genes showed significant positive selection under all three comparisons: CCDC82 and RFX5. Additionally, CGN showed significant positive selection under M1a vs. M2a and M8a vs. M8, and THADA was significant under M8a vs. M8. We have included this last gene in further discussions since this model comparison is the most realistic (Swanson et al. 2003).

Two of these genes are involved in human disease/immunity. THADA, which has been shown to be involved in beta-cell function (Simonis-Bik et al. 2010), is located close to a potential susceptibility locus of type II diabetes (Zeggini et al. 2008), and a SNP within THADA has been shown to be associated with type II diabetes (Schleinitz et al. 2010). RFX5 is involved in MHC-II expression through interferon gamma (Xu et al. 2003; Garvie et al. 2008). Genes involved in immunity are among the most highly represented in scans for positive selection.
(Yang 2005), with several studies finding significant evidence for positive selection within the antigen recognition site of MHC-I (Hughes and Nei 1988; Yang et al. 2002) and MHC-II (Hughes and Nei 1989). The other two genes, *CCDC82*, and *CGN*, are not as well characterized, and any inference about their evolutionary significance would be purely speculative.

(The Table 2 around here)

The codeml sites model also makes predictions regarding the most likely sites experiencing positive selection according to a Bayes empirical Bayes method (Yang et al. 2005). For each codon in a DNA sequence that is analyzed, the probability that $\omega > 1$ at that particular site is computed. A probability of greater than 0.95 was used to determine a site that showed significant positive selection. Of the four genes that were significant for at least two tests of selection under the sites model, two such sites were identified in *CCDC82*, *CGN*, and *THADA*; four sites were identified in *RFX5* (Figure 2). In all cases, sites display accelerated rates of evolution across the species tree, but do not contain human-specific changes.

(Figure 2 around here)

Additionally, we performed two branch tests in codeml, which specifically test for higher than expected $d_s/d_\omega$ along a single branch of interest. For this analysis we tested the human branch and the branch ancestral to humans, Neanderthals, and Denisovans. This is achieved, again, by a likelihood ratio comparison between two models where a $d_s/d_\omega$ ratio is assigned to each branch in the tree. Each of the models allows for two values for $d_s/d_\omega$: one for the foreground branch where positive selection is assumed ($\omega_1$), and one for the rest of the background branches ($\omega_0$). In the null model, $\omega_1$ is fixed to equal 1 on the foreground branch, while $\omega_0$ is estimated on the remaining branches. In the alternative model, $\omega_1$ is also estimated from the data.
We found that none of the previous 29 species alignments showed significant positive selection along either the human branch or the branch ancestral to hominins \((p < 0.01)\). However, five genes did reject the null model in favor of the alternative on both branches \((p < 0.01:\) CADPS2, DYRK1A, BACH2, INPPL1, and ZFP36L2\) though \(\omega_1 < 1\).

**Evidence for selection in modern human populations**

To detect recent selective sweeps in human populations we used ascertainment-corrected polymorphism data from Perlegen, in African-American, European-American, and Chinese populations (Williamson et al. 2007). The program SweepFinder (Nielsen et al. 2005) was used to scan for sweeps, given the relatively large size of the genomic regions under consideration. SweepFinder computes the background site frequency spectrum (SFS) for the region in question and then identifies unusual regions relative to this background (Figure 3). A significant cutoff value is determined using neutral simulation (see Methods).

Of the top 20 putative sweep regions from Green et al., three were identified as being consistent with recent selection in modern humans (Figure 3). Sweep region 1 is upstream of ZFP36L2 on chromosome 2 in the European population (Figure 3a). Sweep region 2 is centered around an intron of KCNAB1 on chromosome 3 in the African population (Figure 3b). Finally, sweep region 3 is localized near the last exon of DLK1 on chromosome 14 in the Chinese population (Figure 3c). These sweeps are distinct from those detected in the original dataset for at least two reasons. First, our sweep analysis was performed using population specific data, and thus any selective signal will be unique to a single population; whereas the Green et al. scan was based upon detecting a joint signal from all five populations considered. Second, because of the time restrictions over which a recent sweep can be detected (~100,000 years for Africans), the
time scales of the two statistics are essentially non-overlapping. This scaling becomes even faster for populations of smaller effective population sizes (i.e., $N_e$(Chinese) = 510, $N_e$(Europe) = 1000) (Gutenkunst et al. 2009); thus the time to the oldest detectable sweep ~ 5100 years and ~10,000 years for the Chinese and European populations, respectively. Therefore, these results suggest recurrent selective sweeps along the human-lineage in these regions (i.e., around the human-Neanderthal split, and in modern human populations).

(Figure 3 around here)

In an attempt to localize potential genetic targets of these peak regions, the UCSC genome browser (track SNP 130) and dbSNP were used to identify SNPs specific to the populations under consideration. Since the peak regions in chromosome 2 and 14 were less than 1Kb, an additional 2Kb of human sequence was examined on either side of the peak. One high frequency derived SNP (rs10132598) was identified in the Asian population near the significant peak of chromosome 14 (CHB+JPT= 0.83, YRI= 0.30, and CEU=0.15) according to the 1000 genomes pilot data, phase 1 (Durbin et al. 2010). This agrees well with the SweepFinder result, as the significant peak using the Perlegen dataset was specific to the Chinese population.

Another SNP (rs72875566) was found near the significant peak region of chromosome 2. The significant sweep was detected in the European population, and interestingly, this SNP is at a higher frequency in individuals of European ancestry compared to Yorubans (0.85 vs. 0.61, respectively) according to the phase 1 low coverage data from the 1000 genomes project. No information on this SNP was provided for the Asian populations. This SNP is also located in a CpG island upstream of both ZFP36L2 and another predicted mRNA locus (LOC100129726, Figure 3) that was not in the original table in Green et al. These two genes transcribe in opposite
directions and the CpG island overlaps both genes, suggesting that it may affect expression of
either locus.

Discussion

By examining the candidate selection genes of Green et al. using both divergence and
polymorphism data, we have parsed the list of candidate regions that may have been uniquely
important in differentiating human and Neanderthal, providing an ideal list for functional
validation. The extent of overlap between codeml, SweepFinder, and Green et al. is summarized
in Table 1. Of the 20 original regions, 15 would not have been identified using the methods
tested above (Table 1, red text). This highlights the utility of the Neanderthal genome -
demonstrating power to identify regions that would have been missed by using site frequency
spectrum- or $d_s/d_3$-based methodology alone.

The genetic functions contained within some of these novel regions are of interest in
terms of human evolution. The $HoxD$ gene cluster located on chromosome 2 is involved in both
vertebral and limb development (for review, see Favier and Dollé 1997). Another interesting
genre is $RUNX2$ ($CBFA1$). This is a transcription factor involved in bone development. Mutations
in $RUNX2$ can lead to a skeletal disorder known as Cleidocranial dysplasia (CCD), which is
characterized by short stature, underdeveloped or missing clavicles, and dental and cranial
abnormalities, among other skeletal changes (Mundlos et al. 1997). Thus selection within these
regions could have led to morphological differences in modern humans.

Also of note are $DYRK1A$, $NRG3$, and $CADPS2$. $DYRK1A$ is located in the Down
Syndrome Critical Region on chromosome 21. It is expressed during brain development, and
also in the adult brain, where it is believed to be involved in learning and memory (Hämmerle et
al. 2003). $NRG3$ also has neurological implications. In humans, it is expressed in the
hippocampus, amygdala, and thalamus, and is believed to be a susceptibility locus for schizophrenia (Zhang et al. 1997, Wang et al. 2008). Mutations in CADPS2 have been associated with autism (Sadaka et al. 2010). Selection in these three regions during human evolution could have resulted in characteristic cognitive behavior.

The availability of extinct hominin genomic sequence, such as Neanderthal and Denisova, is an important milestone in the study of human evolution. These genomes provide much greater resolution for the identification of unique human adaptive substitutions, since they serve as a nearer outgroup than chimpanzee (Figure 1). Any human substitutions identified using chimpanzee may be shared among the many ancestors between human and chimpanzee, including Australopithecus and Paranthropus, whereas Neanderthal and Denisova are the two nearest known relatives of Homo sapiens. These two genomes also can provide a more detailed adaptive history of the human species, and, in combination with the selective scan method of Green et al., we now have power to detect adaptive fixations in deeper evolutionary time. Our results show that this method can, in fact, detect adaptive genomic regions that would have been missed using selective scans based on $d_s/d_S$ (i.e., codeml) or site frequency spectrum summary statistics (i.e., SweepFinder). In their analysis, Green et al. compared their regions to two other genomic scans for selection in humans, one using an outlier approach and the other based on Tajima’s $D$ (Tajima 1989). They found no significant overlap between their regions and those of other studies, further suggesting power over separate time frames. There is also no overlap between the 20 regions we examined here and the SweepFinder scan performed by Williamson et al. (2007).

It is not unexpected that the majority of genes we examined within these 20 candidate regions do not contain significant $d_s/d_S$. The codeml sites model requires that there be excessive
\( d_n \) across all species at a particular site in order to infer positive selection, and the human branch is short relative to other apes. Thus, the non-synonymous changes are more likely to pre-date humans. Additionally, the codeml branch model averages \( d_n/d_s \) across an entire sequence, and this leads to reduced power to detect selection, as discussed above. Moreover, Green et al. identified 78 fixed non-synonymous amino acid changes in humans that were ancestral in Neanderthal, and none of the genes containing these fixed changes overlapped with the genes in the top 20 candidate regions for a selective sweep. It may well be that the target of these sweeps was not non-synonymous (e.g., a synonymous or non-coding change, or that a non-synonymous change in humans was unable to be determined due to the variable depth in sequence coverage of the Neanderthal genome). In fact, 5 of the 20 candidate regions contain no annotated coding sequence (Table 1), and Green et al. found an additional 232 human-specific substitutions in 5’- and 3’- UTR regions, suggesting that non-coding sites may have been targeted.

**Conclusion**

Here we have shown that using an ancient hominin genomic sequence to scan for positive selection in humans (as performed by Green et al.) has elucidated a novel list of candidate selection regions that would not have been discovered using currently available methods of detecting selection. Of the 15 novel regions from the Green et al. scan, 5 contained genes with interesting relations to human morphological and cognitive traits. Therefore, we conclude that using an ancient hominin genome to scan for selection in conjunction with already established methods could offer a more complete picture of how positive selection has shaped modern humans.

**Methods**

*Multiple species alignment for codeml*
Human mRNA sequences were obtained from Ensembl. Only sequences with CCDS citations were used. If there was more than one transcript, the one with the longest amino acid sequence was chosen. Macaque, chimp, gorilla, and orangutan sequences were retrieved from Ensembl using BioMart. Briefly, using the list of human gene IDs, orthologous Ensembl gene IDs for each species were obtained from the Ensembl Genes 58 human dataset using the homologs filter under Multi-species Comparisons. These IDs were then queried to get orthologous coding transcript sequences from each species using the sequences attribute. In cases where more than one transcript variant was returned, the longest was chosen. Only genes showing 1:1 homology with orthologues in all five species were used for codeml analysis.

Sequences were aligned using PRANK (Loytynoja and Goldman 2005). The codon option was used, which uses the empirical codon model (ECM; Kosiol et al. 2007) to align individual codons while preserving the reading frame. The guide tree was estimated by the program and all other parameters were left as default. This method of alignment was shown by Fletcher and Yang (2010) to be the most accurate at preserving true sequence alignment in the presence of insertions and deletion when using the PAML branch-site test.

**codeml analysis**

The codeml program in PAML version 4.4 (Yang 2007) was used to test for positive selection across apes (with the exception of macaque, which was included even though it is an Old World Monkey). Three different sites model tests were examined: M1a vs. M2a, M7 vs. M8, and M8 vs. M8a (see PAML documentation for parameters). A likelihood ratio test was used to determine significance. A Bonferroni corrected \( p \) value assuming 29 tests (0.05/29) is equal to 0.0018. We also compare with the uncorrected \( p \) value of 0.01 to determine significance. For
both the sites and human-specific branch tests, an alignment of 5 primate species is used (human, chimpanzee, gorilla, orangutan, macaque). For the human-Neanderthal ancestral branch test an alignment of 7 species was used that included the Neanderthal and Denisovan sequences. These two sequences were excluded from sites test due to the variable coverage of both genomes, as codeml ignores sites with missing data.

Neanderthal and Denisova sequence construction

The BAM files for Neanderthal and Denisova can be found at:
ftp://ftp.ebi.ac.uk/pub/databases/ensembl/neandertal and
http://hgdownload.cse.ucsc.edu/downloads.html, respectively. SAMtools (Li et al. 2009) was used to retrieve the reads corresponding to each gene sequence from the Neanderthal and Denisova BAM files using the chromosomal locations. These reads were mapped back to hg18 using Geneious version 5.3.2 (Drummond et al. 2011). A Phred scaled confidence score cutoff of 30 was applied for all sites where these sequences differed from hg18.

SweepFinder Analysis

The data used for this analysis was the same Perlegen SNP dataset as in Williamson et al 2007. The SNPs for each region were analyzed using SweepFinder (Nielsen et al. 2005), which computes the background site frequency spectrum (SFS) for a region using SNP data. It uses a likelihood framework (Kim and Stephan 2002) to compare the background SFS with that expected under a model of a selective sweep at a predetermined set of sites along the region. The number of sites is designated by the gridsize parameter, and was set to the number of nucleotides in the region. The cut off value was determined by simulating 1000 replicates in ms (Hudson
2002) under the standard neutral model for each region. The parameters for each simulated
region consisted of the same SNP density (by setting the “S” parameter in ms equal to the
number of SNPs from the Perlegen dataset present in the region) and gridsize as the actual
region. For ms style input, SweepFinder returns the maximum LR value for each replicate. To
determine significance, the top 99.995% of LR values ($p = 5 \times 10^{-5}$) were considered significant.
This $p$ value reflects a Bonferroni correction for 1000 tests.

Acknowledgements

The authors would like to thank Udo Stenzel, Melissa Hubisz, and Ed Green for their help in
obtaining and utilizing datasets, and Svante Pääbo for helpful comments and discussion.
This work was funded by grants from the National Science Foundation [DEB-1002785] and
Worcester Foundation to JDJ.

References

Cai, J.J., Macpherson, J.M., Sella, G., and Petrov, D. a. 2009. Pervasive hitchhiking at coding
and regulatory sites in humans. PLoS Genet 5: e1000336.
Drummond AJ, et al. 2011. Geneious v5.4, Available from http://www.geneious.com/
Durbin, R.M. et al. 2010. A map of human genome variation from population-scale sequencing.
Nature 467: 1061-1073.
Favier, B., and Dolle, P. 1997. Developmental functions of mammalian Hox genes. Mol Hum
Reprod 3: 115-131.
Fletcher, W., and Yang, Z. 2010. The Effect of Insertions, Deletions and Alignment Errors on the
Branch-Site Test of Positive Selection. Mol Biol Evol 27: 2257-2267.
Garvie, C.W., and Boss, J.M. 2008. Assembly of the RFX complex on the MHCII promoter: role of RFXAP and RFXB in relieving autoinhibition of RFX5. Biochim Biophys Acta 1779: 797-804.

Green, R.E. et al. 2010. A draft sequence of the Neandertal genome. Science 328: 710-722.

Gutenkunst, R., Hernandez, R.D., Williamson, S., and Bustamante, C. D. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5: e1000695.

Hämmerle B, Elizalde C, Galceran J, Becker W, Tejedor FJ. 2003. The MNB/DYRK1A protein kinase: neurobiological functions and Down syndrome implications. J Neural Transm Suppl (67):129-37.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Hughes, A.L., and Nei, M. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. P Natl Acad of Sci USA 86: 958-962.

Hughes, A.L., and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167-170.

Kaplan, N.L., Hudson, R R, and Langley, C.H. 1989. The "hitchhiking effect" revisited. Genetics 123: 887-899.

Kim, Y.H., and Stephan, W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765-777.

Kosiol, C., Holmes, I., and Goldman, N. 2007. An empirical codon model for protein sequence evolution. Mol Biol and Evol 24: 1464-79.
Li, H., et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-79.

Locke, D.P. et al. 2011. Comparative and demographic analysis of orang-utan genomes. Nature 469: 529-533.

Löytynoja, A., and Goldman, N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. P Natl Acad Sci USA 102: 10557-62.

Maynard Smith, J., and Haigh, J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23: 23-35.

Mundlos, S. et al. 1997. Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia. Cell 89: 773-779.

Nielsen, R. et al. 2005. Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566-75.

Reich, D. et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053-60.

Sadakata, T., and Furuichi, T. 2010. Ca(2+)-dependent activator protein for secretion 2 and autistic-like phenotypes. Neurosci Res 67: 197-202.

Schleinitz, D. et al. 2010. Lack of significant effects of the type 2 diabetes susceptibility loci JAZF1, CDC123/CAMK1D, NOTCH2, ADAMTS9, THADA, and TSPAN8/LGR5 on diabetes and quantitative metabolic traits. Horm Metab Res 42: 14-22.
Simonis-Bik, A.M. et al. 2010. Gene variants in the novel type 2 diabetes loci CDC123/CAMK1D, THADA, ADAMTS9, BCL11A, and MTNR1B affect different aspects of pancreatic beta-cell function. Diabetes 59: 293-301.

Swanson, W.J., Nielsen, R., and Yang, Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20: 18-20.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

The Chimpanzee Sequencing And Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.

Wang, Y.-C., Chen, J.-Y., Chen, M.-L., Chen, C.-H., Lai, I.-C., Chen, T.-T., Hong, C.-J., Tsai, S.-J., and Liou, Y.-J. 2008. Neuregulin 3 genetic variations and susceptibility to schizophrenia in a Chinese population. Biol Psychiat 64: 1093-6.

Williamson, S.H. et al. 2007. Localizing recent adaptive evolution in the human genome. PLoS Genet 3: e90.

Wong, W.S.W., Yang, Z., Goldman, N., and Nielsen, R. 2004. Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. Genetics 168: 1041-51.

Xu, Y., Wang, L., Buttice, G., Sengupta, P.K., and Smith, B.D. 2003. Interferon gamma repression of collagen (COL1A2) transcription is mediated by the RFX5 complex. J Biol Chem 278: 49134-44.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution 15: 568-573.
Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-91.

Yang, Z. 2005. The power of phylogenetic comparison in revealing protein function. P Natl Acad Sci USA 102: 3179-80.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M.K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449.

Yang, Z., Wong, W.S.W., and Nielsen, R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107-18.

Zeggini, Eleftheria et al. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40: 638-645.

Zhang, D. et al. 1997. Neuregulin-3 (NRG3): a novel neural tissue-enriched protein that binds and activates ErbB4. P Natl Acad Sci USA 94: 9562-7.
Figure Legends

Figure 1. Summary of methods. A graphical representation of the evolutionary timescale over which the methods for detecting positive selection are effective. Branch lengths are not drawn to scale. Divergence-based methods can detect positive selection across a phylogenetic tree or along a single branch; polymorphism-based methods are effective within a single population; the Green et al. method using the Neanderthal genome finds selection in humans that occurred shortly after the human-Neanderthal split.

Figure 2. Mutations at significant sites across the primate tree. For genes that showed significant positive selection by at least two tests in the codeml sites model, the nucleotide changes within the candidate sites for selection were mapped. In cases where there were two possible scenarios that could describe how a change originated, the simplest was assumed. Branch lengths are not drawn to scale, and the spacing and ordering of the mapped substitutions on a given branch are arbitrary.

Figure 3. Sweep regions. The three regions identified from the Green et al. dataset as showing evidence of a selective sweep in a modern human population using SweepFinder. The horizontal dashed line represents a Bonferroni corrected LR cutoff ($p < 5 \times 10^{-5}$). Approximate region lengths correspond to the significant portion of the peak. Population-specific high frequency derived SNPs are marked with an arrow along the x-axis. a) A region of 531bp upstream of ZFP36L2 and LOC100129726 in the European-American population. b) A region of ~11Kb within an intron of KCNAB1 in the African-American population. c) A region of 121bp within an intron of
DLK1 in the Chinese population. For these plots, the coordinates for chromosomal location along the x-axis correspond to the hg16 genome annotation.
Figure 1.
Figure 2.
Figure 3.
| Region (hg18)    | Width (CM) | Genes                                                                 |
|-----------------|------------|----------------------------------------------------------------------|
| chr2:43265008-43601389 | 0.5726     | **ZFP36L2**, **THADA**, **LOC100129726**<sup>a</sup>                  |
| chr11:95533088-95867597 | 0.5538     | **JRKL**, **CCDC82**, **MAML2**                                      |
| chr10:62343313-62655667 | 0.5167     | **RHOBTB1**                                                          |
| chr21:37580123-37789088 | 0.4977     | **DYSK1A**                                                           |
| chr10:83336607-83714543 | 0.4654     | **NRG3**                                                             |
| chr14:100248177-100417724 | 0.4533     | **MIR337**, **MIR665**, **DLK1**, **RTL1**, **MIR431**, **MIR493**, **MEG3**, **MIR770** |
| chr3:157244328-157597592 | 0.425      | **KCNA1**                                                            |
| chr11:30601000-30992792 | 0.3951     |                                                                     |
| chr2:176635412-176978762 | 0.3481     | **HOXD11**, **HOXD8**, **EVX2**, **MTX2**, **HOXD1**, **HOXD10**, **HOXD13**, **HOXD4**, **HOXD12**, **HOXD9**, **MIR108**, **HOXD3** |
| chr11:71572763-71914957 | 0.3402     | **CLPB**, **FOLR1**, **PHOX2A**, **FOLR2**, **INPPL1**               |
| chr7:41537742-41838097 | 0.3129     | **INHBA**                                                            |
| chr10:60015775-60262822 | 0.3129     | **BICC1**                                                            |
| chr6:45440283-45705503 | 0.3112     | **RUNX2**, **SUPT3H**                                                |
| chr1:149553200-149878507 | 0.3047     | **SELENBP1**, **POGZ**, **MIR554**, **RFXS**, **SNX27**, **CGN**, **TUFT1**, **PI4KB**, **PSMB4** |
| chr7:121763417-122282663 | 0.2855     | **RNF148**, **RNF133**, **CADPS2**                                   |
| chr7:93597127-93823574 | 0.2769     |                                                                     |
| chr16:62369107-62675247 | 0.2728     |                                                                     |
| chr14:48931401-49095338 | 0.2582     |                                                                     |
| chr6:90762790-90903925 | 0.2502     | **BACH2**                                                            |
| chr10:9650088-9786954  | 0.2475     |                                                                     |

<sup>a</sup>The significant results using each method are either colored green (overlap between Green et al. and SweepFinder) or blue (overlap between Green et al. and codeml). Regions colored in red contain no overlap with the tested methods, and represent a novel list of genes unique to the Green et al. scan using Neanderthal.

<sup>b</sup>For codeml, genes that were significant for at least two tests of selection are underlined (p<0.01).

<sup>c</sup>LOC100129726 was not listed in Green et al. Table 3.
Table 2. Summary of codeml results

| Genes    | $2\Delta\ln L_{\text{M8a-M2a}}$ | $2\Delta\ln L_{\text{M7-M8}}$ | $2\Delta\ln L_{\text{M8a-M8}}$ | $\omega(\Pr(\omega > 1))$ | $p_{\text{sites}}$ | $\omega > 1$ |
|----------|-------------------------------|-------------------------------|-------------------------------|---------------------------|---------------------|-------------|
| BACH2    | 0.00                          | 0.00                          | 0.00                          |                           | 0.130               | 0.00        |
| BICC1    | 4.78                          | 5.31                          | 4.78                          |                           |                     |             |
| CADPS2   | 2.28                          | 2.50                          | 2.28                          |                           |                     |             |
| CCDC82   | 8.34**                        | 13.03**                       | 8.34**                        | 5.781/0.980               | 0.130               | 0.00        |
| CGN      | 6.55**                        | 6.55                          | 6.55**                        | 4.186/0.952               | 0.376               |             |
| CLPB     | 0.50                          | 0.61                          | 0.50                          |                           |                     |             |
| DLK1     | 1.85                          | 2.57                          | 1.83                          |                           |                     |             |
| DRYK1A   | 0.00                          | -0.18                         | -0.18                         |                           |                     |             |
| EVX2     | 2.19                          | 2.51                          | 2.17                          |                           |                     |             |
| FOLR1    | 0.00                          | 0.00                          | 0.00                          |                           |                     |             |
| HOXD1    | 4.43                          | 4.81                          | 4.41                          |                           |                     |             |
| HOXD4    | 0.20                          | 0.47                          | 0.20                          |                           |                     |             |
| HOXD8    | 3.00                          | 3.00                          | 3.00                          |                           |                     |             |
| HOXD9    | 0.00                          | 0.00                          | 0.00                          |                           |                     |             |
| HOXD10   | 0.00                          | 0.00                          | 0.00                          |                           |                     |             |
| INHBA    | 0.00                          | -0.32                         | -0.32                         |                           |                     |             |
| INPPL1   | 1.77                          | 1.94                          | 1.76                          |                           |                     |             |
| KCNAB1   | 4.72                          | 8.63**                        | 4.29                          | 2.121/0.934               | 0.003               |             |
| MAML2    | 0.03                          | 0.13                          | 0.03                          |                           |                     |             |
| NRG3     | 0.00                          | 0.00                          | 0.00                          |                           |                     |             |
| PHOX2A   | 0.00                          | 0.00                          | 0.00                          |                           |                     |             |
| PK4B     | -6.26                         | -0.32                         | -1.98                         |                           |                     |             |
| PSMB4    | 0.63                          | 0.53                          | 0.51                          |                           |                     |             |
| RFX5     | 13.03**                       | 13.05**                       | 13.03**                       | 7.898/0.993               | 0.050               |             |
| SNX27    | 0.00                          | 0.00                          | 0.00                          |                           |                     |             |
| SUPT3H   | 1.70                          | 2.03                          | 1.70                          |                           |                     |             |
| TUBA1A   | 6.35**                        | 7.11                          | 6.35**                        | 3.720/0.965               | 0.108               |             |
| TUFT1    | 0.14                          | 0.19                          | 0.14                          |                           |                     |             |
| ZFP36L2  | 0.05                          | 0.41                          | 0.05                          |                           |                     |             |

** ** p<0.01

<sup>a</sup>Significance for each test was determined from a chi-squared distribution with degrees of freedom =1 for M8a vs. M8, and df=2 for M1a vs. M2a and M7 vs. M8.

<sup>b</sup>The probability that $\omega$ is greater than one at a given site in the sequence based on the BEB posterior probability For each gene showing evidence of positive selection. The highest probability observed is given with its corresponding $\omega$ value.

<sup>c</sup>The proportion of sites examined per sequence that fall in the category of $\omega$ being greater than one.