# Selective Sweeps across Twenty Millions Years of Primate Evolution

Kasper Munch,*[,1] Kiwoong Nam,[2] Mikkel Heide Schierup,[1] and Thomas Mailund[1]

[1]Bioinformatics Research Centre, Aarhus University, C. F. Møllers Alle, Denmark

[2]UMR INRA, Université Montpellier 2 Place Eugène Bataillon, France

***Corresponding author:** E-mail: kaspermunch@birc.au.dk.

**Associate editor:** Anna Di Rienzo

## Abstract

**The contribution from selective sweeps to variation in genetic diversity has proven notoriously difficult to assess, in part because polymorphism data only allows detection of sweeps in the most recent few hundred thousand years. Here, we show how linked selection in ancestral species can be quantified across evolutionary timescales by analyzing patterns of incomplete lineage sorting (ILS) along the genomes of closely related species. We show that sweeps in the human–chimpanzee and human–orangutan ancestors can be identified as depletions of ILS in regions in excess of 100 kb in length. Sweeps predicted in each ancestral species, as well as recurrent sweeps predicted in both species, often overlap sweeps predicted in humans. This suggests that many genomic regions experience recurrent selective sweeps. By comparing the ILS patterns along the genomes of the closely related human–chimpanzee and human–orangutan ancestors, we are further able to quantify the impact of selective sweeps relative to that of background selection. Compared with the human–orangutan ancestor, the human–chimpanzee ancestor shows a strong excess of regions depleted of ILS as well as a stronger reduction in ILS around genes. We conclude that sweeps play a strong role in reducing diversity along the genome and that sweeps have reduced diversity in the human–chimpanzee ancestor much more than in the human–orangutan ancestor.**

***Key words:*** **incomplete lineage sorting, selective sweeps, background selection, primate evolution.**

## Introduction

The genetic variation across the genome is shaped by mutation and natural selection. This means that the action of selection can be inferred from patterns of diversity across the genome if variation in mutation rate is controlled for. Positive and negative selection both reduce genetic variation in the targeted regions because the fitness effect associated with the selected variant increases the variance in the number of offspring among individuals. The magnitude of this reduction in diversity, and the size of the genomic region that is affected, depend on the type of selection and the local recombination rate. Positive selection in the form of hard selective sweeps are mainly identified as strong reductions in diversity in regions linked to selected variants, as well as from associated distortions to the site frequency spectrum. Negative selection removes variation linked to deleterious variants in a process referred to as background selection. Here each deleterious variant only has a small effect on linked diversity, but the cumulated effect may be strong in regions with many targets of purifying selection. It remains an open question whether positive or negative selection exerts the strongest influence on genetic diversity.

A recent study that contrasted diversity data across distantly related species, showed that the joint effect of background selection and selective sweeps increases with effective population size (Corbett-Detig et al. 2015) but the study was unable to quantify a separate effect of selective sweeps because distantly related species cannot be assumed to show similar levels of background selection. Another recent study attempted to disentangle selection effects by assuming that the effect of background selection scales with the density of sites under purifying selection whereas positive selection scales with the proportion of non-synonymous substitutions (Elyashiv et al. 2015). This study found that the combined effect of sweeps and background selection reduces genome-wide diversity by more than 50% compared with the diversity expected in the absence of selection. The study concluded that classic selective sweeps driven by selection on new protein coding variants play a relatively small role in Drosophila, but that the combined effect of other types of positive selection, such as selection on standing variation, may be similar to the effect of background selection. Two recent analyses that separately address the effect of background selection also report a significant role of sweeps in shaping human diversity (Enard et al. 2014; Fagny et al. 2014).

Here, we present a different approach that quantifies the effects of sweeps and background selection by measuring their effects on incomplete lineage sorting (ILS) patterns. This allows us to identify individual selective sweeps in ancestral species and, by contrasting ILS patterns among closely related species, we are able to disentangle the impact that positive and negative selection have on genomic diversity.







Article

Fast Track

ILS arises when the time from one speciation event to the next is not long enough for all lineages to have found common ancestry. This leads to gene trees with topologies that are different from that of the species tree, here termed non-canonical topologies (NCT). ILS is ubiquitous on the internal branches of the great apes species tree (Mailund et al. 2014), and the two species trios investigated here (human, chimpanzee, gorilla and human, orangutan, gibbon) both show high levels of ILS. The probability that a position in a genomic alignment of three closely related species shows a NCT can be expressed using standard coalescent theory (Takahata 1989) as $\exp[-(t_2 - t_1)/2N_e g] * 2/3$, where $g$ is generation time, $t_1$ and $t_2$ are the two speciation times, and $N_e$ is the local effective population size of the ancestor of the two most closely related species (marked with blue and green dots in fig. 1A). NCTs are called for each position of the genome by posterior decoding of a fitted coalescent hidden Markov model (Hobolth et al. 2007; Dutheil et al. 2009). The demographic model used is a three species isolation model with constant population sizes. Demographic model parameters include times of the two speciation events and population sizes of the two ancestral populations (see Methods). The model has four hidden states that represent alternative gene trees, two of which represent NCTs (red and orange in fig. 1B).

Since the speciation time is the same along the genome, the proportion of sites showing NCT in a genomic region reflects the average coalescence time and thus the diversity of the ancestral species. In contrast to diversity estimates using polymorphism data, the estimates using NCT proportions are not affected by variation in mutation rate across the genome. Background selection in the ancestral species will reduce $N_e$ locally and thus reduce the regional proportion of NCT. A selective sweep in the ancestral species will induce a rapid coalescence of linked sequence and thus eliminate the possibility of NCT in a region around the selected site. The proportion of NCT in a genomic region thus reflects the collective effect of background selection and selective sweeps across the entire span of time between the two speciation events. This unique property allows us to evaluate the effects of linked selection on a much longer time span than that typically addressed by polymorphism data of extant species. Further, since the contribution of background selection to variation in NCT proportions across the genome is expected to be highly similar between closely related species, we can attribute strong differences in NCT proportions to the action of selective sweeps.

In our analysis of ILS patterns, we apply two separate approaches to gauge the effects of linked selection. The first is a scan for possible selective sweeps in the human–chimpanzee and human–orangutan ancestors by identifying very long genomic regions depleted of NCT. The second approach contrasts the proportions of NCT across orthologous genomic regions of the two closely related (but independent) ancestral species and measures the separate effects of background selection and sweeps. The two ancestral species have independent diversity, yet are only separated by a few million years of evolution. Purifying selection is thus expected to affect the same genomic positions. Theoretically, background selection depends on the rate of deleterious mutations, the recombination rate, and the distribution of fitness effects of deleterious mutations and their dominance, all of which are expected to be similar among closely related species. Importantly, background selection is only slightly dependent on demographic differences and it is therefore expected to produce very similar local reductions in NCT along the genomic alignments of each species trio. This means that strong reductions in NCT proportion that are observed in only one species serve as strong indications of species-specific selective sweeps. It also means that we can quantify the impact of selective sweeps in each ancestral species from systematic differences in the levels of NCT along the genome in the two ancestors. Here we apply this idea to identify specific selective sweeps in the human–chimpanzee and human–orangutan ancestors and to gain new insight into the rate of selective sweeps across the last 20 million years of human evolution.

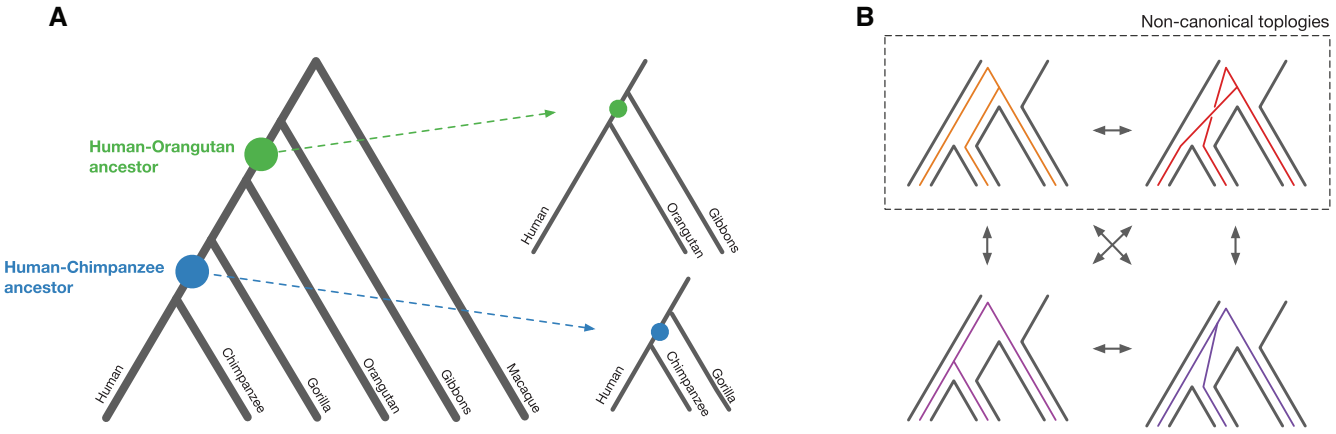


FIG. 1. (A) Cladogram of the great ape species tree is shown left. Each colored dot represents an ancestral species that is analyzed through comparative genomic analysis of a species trio shown to the right. (B) The four states of the coalescent hidden Markov model. Arrows show possible transitions. The two bottom states represent canonical topologies where the first coalescence event occurs in either the first or second ancestral species. The two top states represent non-canonical topologies (NCTs).

## Results

### Ancestral Population Sizes and Speciation Times On the Human Lineage

With the gibbon genome now available (Carbone et al. 2014), we were able to estimate speciation times and effective population sizes for two ancestral species on the great ape lineage leading to humans. In the human–orangutan ancestor, we estimate a mean proportion of NCT of 29%. This is almost identical to the mean proportion of NCT in the human–chimpanzee ancestor that we estimate to 30%, in line with previous estimates (Scally et al. 2012; Prado-Martinez et al. 2013). The almost identical proportions of NCT ensure that we have the same power to measure variation in diversity along the genomes of the two ancestral species. Assuming a per generation mutation rate of 6e−10 per year and a generation time of 20 years, we estimate the median speciation times in millions of years (with 0.25 and 0.75 quartiles) to 6.6 (6.2–7.1) for human–chimpanzee, 10.3 (9.9–10.9) for human–gorilla, 22.3 (21.2–24.0) for human–orangutan, and 26.6 (25.5–28.2) for human–gibbon (fig. 2A). Divergence between human and macaque is estimated to 54.6 (51.9–58.8). Using this divergence as calibration, our estimate of the human–gibbon speciation time is smaller than that obtained by Carbone et al. (2014). The speciation times for human–chimpanzee and human–gorilla were estimated using the same method and reference genomes as part of the initial gorilla genome analysis and these estimates fall within our confidence intervals (Scally et al. 2012). Our analysis is the first to exploit the large amount of ILS between orangutan and gibbon and we expect our estimate of the human–orangutan speciation to be more reliable than the 18 myr previously estimated based on only 1% ILS between human, chimpanzee, and orangutan (Hobolth et al. 2011).

The mean $N_e$ of the human–chimpanzee ancestor is estimated to 120,000 (96,000–150,000) and that of the human–orangutan ancestor to 140,000 (118,000–159,000). The smaller $N_e$ of the human–chimpanzee ancestor is associated with a shorter span of time between speciation events (3.7e6

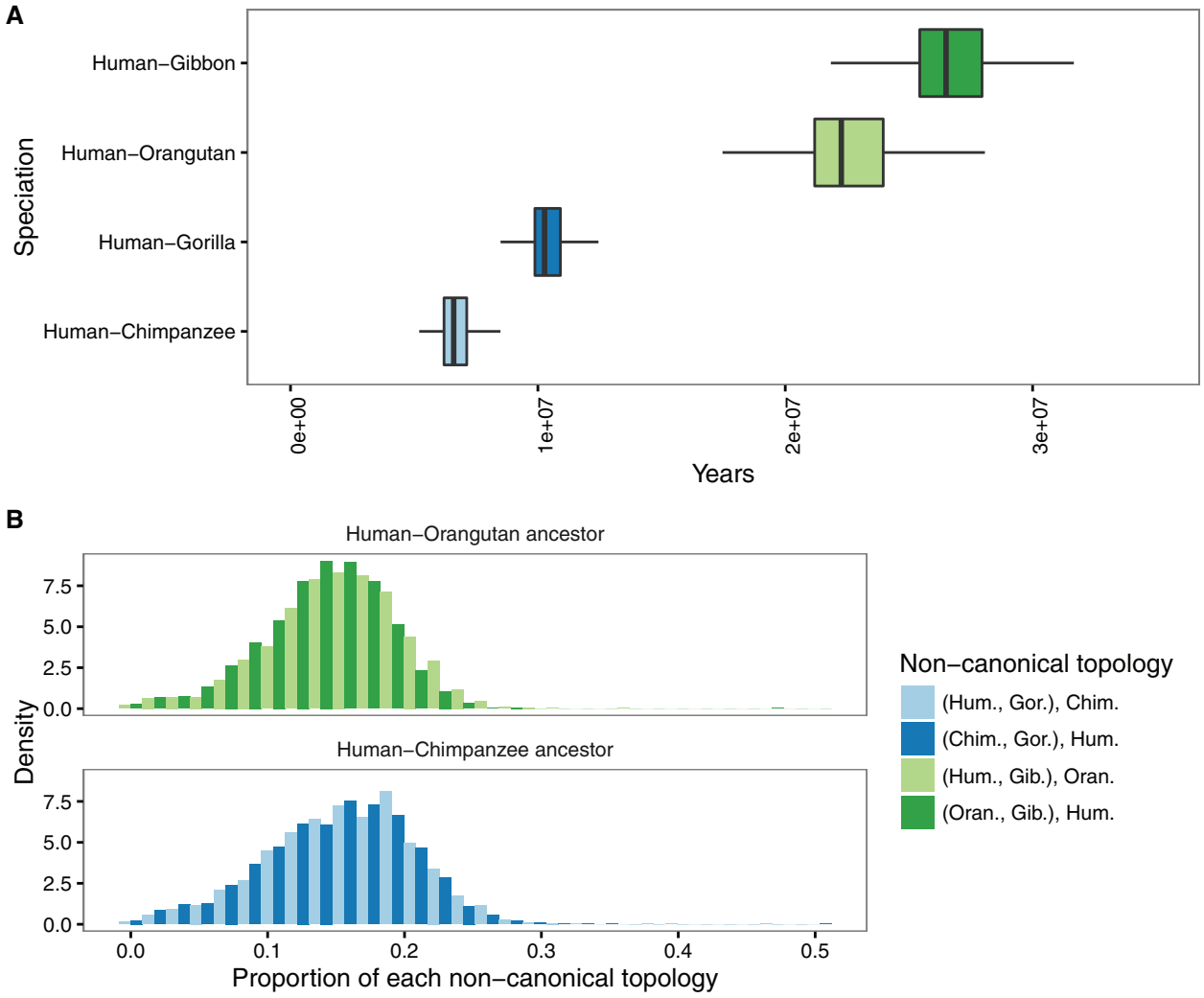


FIG. 2. (A) Distribution of estimated speciation times (outliers not shown) across individual coalescence HMM analyses of 1Mb of genomic alignment. (B) Distributions of the proportion of each non-canonical topology in 1 Mb windows for the species trio defining human–orangutan ancestor (top) and the human–chimpanzee ancestor (bottom).

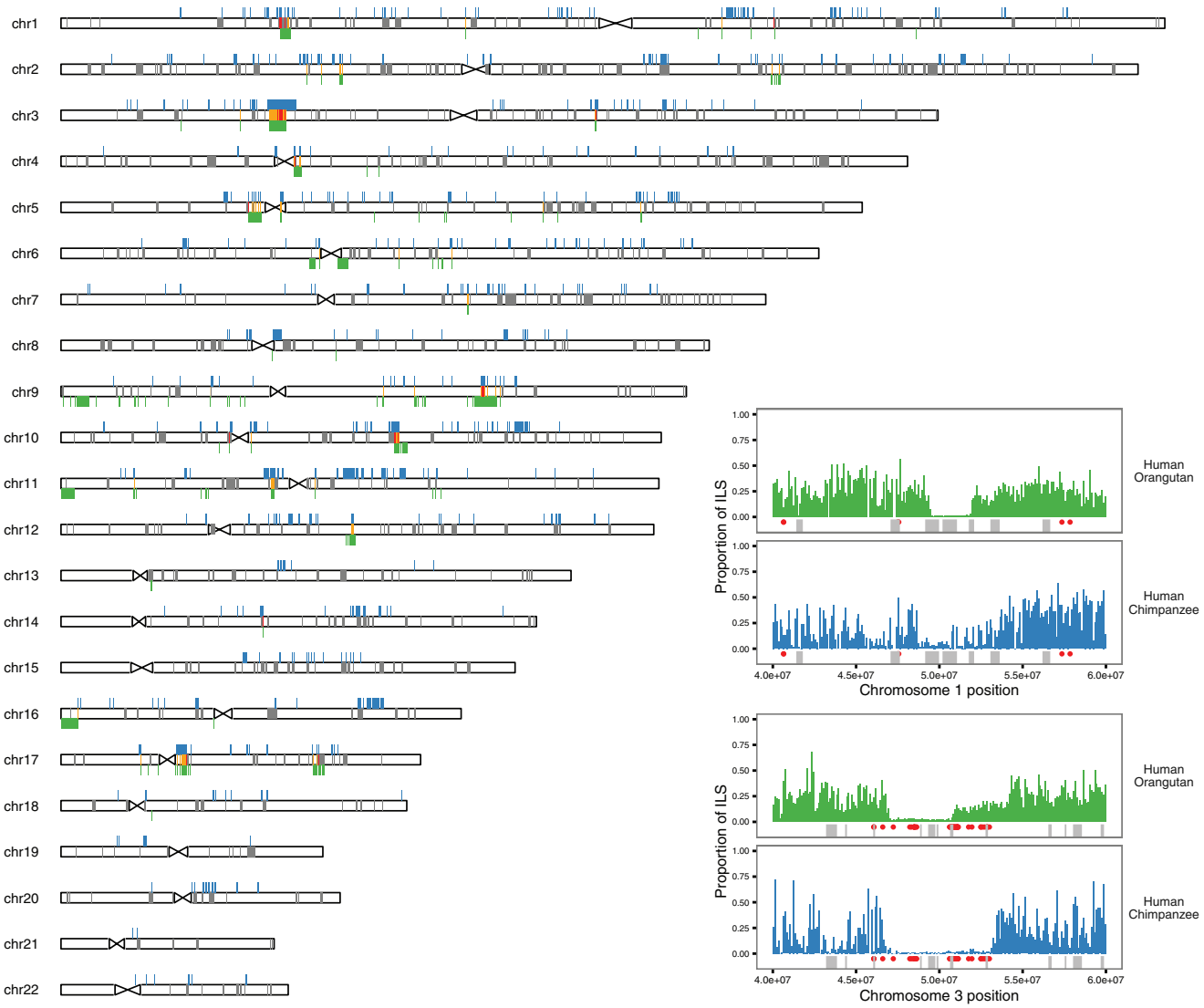


**FIG. 3.** Chromosome ideograms showing the genomic distribution of identified low-NCT regions. Regions in the human–chimpanzee ancestor are shown in blue above each chromosome. Regions in the human–orangutan ancestor are shown below in green. Predicted selective sweeps in humans are shown as gray regions on chromosomes. Overlap between blue and green regions are shown in orange and overlaps between blue, green and gray regions are shown in red. Inserts show proportions of NCT in 100 kb windows across a region of chromosome 1 (top) and chromosome 3 (bottom); here gray bars show predicted human sweeps and red dots show SNPs associated with positive selection in humans.

*vs* 4.3e6), which results in very similar distributions of the proportion of NCT on a 1Mb scale (fig. 3*B*). To check the consistency of our model, we computed the ancestral effective population sizes expected from the estimated split times and global mean proportions of called NCT. This is readily done by rearranging the formula given in the introduction, and yields effective population sizes for human–chimpanzee and human–orangutan of 117,000 and 131,000, congruous with model estimates.

## Strong and Recurrent Selective Sweeps in the Ancestral Species

To identify genomic regions subject to selective sweeps in the two ancestral species, we searched for contiguous genomic regions where the NCT proportion is reduced to less than 5% in regions of >100 kb (corresponding to >70% reduction in $N_e$). We identify 571 such low-NCT regions in the human–chimpanzee ancestor and 139 in the human orangutan ancestor. 28 human–chimpanzee and 21 human–orangutan regions are larger than 500 kb. The low-NCT regions in the two species show very similar length distributions (supplementary fig. S1, Supplementary Material online). This approach conservatively identifies only the most extreme cases of linked selection, which are expected to be enriched for regions with lower recombination rate. As expected, we find that the mean human (HapMap) recombination rate in low NCT regions (0.55) is about half the global average. To test if long low-NCT regions may arise from the variance in the neutral coalescence process, rather than from variation in the strength of linked selection, we repeated our analysis on two genomic alignments simulated assuming the estimated model parameters, a uniform recombination rate of 1 cM/Mb,

and with a distribution of analyzed alignment across the genome that mirrors the true alignments. In both cases, we do not find any regions that meet the criteria defining a low-ILS region.

The genomic positions of low-NCT regions are displayed in figure 3 together with the genomic position of regions reported to be under positive selection in humans (Akey 2009). Two of the most striking examples are shown in more detail: the bottom insert shows a region on chromosome 3 with very wide and strongly delineated depletions of NCT, which span several megabases in both ancestral species. The depletion extends further along the chromosome in the human–chimpanzee ancestor, as would be expected from the action of independent sweeps. The same region has previously been associated with zero NCT in the bonobo-chimpanzee ancestor (Prüfer et al. 2012), and with strong selection and extremely long admixture tracts between the Bantu and Western African pygmy populations (Jarvis et al. 2012). Red dots represent nucleotide polymorphisms (SNPs) reported as extreme by at least one of four different methods for inference of selection in these African populations. The top insert in figure 3 shows a region on chromosome 1 depleted of NCT in the human–orangutan ancestor only.

The visual impression of figure 3 suggests that the positions of sweeps in the two ancestral species are shared to an appreciable extent, and intersecting the two sets of regions indeed reveals a significant overlap (Jaccard *P*-value: $<$1e$-$4). To investigate if this reflects recurrent positive selection, we tested if the low-NCT regions of the ancestors are enriched for the set of regions under positive selection in humans compiled by Akey (2009). We find a significant enrichment in low-NCT regions in both ancestral species. Sweeps reported in humans overlap 159 of the 571 low-NCT regions in the human–chimpanzee ancestor (Jaccard *P*-value: $<$1e$-$4) and 45 of the 139 regions in human–orangutan (Jaccard *P*-value: 0.0141). However, scans for positive selection in humans only represent the most recent few hundred thousand years and may thus not identify selection in regions that only rarely experience sweeps. To accommodate this bias, we focused on low-NCT regions identified in both ancestral species, as these are more likely to represent regions subject to a higher frequency of sweeps. We identify 71 such regions (totaling 15 Mb) of which 20 regions overlap reported sweeps in humans (Jaccard *P*-value: 0.0055). The median recombination rates of the low-NCT regions are 0.24 cM/Mb for human–chimpanzee regions and 0.34 cM/Mb for human–orangutan regions. Although this is considerably lower than the genome average (e.g., the median recombination rate of non-overlapping 100 kb windows along the genome is 0.8 mM/Mb), such lower recombination rates are not confined to a small part of the genome (e.g. ~25% of non-overlapping 100 kb windows along the genome have mean recombination rates below 0.3). Hence, sharing of low-recombining regions between the ancestral species is thus unlikely to solely explain the highly significant overlaps between low-ILS regions. Our observations thus not only indicate that depletions of NCT have the power to identify genomic regions subject to ancestral sweeps, but also that many genomic regions are recurrently affected by sweeps.

Individual hard sweeps are expected to completely abolish ILS in the swept regions. To assess how often such strong sweeps contribute to low-NCT regions, we identified all consecutive genomic regions larger than 100 kb with a NCT proportion of ~0 (allowing rare and very short NCT segments that may be falsely called). In the human–chimpanzee ancestor, we identify 742 such regions overlapping 370 of the 571 low-NCT regions. In the human–orangutan ancestor, we identify 25 such regions overlapping 17 of the 139 low-NCT regions. These results suggest that strong hard sweeps contribute to a substantial part of the low-NCT regions in both species.

## The Genome-Wide Impact of Selective Sweeps Is Stronger in the Human–Chimpanzee Ancestor

An alternative way to investigate the intensity and genomic reach of linked selection is to compute the mean proportion of NCT as a function of distance to the closest human protein-coding RefSeq gene (Scally et al. 2012). As expected, we find that the NCT proportion is strongly reduced close to genes, and that this reduction reaches far away from genes (fig. 4). In the human–chimpanzee ancestor, the proportion of NCT increases with distance at least 500 kb away from genes. As the median and mean distances between the analyzed genes is only ~120 and ~370 kb, this suggests that levels of NCT in most intergenic regions are affected by linked selection.

The two ancestral species show striking differences in the relation between NCT reduction and physical distance: the reduction in NCT proportion is stronger and reaches further away from genes in the human–chimpanzee ancestor than in the human–orangutan ancestor, suggesting that positive selection was stronger or more frequent in the human–chimpanzee ancestor. The few genomic regions further than 1 Mb from any gene experience the smallest effect of linked selection and the $N_e$ computed from NCT proportions in these regions thus best approximates the number of randomly mating individuals in each ancestral species. In these regions, the human–chimpanzee ancestor shows an NCT proportion of 38%, which drops to 25% closest to genes. The human–orangutan ancestor shows a less pronounced drop from 34% to 26%. The effective population sizes, which correspond to these NCT proportions, can be calculated using the formula given in the introduction. In the human–chimpanzee ancestor, $N_e$ drops from 163,000 $>$1 Mb from genes to 93,000 at the gene border, corresponding to a 43% reduction in $N_e$. In the human–orangutan ancestor $N_e$ drops only 30%, from 161,000 to 113,000. We note that despite strong differences in linked selection close to genes, the mean effective population sizes estimated across the entire genome remain very similar between the two species.

The difference in reduction of NCT around genes is informative of the relative contributions of selective sweeps and background selection. Since we compare closely related

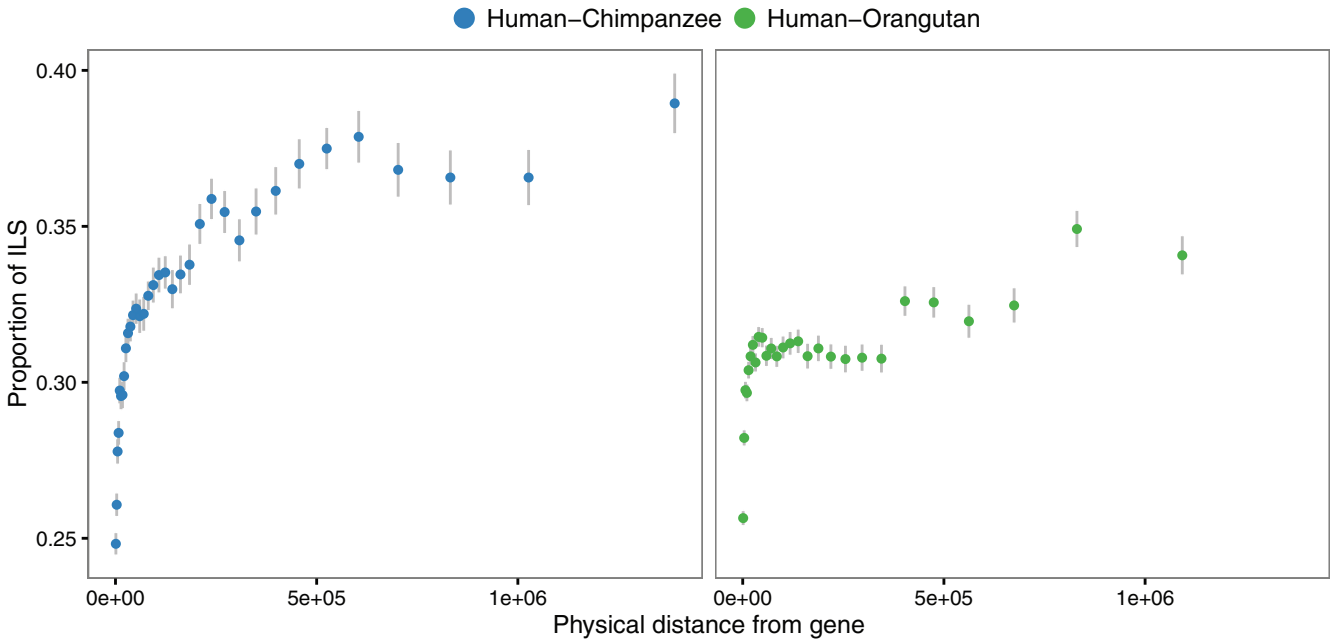


FIG. 4. The NCT proportion as a function of physical distance to the nearest gene. The human–chimpanzee ancestor is shown left and the human–orangutan ancestor right. Bins are selected to put approximately the same amount of alignment in each bin. Error bars show 1.96 * standard error where the number of independent observations is adjusted by correcting for first order autocorrelation.

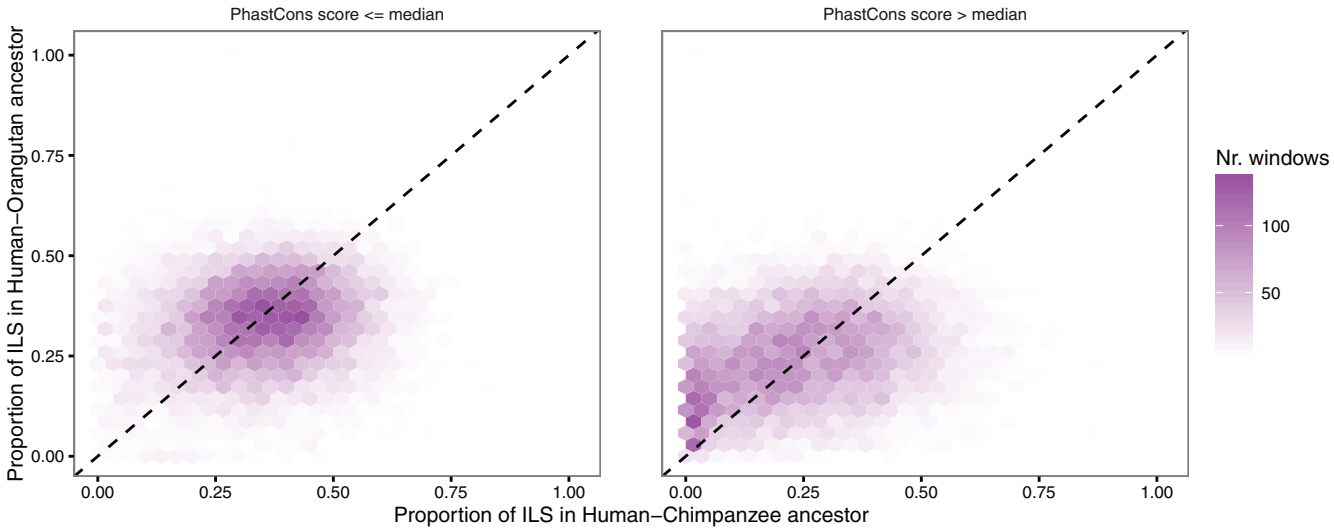


FIG. 5. Two-dimensional histograms each showing the joint distribution of NCT proportions in the two species across orthologous 100 kb windows. Windows are partitioned equally according to PhastCons score showing windows with lower PhastCons scores to the left and windows with higher score to the right.

species that are expected to show highly similar levels of background selection, we can identify upper and lower bounds for the contribution from selective sweeps in the human–chimpanzee ancestor. If we assume that background selection does not contribute to the reduction in human–chimpanzee $N_e$, then sweeps are responsible for the entire 43% reduction near genes. If background selection completely explains the reduction in human–orangutan $N_e$, then sweeps must be responsible for only the difference between the two species, which amounts to a 13% reduction in $N_e$.

To more generally address the impact of linked selection across individual genomic regions we compared the proportions of NCT in non-overlapping orthologous 100 kb windows along the genomes of the human–chimpanzee and human–orangutan ancestors (see Methods). For each window, we also computed the mean PhastCons score (Siepel et al. 2005), which represents the posterior probability that a genomic position is evolutionarily conserved across 46 vertebrate species. Here we use it as a proxy for the functional importance of genome sequence. We partitioned all 100 kb windows into two equal-sized data sets with low and high

PhastCons scores respectively. The two-dimensional histograms in figure 5 show the joint distributions of NCT proportions in orthologous 100 kb windows in the human–chimpanzee and human–orangutan ancestors. The left histogram shows the distribution for windows with lower PhastCons scores (scores below median score), and the right shows the distribution for higher scores. The distribution of NCT proportions for windows with lower PhastCons scores is symmetric, which suggests that the action of linked selection is very similar in genomic regions with a low density of functional sequence. It further provides assurance that our inference of NCT proportion in the two species does not itself produce systematic differences. In contrast, the distribution of NCT proportions for windows with higher PhastCons scores is strongly asymmetric and reveals a strong excess of 100 kb windows showing very low NCT proportions only in the human–chimpanzee ancestor. This observation further supports a stronger role of selective sweeps in the human–chimpanzee ancestor than in the human–orangutan ancestor.

The results presented in figure 5 shows that the action of linked selection may be revealed by the relationship between PhastCons score and difference in NCT proportion between orthologous regions. In regions of extreme background selection and in regions subject to recurrent selective sweeps, we expect that the difference between NCT proportions will be reduced because the NCT proportions are depressed in both species. In contrast, if a region was subject to a sweep in only of the ancestral species. In contrast, sweeps depleting NCT proportions in only one species will result in a larger difference in NCT proportion. Hence, whereas both recurrent sweeps and background selection will associate functional regions (high PhastCons scores) with smaller differences in NCT proportion, only sweeps unique to one species will associate such regions with larger differences in NCT proportion. To test this prediction, we partitioned all orthologous 10 kb windows according to the difference in NCT proportions between the two ancestral species and plotted PhastCons score as a function of this difference (fig. 6). The resulting curve shows a distinct U-shape, revealing that both regions with low and high difference in NCT proportion associates with higher PhastCons score. We see no way to explain this pattern that does not invoke a significant contribution from sweeps unique to each species.

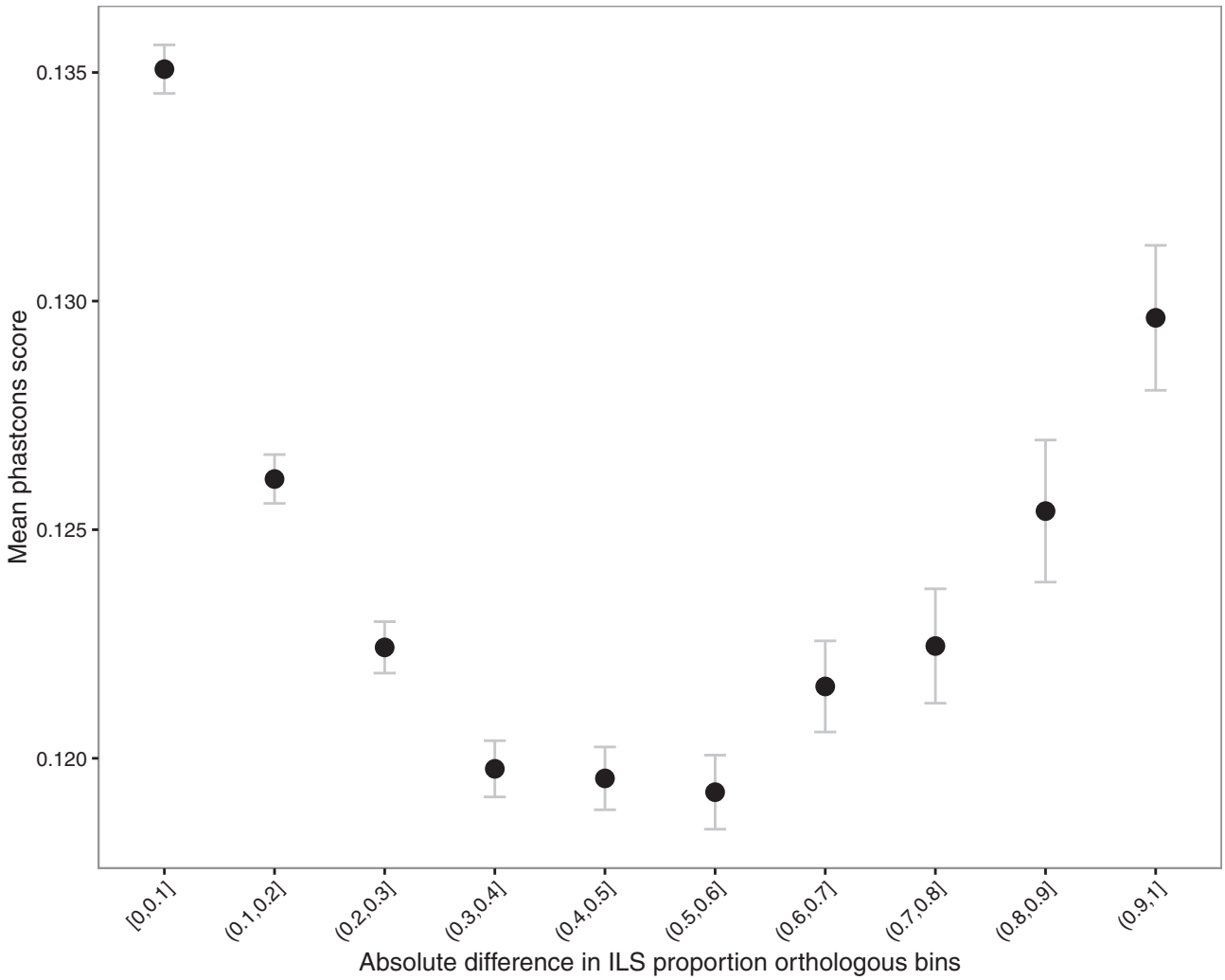


FIG. 6. Mean PhastCons score for 10kb windows grouped by absolute difference in NCT proportion between orthologous windows. Error bars represent 1.96 * standard error.

## Discussion

We have presented a new approach to detect selective sweeps in ancestral species. Through inference of ILS between three closely related species, the method interrogates the effect of linked selection on ancestral diversity and identifies selective sweeps as wide regions depleted of NCTs. As opposed to polymorphism-based methods used to infer sweeps in extant species, this method is not biased by mutation rate variation. More importantly, the method is evenly sensitive to sweeps across the span of time between the two speciation events that define an ancestral species and is thus able to address the effect of sweeps across evolutionary time scales. The method is augmented by a comparative approach, which contrasts evidence of linked selection in two closely related ancestral species expected to show very similar levels of background selection. Together, the two approaches allow us to evaluate the relative impact of sweeps and background selection on ancestral diversity across the genome.

We find that strong sweeps were common in the two ancestral species but that the human–chimpanzee ancestor saw a larger number of sweeps than the human–orangutan ancestor. We report several cases where swept regions span several megabases, likely the result of multiple sweeps. We further find that strong sweeps have recurrently hit the same genomic regions in the two ancestral species suggesting that many regions of our genome have been subject to recurrent sweeps over the past 20 million years. If the low-ILS regions were random outcomes of highly turbulent demographic histories this overlap would not be observed. As expected for independent sweeps, the regions swept in each ancestral species overlap but are not identical in physical extent. In addition to the larger number of low-NCT regions in the human–chimpanzee ancestor, a notable difference between the two species is that the human–orangutan ancestor shows a few very long regions (fig. 3 and supplementary fig. S1, Supplementary Material online) half of which overlap low-NCT regions in the human–chimpanzee ancestor. One of several possible explanations is that these regions correspond to a small number of very strong sweeps in the human–orangutan ancestor.

We find that selective sweeps have had a strong impact on diversity across the genomes of the two ancestral species and that the different extent to which diversity is reduced around genes and other functional regions is determined by the frequency of sweeps. This conclusion rests on the observation that the two species were subject to very different levels of linked selection. The human–chimpanzee ancestor shows a stronger and more far reaching reduction in NCT proportion around genes, and regions with very low proportions of NCT in only one species are much more frequent in the human–chimpanzee ancestor.

Throughout this study we have assumed that orthologous regions in the two ancestral species were exposed to very similar levels of background selection but this assumption can be relaxed without compromising our findings. Although a low proportion of NCT at each individual genomic locus may be explained by background selection under highly accommodating assumptions, it cannot consistently explain the generality of our findings when these are considered together. The stronger depression in NCT proportions around genes and the larger number of regions with very low ILS in the human–chimpanzee ancestor cannot be explained by background selection without the temporary emergence of a very large number of recombination cold-spots in the human–chimpanzee ancestor not observed in either the human–orangutan ancestor or in humans and chimpanzees. A large number of segregating inversions in the human–chimpanzee ancestor could potentially mediate such cold-spots. However, between the common ancestor of human and orangutan and that of humans and chimpanzees, only 40 breakpoints are involved in structural rearrangements larger than 100 kb (Locke et al. 2011). Even if all these rearrangements only segregated in the human–chimpanzee ancestor, and if they all drastically lowered recombination rates, they would only explain a small fraction of the low NCT regions we observe. In addition, it does not seem likely that genomic instability or other mechanisms producing cold-spots would arise temporarily in the human–chimpanzee ancestor to then be lost in both humans and chimpanzees. In contrast, all our findings are readily explained by a strong influence of selective sweeps acting at a higher frequency in the human–chimpanzee ancestor than in the human–orangutan ancestor.

In a recent study, Dutheil et al. (2015) used simulations to show that the ILS-depleted regions produced by soft sweeps are much smaller than those produced by hard sweeps. This is because the two sampled lineages (e.g., human and chimpanzee) are not guaranteed to coalesce during a soft sweep (as is the case in a hard sweep) and this leaves much more time for sequence flanking the selected variant to escape by recombination. A very large number of soft sweeps, which each produce a short region without ILS, may contribute to the variation and interspecies differences in NCT proportions, which we observe. However, soft sweeps do not easily explain the wide (>100 kb) regions without ILS that we observe in higher numbers in human–chimpanzee, the longest of which cover several megabases. In addition, the much shorter ILS depleted regions produced by soft sweeps cannot explain that the depression in NCT proportions extends several hundred kilobases away from genes.

A higher frequency of sweeps in the human–chimpanzee species may result from a transition of environment prompting an accelerated adaptive evolution. An alternative explanation is that adaptive selection is limited by the introduction of new mutations. In contrast to background selection, for which theoretical results predict that the effect of background selection do not depend on either census size or effective population size (Charlesworth et al. 1993; Hudson and Kaplan 1995; Nordborg et al. 1996), the frequency of selective sweeps will increase linearly with census population size. The two ancestral species in this study show very similar effective

population sizes, but these reflect unknown contributions from census size, species demography, population structure, and any sex ratio or mating behavior affecting the variance in number of viable offspring. It is thus possible that the census size of the human–chimpanzee ancestor much exceeded that of the human–orangutan ancestor.

The demographic model used in this study assumes instantaneous speciation events without subsequent gene flow. However, evidence of gene flow during speciation has been detected for the human–chimpanzee split (Mailund et al. 2012). Our model will respond to this model violation by estimating more recent split times and by estimating larger ancestral effective population sizes to accommodate some of this gene flow as population structure. However, estimated proportions of NCT across the genome are not sensitive to this model violation. Our model also assumes a constant $N_e$ in the ancestral species and will accommodate alternative demographic scenarios by estimating an $N_e$ in line with the proportions of ILS in the data.

ILS is a common phenomenon in species groups where speciation events follow in rapid succession and the approach presented here offers the opportunity to analyze the effect of linked selection across multiple internal branches of such species trees, providing new insights into how the strength of adaptive selection varies across evolutionary time scales.

## Methods

### Preparation of Alignment Data

To quantify linked selection in the human–chimpanzee ancestor we analyzed the genomic alignment of human, chimpanzee, and gorilla using orangutan as out-group. The same analysis is performed on the human–orangutan ancestor using the genomic alignment of human, orangutan, and gibbon, with macaque as out-group (fig. 1A). We downloaded the publically available Multiz 11-way alignments from the UCSC genome browser, which uses reference assemblies of gorilla, chimpanzee, human, gibbon, orangutan, baboon, macaque, marmoset, tarsier, mouse lemur, and bush baby. We sorted alignment blocks according to hg19 coordinates and removed overlapping blocks. In processing the alignment for CoalHMM analysis, we extracted the four relevant genomes from each alignment block. The resulting blocks that are separated by no more than 100 alignment positions for the in-group species are concatenated into larger syntenic blocks where possible. These blocks were only included in the subsequent analysis if they contain at least 500 alignment positions and if the content of 'N' characters is below 20%. The total lengths of the resulting human–chimpanzee–gorilla–orangutan and human–orangutan–gibbon–macaque blocks were 1,192 and 1,003 Mb respectively.

### CoalHMM Analysis

We infer population genetic parameters and patterns of ILS using the coalescent hidden Markov model (HMM) framework (CoalHMM). In this HMM the hidden states along the alignment are gene trees with separate topologies and separate coalescent times. The model is applied to a genomic alignment of human, chimpanzee, gorilla and orangutan (using orangutan as outgroup) and to a genomic alignment of human, orangutan, gibbon, and macaque (using macaque as outgroup). The demographic model is a three species isolation model and the demographic parameters are the two ancestral population sizes and two speciation times. These parameters are all scaled with the substitution rate. The HMM has four trees as hidden states: the two most closely related species may find a common ancestor in their ancestral population (fig. 1B, bottom left) or in the population ancestral to all three species (fig. 1B, bottom right), and each of the two most closely related species may find common ancestry with the third species before they find common ancestry with each other (fig. 1B, top left and right). The latter two cases represent non-canonical topologies (NCT). The transition matrix is parameterized using coalescent theory. The probability of emitting an alignment column from a given state is computed as the probability of the state tree given the four bases. Given a set of model parameters, the likelihood is calculated using the forward algorithm and the maximum likelihood is found using a modified Newton–Raphson algorithm.

A separate CoalHMM analysis was performed on approximately 1 Mb of alignment blocks, restarting the HMM between alignment blocks more than 100 bp apart. The resulting estimates are associated with a known bias that is corrected as described in (Dutheil et al. 2009). Briefly, we simulated hundred 1 Mb alignments for combinations of model parameters. To correct the human–chimpanzee analysis, we used (see Dutheil et al. for parameterization): $Ne_{12}$: 50,000, 110,000, 170,000; $Ne_{123}$: 50,000, 70,000, 90,000; T1: 5,000,000, 6,500,000, 8,000,000; T12: 10,000,000, 10,500,000, 11,000,000. To correct the human–orangutan analysis, we used: Ne12: 1e5, 1.5e5, 2e5; Ne123: 120,000, 140,000, 160,000, T1: 19,000,000, 21,000,000, 23,000,000; T12: 26,000,000, 27,000,000, 28,000,000. In all simulations, we used a mutation rate per year of 6e−10, a generation time of 20 years and a recombination rate of 1e−8. Each simulated alignment was subjected to a CoalHMM analysis to estimate model parameters, and the estimation bias on each parameter was computed as the deviation of from the true value. Estimation bias was then corrected using a linear model fitted to explain bias from known values of parameters and their interaction.

### Calling of NCT

Following successful CoalHMM parameter estimation of each 1 Mb of our data, we performed a posterior decoding to obtain posterior probabilities of each hidden state at each alignment position. NCTs were called as alignment segments where a hidden state that represents NCT is associated with the highest posterior probability, and these are reported in hg19 coordinates. The proportions of NCT were reported in 10 kb, 100 kb, and 1 Mb windows in hg19 coordinates. Windows for which less than 10% of the spanned genome is represented by analyzed alignment are not included in subsequent analyses.

### Scan for Low-NCT Regions

To identify regions with very low NCT proportions, we computed the running proportion of NCT in all 100 kb windows

shifted by 1 kb across the genome. We first identified regions as consecutive genomic regions that did not overlap a 100 kb window with more than 5% NCT or with less than 10 kb analyzed alignment. Such regions of at least 100 kb in length are reported as low-NCT regions. To identify regions with ~0 NCT, we first identified consecutive non-NCT regions only interrupted by NCT-regions shorter than 50 kb. Of these regions, we only report those longer than 100 kb, with an NCT proportion of at most 0.1%, and with coverage of alignment blocks of at least 20%. The statistical significance of overlaps between genomic regions was evaluated using the Jaccard-measure implemented in the R package GenometriCorr.

## Supplementary Material

Supplementary figure S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

This work was supported by the Danish Council For Independent Research (grant number 12-125062 and 4181-00358).

## References

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* 19:711–722.

Carbone L, Harris AR, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* 13:195–201.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. *PLOS Biol.* 13:e1002112.

Dutheil JY, Ganapathy G, Hobolth A, Mailund T, Uyenoyama MK, Schierup MH. 2009. Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* 183:259–274.

Dutheil JY, Munch K, Nam K, Mailund T, Schierup MH. 2015. Strong selective sweeps on the X chromosome in the human-chimpanzee ancestor explain its low divergence. *PLoS Genet.* 11:e1005451.

Elyashiv E, Sattath S, Hu T, Strustovsky A, McVicker G, Andolfatto P, Coop G, Sella G. 2015. A genomic map of the effects of linked selection in *Drosophila*. arXiv:1408.5461.

Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. *Genome Res.* 24:885–895.

Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. 2014. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Mol Biol Evol.* 31:1850–1868.

Hobolth A, Christensen OF, Mailund T, Schierup MH. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* 3:e30007.

Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* 21:349–356.

Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. *Genetics* 141:1605–1617.

Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, Froment A, Bodo J-MM, Beggs W, Hoffman G, et al. 2012. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet.* 8

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang S-PP, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orangutan genomes. *Nature* 469:529–533.

Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, Lunter G, Prüfer K, Scally A, Hobolth A, et al. 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genet.* 8:e1003125.

Mailund T, Munch K, Schierup MH. 2014. Lineage sorting in apes. *Annu Rev Genet.* 48:519–535.

Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. *Genetical Res.* 67:159–174.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* 499:471–475.

Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486:527–531.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483:169–175.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15: 1034–1050.

Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122:957–966.