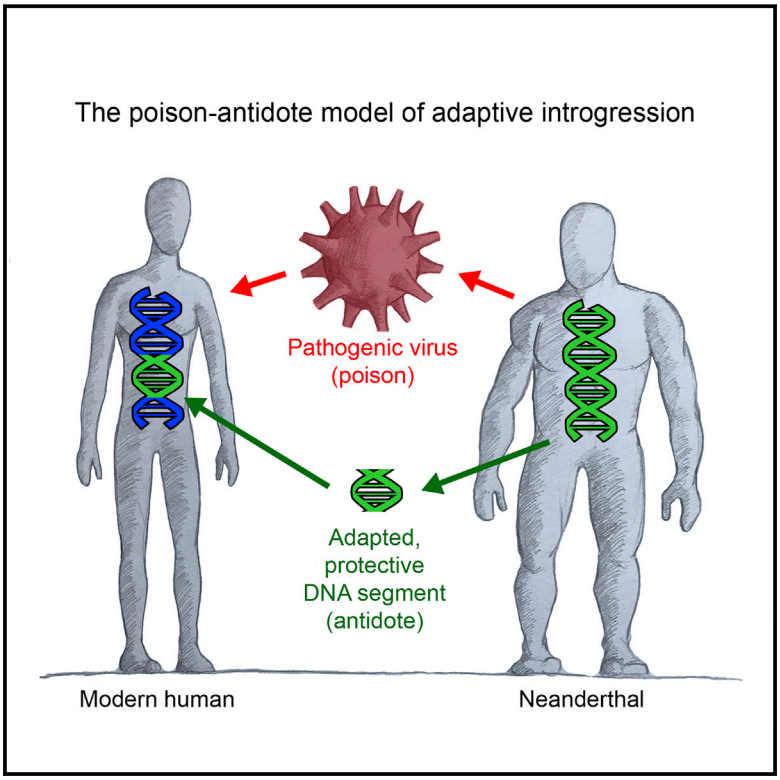
Article

Cell

# Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans

## Graphical Abstract




## Authors

David Enard, Dmitri A. Petrov

## Correspondence

denard@email.arizona.edu

## In Brief

Human genome evolution after Neanderthal interbreeding was shaped by viral infections and the resulting selection for ancient alleles of viral-interacting protein genes.

## Highlights

- Neanderthals and modern humans interbred and exchanged viruses
- Neanderthal DNA introgressed in modern humans helped them adapt against viruses
- Neanderthal DNA-based adaptation was particularly strong against RNA viruses in Europeans
- Ancient epidemics can be detected through the lens of abundant host genomic adaptation

Enard & Petrov, 2018, Cell *175*, 360–371
October 4, 2018 
https://doi.org/10.1016/j.cell.2018.08.034

CellPress

Cell

Article

# Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans

David Enard[1,3,*] and Dmitri A. Petrov[2]
[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA
[2]Department of Biology, Stanford University, Stanford, CA, USA
[3]Lead Contact
*Correspondence: denard@email.arizona.edu
https://doi.org/10.1016/j.cell.2018.08.034

## SUMMARY

**Neanderthals and modern humans interbred at least twice in the past 100,000 years. While there is evidence that most introgressed DNA segments from Neanderthals to modern humans were removed by purifying selection, less is known about the adaptive nature of introgressed sequences that were retained. We hypothesized that interbreeding between Neanderthals and modern humans led to (1) the exposure of each species to novel viruses and (2) the exchange of adaptive alleles that provided resistance against these viruses. Here, we find that long, frequent—and more likely adaptive—segments of Neanderthal ancestry in modern humans are enriched for proteins that interact with viruses (VIPs). We found that VIPs that interact specifically with RNA viruses were more likely to belong to introgressed segments in modern Europeans. Our results show that retained segments of Neanderthal ancestry can be used to detect ancient epidemics.**

## INTRODUCTION

After their divergence 500,000 to 800,000 years ago, modern humans and Neanderthals interbred at least twice: the first time ~100,000 years ago (Kuhlwilm et al., 2016) and the second ~50,000 years ago (Fu et al., 2015; Green et al., 2010; Pääbo, 2015; Sankararaman et al., 2012, 2014). The first interbreeding episode left introgressed segments (IS) of modern human ancestry within Neanderthal genomes (Kuhlwilm et al., 2016), as revealed by the analysis of ancient DNA from a single Altai Neanderthal individual sequenced by Prüfer et al. (2014). This first interbreeding event appears not to have left any detectable segments of Neanderthal ancestry in extant modern human genomes (Kuhlwilm et al., 2016). In contrast, the second interbreeding episode left detectable IS of Neanderthal ancestry within the genomes of non-African modern humans (Fu et al., 2015; Green et al., 2010; Prüfer et al., 2014; Sankararaman et al., 2014; Vernot and Akey, 2014).

Recent advances in the detection of introgression have led to the discovery that the majority of genomic segments initially introgressed from Neanderthals to modern humans were rapidly removed by purifying selection. Harris and Nielsen (2016) estimated that the proportion of Neanderthal ancestry in modern human genomes rapidly fell from ~10% to the current levels of 2%–3% in modern Asians and Europeans (Fu et al., 2015; Juric et al., 2016).

This history of interbreeding and purifying selection against IS raises several important questions. First, among the introgressed sequences that were ultimately retained, can we detect which sequences persisted by chance because they were not as deleterious or not deleterious at all to the recipient species, and which persisted not despite natural selection but because of it—that is, which IS increased in frequency due to positive selection? If any of the introgressed sequences were indeed driven into the recipient species due to positive selection, can we determine which pressures in the environment drove this adaptation?

Recently we found that proteins that interact with viruses (virus-interacting proteins [VIPs]) evolve under both stronger purifying selection and tend to adapt at much higher rates compared to similar proteins that do not interact with viruses (Enard et al., 2016). We estimated that interactions with viruses accounted for ~30% of protein adaptation in the human lineage (Enard et al., 2016). Because viruses appear to have driven so much adaptation in the human lineage, and because it is plausible that when Neanderthals and modern humans interbred they also exchanged viruses either directly by contact or via their shared environment, we hypothesized that some introgressed sequences might have provided a measure of protection against the exchanged viruses and were driven into the recipient species by positive directional selection. Consistent with this model, several cases of likely adaptive introgression (Gittelman et al., 2016; Racimo et al., 2015, 2017) from Neanderthals to modern humans involve immune genes that are specialized to deal with pathogens including viruses (Abi-Rached et al., 2011; Dannemann et al., 2016; Deschamps et al., 2016; Houldcroft and Underdown, 2016; Mendez et al., 2012, 2013; Nedelec et al., 2016; Quach et al., 2016; Sams et al., 2016).

Here, we test this hypothesis by assessing whether VIPs are enriched in IS overall and, more specifically, in longer and more frequent IS that are more likely to have been driven into the recipient genome by positive directional selection. Because

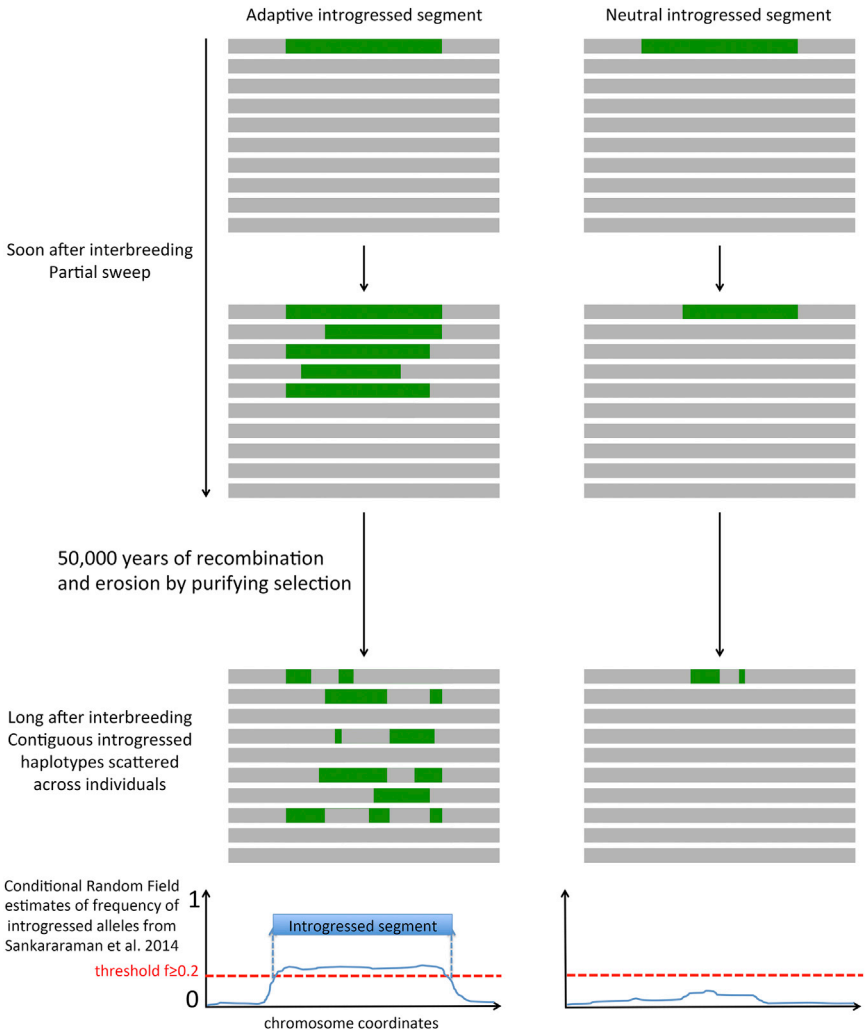



**Figure 1. Higher Frequency and Longer Adaptive Introgressed Segments Compared to Neutral Ones**

Green: introgressed Neanderthal segment. Grey: genomic background of the receiving population. Contiguous fragments of IS result in a longer region with a higher probability/frequency of introgressed alleles. The frequencies of alleles inherited from Neanderthals were estimated by Sankararaman et al. (2014) using the conditional random field approach. Note that Figure 1 is a schematization meant to highlight the differences between the neutral and the adaptive introgression scenario. The represented frequencies or length of IS are meant for illustration purpose only and do not represent actual cases present in our dataset.

See also Figure S1.

purifying selection strongly affects the probability of introgressed sequences being retained by chance, we test introgression enrichments at VIPs after controlling for the stronger purifying selection at VIPs as well as many other potentially confounding factors.

The basic logic of the analysis is as follows. If positive directional selection occurs soon after interbreeding, adaptive Neanderthal introgressed haplotypes are expected to rapidly increase in frequency before being fragmented by recombination and thus should lead to the presence of long and frequent IS as a result (Figure 1). Over time, recombination is expected to break up IS while purifying selection should remove deleterious alleles that hitchhiked together with the adaptive variant(s). As a result, the signal should erode over time. However, because IS scattered across multiple individuals by recombination can be identified and aggregated into single contiguous genomic regions, as was done by Sankararaman et al. (2014) and shown schematically in Figure 1, the originally adaptive introgressed segment of Neanderthal ancestry might still be identifiable as aggregated segments of Neanderthal ancestry. Furthermore, the frequency and length of such retained regions of Neanderthal ancestry can be assessed (Figure S1; STAR Methods).

Here, we gathered a large dataset of thousands of VIPs and showed that they are strongly enriched within longer and more frequent IS of Neanderthal origin in modern human genomes, as well as in the longer IS of modern human origin in the Neanderthal genome. Furthermore, we found that VIPs that specifically interact with RNA viruses are particularly enriched in Neanderthal IS in modern European genomes compared to VIPs that interact with DNA viruses. We provide a number of arguments suggesting that it is specifically adaptation in response to viruses that drove these enrichments. We next identify several viruses as likely agents of selection, as well as a number of specific VIPs as likely targets of adaptation. Finally, we estimate that adaptation overall, and specifically adaptation in response to viruses, was an important force in the history of those Neanderthal IS that were ultimately retained in modern human genomes.

## RESULTS

### VIPs and Introgression Data

We focused on 4,534 VIPs (~20% of the human proteome; Table S1) that engage in defined physical interactions with many viruses, including 20 human viruses known to interact with at least 10 VIPs (Table S1; STAR Methods). VIPs were annotated based on interactions with modern viruses, but these can be thought of as proxies for related viruses in ancient populations. This extension is supported by the fact that related viruses tend to use similar host VIPs (Enard et al., 2016). For example, VIPs interacting with HIV are also likely to interact with other lentiviruses. Thus, if enrichment of adaptive introgressions of HIV-interacting VIPs is observed, this presents evidence of past adaptation related to a lentivirus rather than to HIV itself.

To estimate enrichment of introgression at VIPs, we used the IS of Neanderthal ancestry in East Asian and European modern human genomes identified by Sankararaman et al. (2014). These authors used a conditional random field (CRF) approach to estimate the frequencies and lengths of IS, and here, we simply reuse these estimates (STAR Methods). In brief, for each position in the genome that is marked by a SNP, the CRF model provides a posterior probability that any randomly sampled modern haplotype contains an allele of Neanderthal origin. Smoothed over a set of contiguous SNPs, this also provides a regional estimate of the frequency of Neanderthal ancestry (Figures 1 and S1; STAR Methods). The method also generates a list of high-confidence Neanderthal haplotypes present in some individuals that we also utilized in this paper. We analyzed East Asian and European modern human populations separately because they had distinct histories of interbreeding with Neanderthals (Kim and Lohmueller, 2015; Vernot and Akey, 2015).

## Differences in Confounding Factors between VIPs and Non-VIPs

Of the 4,534 VIPs, 1,920 VIPs were identified by low-throughput approaches and hand-curated from the virology literature (LT-VIPs), whereas 2,614 VIPs were identified by high-throughput approaches (Guirimand et al., 2015) (HT-VIPs; STAR Methods). Previously, using a smaller set of VIPs, we showed that they tend to be unusually highly conserved (Enard et al., 2016). Before studying patterns of introgression between modern humans and Neanderthals, we first confirmed our previous findings with the current, expanded set of VIPs. Using the full set of 4,534 human VIPs, we showed that compared to non-VIPs, VIPs do exhibit (Table S2; STAR Methods): (1) a lower average ratio of nonsynonymous to synonymous polymorphisms, (2) a higher proportion of rare, likely deleterious polymorphisms, reflected in the more negative values of Tajima's D, and (3) a higher density of functional and possibly deleterious segregating variants inferred by FUNSEQ (Fu et al., 2014; Khurana et al., 2013). VIPs are also (1) found in regions of the genome with higher densities of coding sequences (Yates et al., 2016), regulatory sequences (ENCODE Project Consortium, 2012) and conserved genomic segments defined by PhastCons (Siepel et al., 2005); (2) are more highly expressed (GTEx Consortium, 2015); and (3) have more interacting protein partners in the network of human protein–protein interactions than non-VIPs (Table S2) (Luisi et al., 2015; Stark et al., 2011).

In summary, we confirmed that the 4,534 VIPs in our set are more conserved and have more segregating deleterious variants than non-VIPs (Enard et al., 2016). The higher levels of conservation and purifying selection of VIPs, and higher loss rate of more constrained sequences of Neanderthal ancestry from the modern human genomes, implies that IS containing VIPs are more likely to have been removed by purifying selection. It is thus essential to control for varying levels of purifying selection to increase the power of detection of enrichment of VIPs in IS. An imperfect control for purifying selection is indeed likely to make the test of the enrichment of VIPs in the introgressed regions overly conservative, as under the null hypothesis of no adaptation preferentially targeting VIPs, we would expect VIPs to be present in the Neanderthal IS less often compared to non-VIPs.

Although we combined LT and HT-VIPs into a single VIP category, throughout the paper, we systematically confirmed that the major results obtained when combining all VIPs also held true when using only the hand-curated LT-VIPs.

## Controlling for Confounding Factors between VIPs and Non-VIPs

To determine if VIPs are enriched in segments introgressed between modern humans and Neanderthals, it is important to first define which other factors, in addition to the levels of constraint, affect the occurrence of IS along the genome independently of interactions with viruses. In the genome, factors that affect the occurrence of IS should differ inside compared to outside IS. We must therefore match VIPs and non-VIPs for genomic factors that (1) differ inside versus outside IS, and (2) also differ between VIPs and non-VIPs (Figure 2A).

We defined all the genomic factors that differed between IS and non-IS regions in both directions, including GC content, the number of human protein-protein interactions, and multiple parameters controlling for levels of deleterious variants (i.e., Tajima's D, FUNSEQ score, and densities of coding, regulatory, and conserved elements) (Figures 2 and S2; Table S2; STAR Methods). Because all of these genomic parameters also varied between VIPs and non-VIPs (Table S2), we used a bootstrap test to first match the VIPs with control non-VIPs for all relevant factors (Figures 2B and 2C; Tables S3 and S4; STAR Methods). We also systematically matched VIPs and non-VIPs with similar recombination rates in the bootstrap test (STAR Methods), and we assessed whether the enrichment of VIPs in the IS becomes more pronounced in regions of higher recombination rate (Hinch et al., 2011). We did this to further confirm that adaptive introgression rather than heterosis explains our results. Indeed, Kim et al. (2017) have recently shown that heterosis can mimic adaptive introgression in regions of low recombination but to a smaller extent in regions of high recombination.

## VIPs Are Enriched for Introgressed Segments from Neanderthals to Modern Humans

The model of positive directional selection of Neanderthal IS predicts an enrichment of Neanderthal ancestry at VIPs. More specifically, positive directional selection should have left Neanderthal IS at VIPs that are longer and at higher frequencies than Neanderthal segments that overlap non-VIPs. Long IS, in particular, are expected at VIPs if positive directional selection occurred not too long after interbreeding. We first used the bootstrap test to show that significantly more Neanderthal IS overlap VIPs than non-VIPs both in East Asia (169 segments overlapping VIPs versus 136 overlapping matched non-VIPs on average, bootstrap test $p < 10^{-3}$) and in Europe (154 segments overlapping VIPs versus 128 overlapping matched non-VIPs on average, bootstrap test $p = 0.003$).

We further used the hypergeometric test to detect a strong and highly significant excess of long and frequent Neanderthal IS encompassing VIPs in both East Asian and European populations (Figure S3). Specifically, the excess of VIPs in the long IS ($\geq$100 kb) is significantly higher than in all (>0 kb) IS both in East Asians (Figure S3A, hypergeometric one-tailed test $p = 1.2 \times 10^{-5}$) and Europeans (Figure S3B, $p = 0.007$). Likewise,

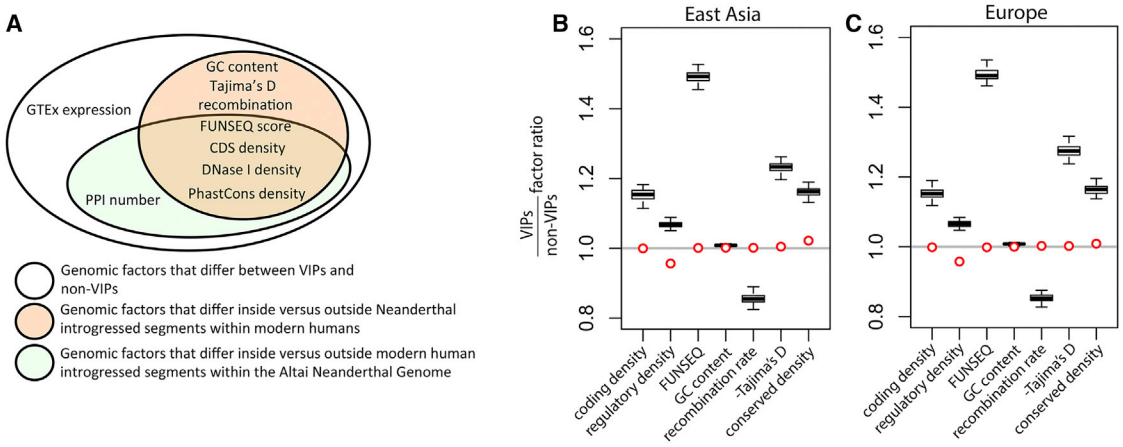



**Figure 2. Confounding Factors Included in the Bootstrap Test**

(A) Venn diagram of the factors that could confound the comparison of VIPs with non-VIPs, that is those that differ both between VIPs and non-VIPs and also inside and outside introgressions.

(B) Bootstrap matching of potential confounding factors between VIPs and control non-VIPs for introgression from Neanderthals to East Asian modern humans. Boxplot intervals represent the discrepancy between confounding factors between VIPs and non-VIPs before bootstrap matching. The red dots represent the difference in confounding factors between VIPs and non-VIPs after bootstrap matching. Note that for the factor "regulatory density" the residual discrepancy is in the conservative direction.

(C) Same as (B) but for introgression from Neanderthals to European modern humans.

See also Figure S2 and Table S2.

the excess of VIPs in the IS at frequencies >15% is higher than that in all IS both in Asians (p = 0.05) and Europeans (p = 0.025). Most importantly, the excess of VIPs in the long (≥100 kb) segments at frequencies >15% is significantly higher than that in all segments (>0 kb) at frequencies higher than 15% both in East Asians (Figure S3A, $p = 2.5 \times 10^{-4}$) and Europeans (Figure S3B, p = 0.034). This significant excess of long and frequent Neanderthal IS is the hallmark of directional selection. These patterns remained when we restricted the analysis to only very high confidence segments of Neanderthal ancestry (Figures S3C and S3D; STAR Methods) and are also robust to variations in the definition of IS (STAR Methods).

## A General Trend toward Longer and More Frequent Neanderthal Introgressed Segments at VIPs

The hypergeometric test we implemented required fixing arbitrary thresholds of length and frequency of the IS. We thus further verified whether we could observe a more general trend toward an increase in Neanderthal ancestry at VIPs as we increased the length and frequency of IS across a wide range of thresholds.

Figure 3A shows that the excess of Neanderthal ancestry at VIPs does tend to progressively increase with larger length thresholds as well as with larger frequency thresholds (see also Figures S4A and S4B). Moreover, the excess of Neanderthal ancestry at VIPs is significantly greater in high-recombination regions of the genome (hypergeometric test using IS larger than 100 kb and at frequencies higher than 15%; East Asia p = 0.016, Europe p = 0.039) (Figure 3B) as expected under the adaptive introgression model. These patterns remained when (1) we restricted the analysis to LT-VIPs (Figure S4C), or (2) we used a different recombination map (Kong et al., 2010) (Figures S4D and S4E), or (3) when we added a control for background selection (McVicker et al., 2009) (Figures S4F and S4G). Furthermore, VIPs and control non-VIPs have very similar numbers of segregating variants (241 segregating variants on average in VIPs and 239 in non-VIPs in East Asia, p = 0.32. 247 in VIPs and 243 in non-VIPs in Europe, p = 0.2) revealing that VIPs and control non-VIPs have similar amounts of highly constrained sites.

## Adaptive Introgressed Loci Are Strongly Enriched among VIPs

Overall, the enrichment of Neanderthal ancestry, and specifically the strong enrichment of long and frequent IS at VIPs, suggest that viruses frequently drove adaptive introgression after interbreeding between Neanderthals and modern humans. It is important to note, however, that so far we have not used information on adaptive introgression at the level of specific loci. Several scans for adaptive introgressed loci previously identified multiple loci with locus-specific evidence of adaptive introgression (Gittelman et al., 2016; Jagoda et al., 2017; Racimo et al., 2017). If the overall enrichment of long and frequent IS reflects the impact of adaptive introgression at VIPs, then VIPs should be particularly strongly enriched in loci previously shown to have undergone adaptive introgression. Here, we used the loci identified by three different scans (Gittelman et al., 2016; Jagoda et al., 2017; Racimo et al., 2017) and estimated their enrichment at VIPs. In line with the overall enrichment of Neanderthal ancestry at VIPs being due to adaptive introgression, we found a very strong excess of adaptive IS at VIPs compared to non-VIPs (Figure S4H). As expected, the excess is very pronounced for long and frequent adaptive IS (Figure S4I). Thus, these results further show that adaptive introgression had a substantial impact at VIPs after interbreeding.

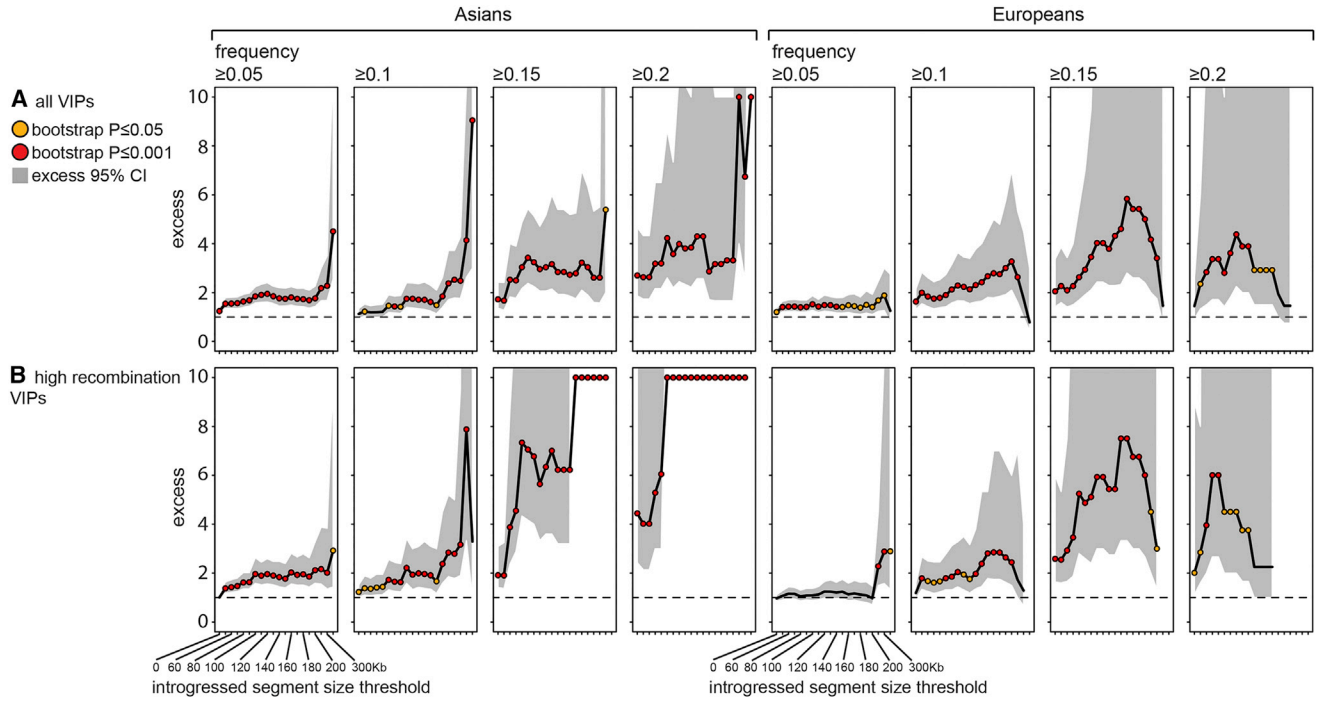



**Figure 3. Excess of Introgression from Neanderthals to Modern Humans at VIPs**

The graphs show the relative excess (y axis) of IS of Neanderthal ancestry within Asian and European modern human genomes as a function of increasing lower segment size threshold (x axis) and increasing lower segment frequency threshold (from left to right). The black line is the observed excess. The gray area is the 95% confidence interval. For representation purposes, any excess >10 is depicted as 10 in the graphs. Segment size thresholds for which the confidence interval is not represented correspond to thresholds beyond which there are no IS overlapping control non-VIPs. Orange dots, bootstrap test $p < 0.05$; red dots, bootstrap test $p < 0.001$. The dashed line indicates an excess of 1. The lower segment size threshold was increased until there were fewer than three remaining IS overlapping VIPs or non-VIPs included in the matching. The points that have no confidence interval are points where VIPs still have several overlapping IS, but where control non-VIPs no longer have any overlapping introgressed segment.

(A) Excess in all VIPs.

(B) Excess in VIPs across high recombination regions (>1.5 cM/Mb, the median recombination rate within IS).

See also Figures S3 and S4 and Tables S3, S4, and S5.

## Correlation between Segment Length and the Number of VIPs

The enrichment of long IS suggests that positive directional selection drove adaptive introgression at VIPs. However, the excess of VIPs in the long IS could also be due to unaccounted clustering of multiple VIPs containing multiple adaptive, balanced alleles, instead of isolated alleles under directional selection. This possibility is unlikely, however, because we would then expect a positive correlation between the number of VIPs within an IS and the length of this IS. We found no such correlation (partial correlations controlling for the total number of genes within an introgressed segment; Europe: Spearman's $\rho = 0.06$, $p = 0.6$; East Asia: $\rho = 0.1$, $p = 0.3$).

## Estimating the Proportion of Adaptive Introgressed Segments

The excess of long and frequent IS at VIPs can be used to estimate the rate of adaptive introgression. The number of long and frequent IS at VIPs above the expected number based on matched non-VIPs is a lower bound for the proportion of adaptive IS. For example, if there were 50 IS at VIPs versus 20 IS at control non-VIPs, we would estimate that the 30 additional long and frequent segments at VIPs were due to adaptive introgression.

Overall we identified 121 (versus 66 expected) segments longer than 100 kb overlapping VIPs in East Asia (bootstrap test $p < 10^{-3}$) and 103 (versus 68 expected) in Europe ($p < 10^{-3}$). For the introgressions that are long ($\geq$100 kb) and at high frequency ($\geq$15%) and thus more likely to be adaptive, the absolute counts are smaller but the enrichment is even more pronounced: 36 (versus 11 expected) segments in Asia ($p < 10^{-3}$), and 19 (versus 6 expected) in Europe ($p < 10^{-3}$).

Based on these numbers, we estimated that out of all long and high-frequency IS from Neanderthals to modern humans, 15% to 32% (54 of 171) in East Asians and 12% to 25% (27 of 105) in Europeans have been positively selected in response to viruses. In total there are 171 and 105 long and high-frequency IS overlapping genes in East Asians and Europeans, respectively. In East Asians, a total of 1,702 VIPs matched three or more control non-VIPs in the bootstrap test. These 1,702 VIPs overlap the 36 IS (versus 11 expected) used to measure enrichments (Figure 3; STAR Methods), leaving us with ~25 adaptive IS. Additional 42 IS overlapping VIPs were not used because the VIPs matched with fewer than three control

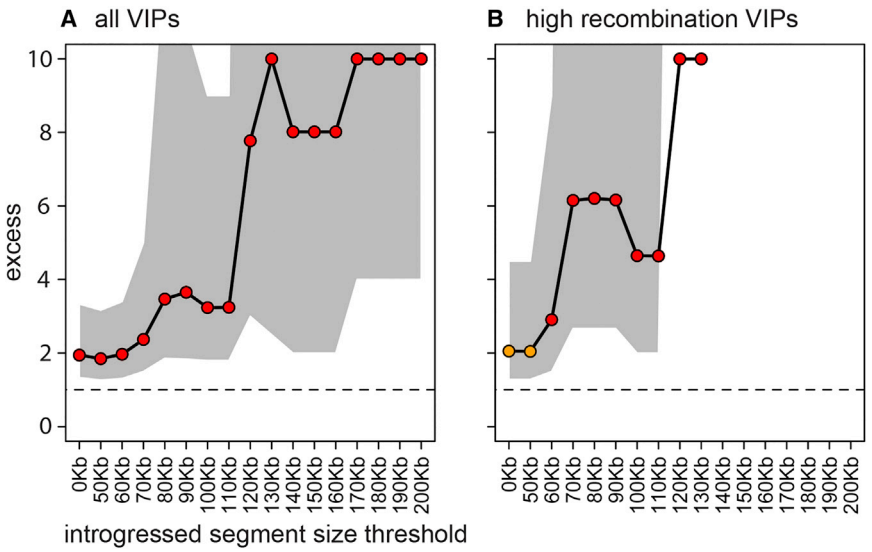



**Figure 4. Excess of Introgression from Modern Humans to Neanderthals at VIPs**

Legend as in Figure 3.

(A) All VIPs.

(B) High recombination VIPs.

See also Figures S5 and S6.

non-VIPs in the bootstrap test (STAR Methods). If we assume that the same proportion was adaptive among the unmatched VIPs, we obtain a total of 54.17 (25 of 36 matched and ~29 of 42 unmatched) positively selected IS, or 32% of all the 171 long, high-frequency IS in East Asians. Using the same extrapolation, we estimated that a total of ~27 or ~25% of all the 105 long, high-frequency IS in Europeans were positively selected in response to viruses.

We could also use these enrichments to estimate false discovery rates (FDR) of adaptive introgression for individual VIPs. VIPs with FDR below 50% are listed in Table S5. Interestingly, several previously published candidate VIP loci for adaptive introgression have low FDR, including the *OAS* gene cluster (FDR = 0.22 in Europe) (Mendez et al., 2013) or the *TLR1/6/10* gene cluster (FDR = 0.17 in Europe) (Dannemann et al., 2016).

## VIPs Are Enriched in Introgressed Segments from Modern Humans to Altai Neanderthals

We next tested for an excess of introgressions from modern humans to Neanderthals, using the data on introgressed genomic regions in a single Altai Neanderthal individual (Kuhlwilm et al., 2016). Because adaptive IS are expected to be longer than neutral ones, we estimated the excess of segments of modern human ancestry in the single Altai Neanderthal individual genome at VIPs as a function of their size. We found a large excess of long segments of modern human ancestry at VIPs (Figure 4A). Furthermore, as predicted, the excess is more pronounced in high-recombination regions of the genome (Figure 4B). We confirmed that this excess was also detected using only high-quality LT-VIPs (Figure S5).

## Identifying Ancient Viruses Responsible for Adaptive Introgression

We next asked if it is possible to identify which ancient viruses might be responsible for the observed enrichments. While such an analysis in the direction from modern humans to Neanderthals is severely underpowered with only 19 VIPs found in IS over 100 kb in the Altai Neanderthal, the number is much larger in modern humans with 152 VIPs found in long (≥100 kb) and frequent (≥15%) Neanderthal IS.

We used the 20 modern human viruses that interact with ten or more VIPs as proxies for the ancient related viruses that infected humans at the time of interbreeding (Table S1). These 20 viruses are evenly distributed between RNA viruses (2,684 VIPs) and DNA viruses (2,547 VIPs) (Table S1). Of the 2,684 RNA VIPs, 1,563 interact with only RNA viruses, while out of 2,547 DNA VIPs, 1,426 interact with only DNA viruses.

We first asked if ancient RNA or DNA viruses were more likely to have been involved, with the expectation that RNA viruses should be more likely to drive adaptive introgression because they are more likely to jump from one species to another (Geoghegan et al., 2017; Kreuder Johnson et al., 2015). In order to determine whether introgression was skewed toward either RNA or DNA viruses, we used the bootstrap test to compare the number of IS at VIPs that interact with only one RNA virus with the number of IS at VIPs that interact with only one DNA virus and are located far from any RNA VIP (≥500 kb) (STAR Methods).

We did not detect any significant skew in favor of RNA-virus VIPs in East Asia (Figure 5A). By contrast, in Europe, we detected a strong bias of RNA-virus VIPs in long, high-frequency IS (Figure 5A). This pattern was more pronounced for introgression in the regions of high recombination (Figure 5B). The enrichment of Neanderthal ancestry at RNA VIPs became even more pronounced (Figures S6A and S6B) when we repeated the comparison after excluding genes known to interact with bacteria, Plasmodium (Ebel et al., 2017), and immune genes annotated as such by the Gene Ontology database (The Gene Ontology Consortium, 2017). Thus, other pathogens appear unable to explain the signal at RNA VIPs. The enrichment was also more pronounced when using only adaptive IS (Gittelman et al., 2016; Jagoda et al., 2017; Racimo et al., 2017) (Figures S6C and S6D). Furthermore, the slightly stronger background selection at RNA VIPs than at control DNA VIPs both in East Asia and Europe (7% stronger in both cases, $p < 10^{-3}$) makes the comparison conservative. RNA VIPs also have slightly fewer segregating variants (9% less in Europe, $p < 10^{-3}$) and thus slightly more sites under strong purifying selection than control DNA VIPs, which is again conservative. The enrichment at RNA VIPs was further confirmed using only LT-VIPs (Figures S6E and S6F) or a different recombination map (Figures S6G and S6H).

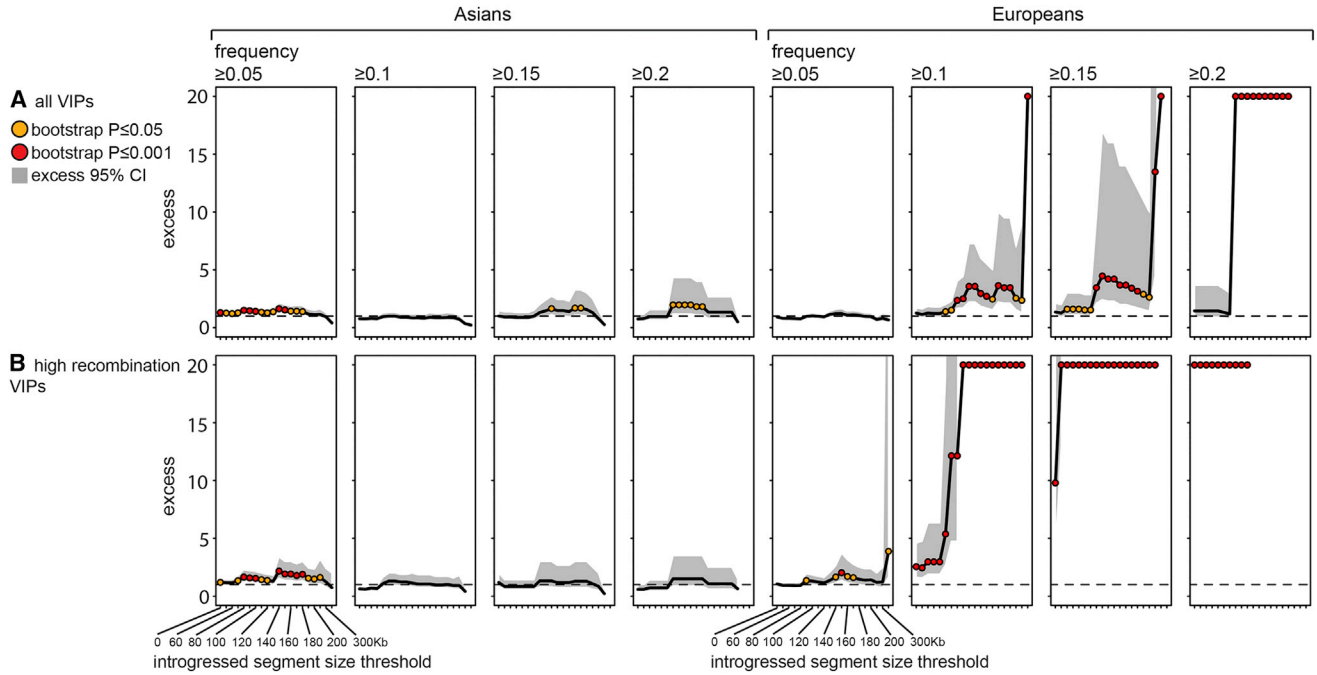



**Figure 5. Excess of Introgression from Neanderthals to Modern Humans at RNA VIPs versus DNA VIPs**

Legend as in Figure 3, except that the y axis represents the excess of introgressions at RNA VIPs versus DNA VIPs rather than VIPs versus non-VIPs. See also Figure S6.

We next tried to identify which families of ancient RNA viruses might explain the observed skew toward RNA VIPs in Europeans. Of the 11 RNA viruses included in this analysis (Table S1), HIV (a lentivirus), influenza A virus (IAV, an orthomyxovirus) and hepatitis C virus (HCV, a flavivirus) have by far the highest numbers of VIPs. It appears that both HIV-only and IAV-only VIPs were each associated with a large excess of high-frequency, long adaptive IS in European modern humans compared to VIPs that interact with only one DNA virus (Figures 6A–6D). The excess was particularly strong for HIV-only and IAV-only VIPs within high-recombination regions (Figure 6B,D). Specifically, we found seven (versus 0.29 expected) high-frequency ($\geq$15%) IS overlapping IAV-only VIPs ($p < 10^{-3}$) and eight (versus 0.83 expected) overlapping HIV-only VIPs ($p < 10^{-3}$). Table S5 lists the specific VIPs found in these IS.

While these results were robust when restricting to HIV-only LT-VIPs (Figures S6I and S6J), we did not detect a significant enrichment at IAV-only LT-VIPs. It is possible that the smaller number of IAV-only LT-VIPs (56 overall versus 195 for HIV and only 15 in high recombination regions) did not provide sufficient power to detect a significant excess of introgression. Indeed, subsampling of HIV-only LT-VIPs to the number of IAV-only LT-VIPs reduced the power enough to eliminate statistical significance (bootstrap test $p > 0.05$ for IAV-only LT VIPs and all ten random subsamples and all introgression lengths and frequencies).

Although we detected no significant enrichment of IS at HCV-only VIPs (Figures 6E and 6F), this might also be an issue of statistical power because HCV has far fewer unique VIPs than HIV and IAV (157 versus 405 and 490, respectively). Indeed, subsampling of HIV-only and IAV-only VIPs to the small number of HCV-only VIPs results in insufficient power to detect any excess due to a small sample size (Figure S6K), leaving open the possibility that HCV-like viruses might also have been involved.

## VIPs and Specific Host Functions

The enrichments of VIPs in IS represent statistical associations, and more evidence is required to demonstrate causality. One particular worry is that host biological functions that are significantly enriched among VIPs might also be enriched in adaptive IS independently of their interactions with viruses. In such a case, the observed enrichment of adaptive introgression at VIPs would not be due to interactions with viruses but would be due, instead, to the uneven representation of diverse host functions between VIPs and non-VIPs.

Consider a hypothetical example in which genes that have the biological Gene Ontology (GO) function "cell cycle" tend to be enriched in IS independently of their interactions with viruses. If the genes with this function are also enriched among VIPs, this could by itself lead to the enrichment of VIPs in the IS. We test this possibility by assessing if the biological functions over-represented among VIPs ($\geq$1.5 enrichment, permutation test $p \leq 0.05$) are enriched within IS overall or within particularly long and/or frequent IS (Figure 7). If they are not—as schematically shown in Figure 7A—then this is an unlikely explanation for our results. If they are—as schematically shown in Figure 7B—then this would be worrisome and would require us to carry out additional analyses to control for this bias. Figures 7C–7F show no evidence that biological functions enriched in

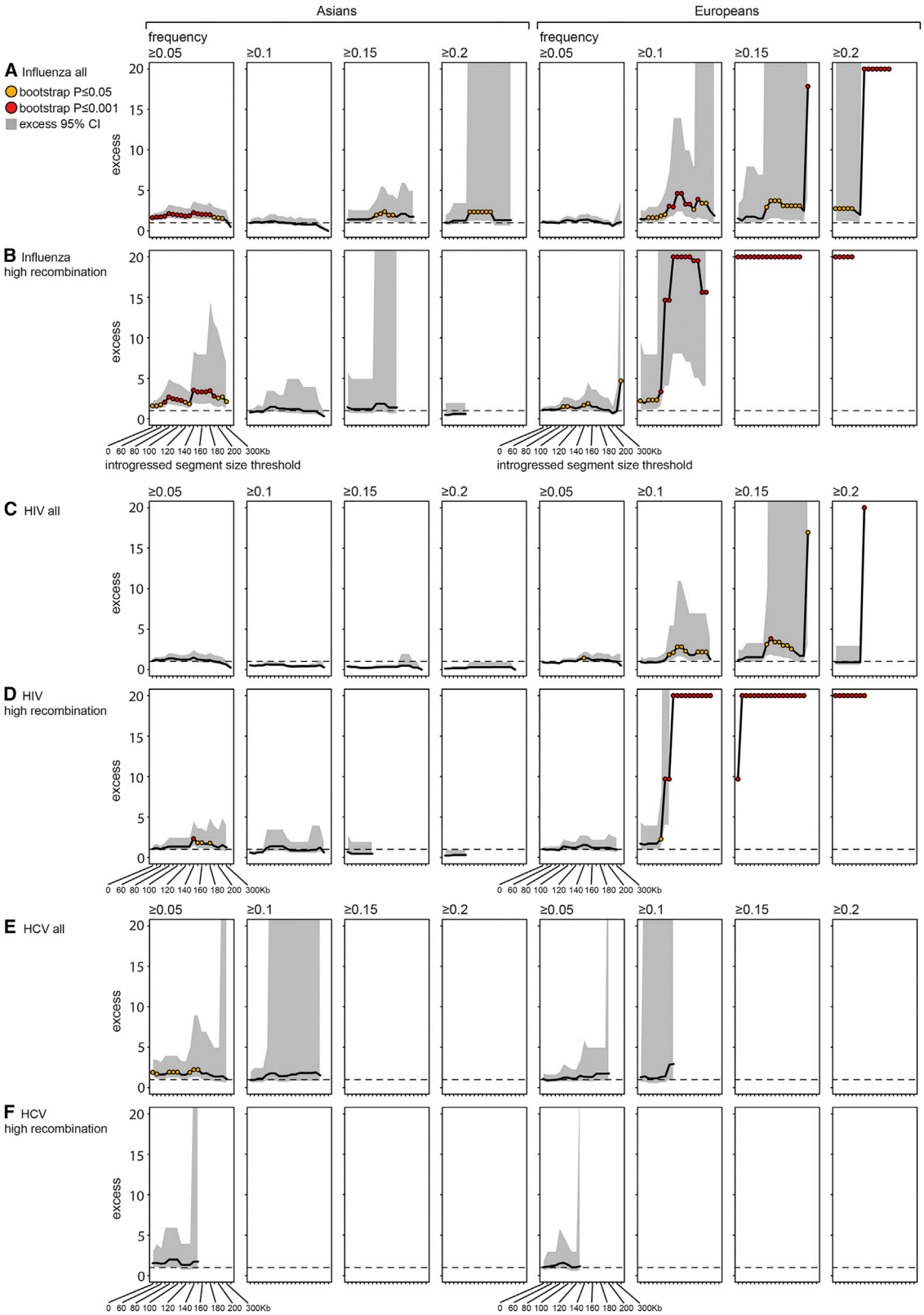



**Figure 6. Excess of Introgression from Neanderthals to Modern Humans at IAV-Only, HIV-Only, and HCV-Only VIPs**
Legend as in Figure 3.

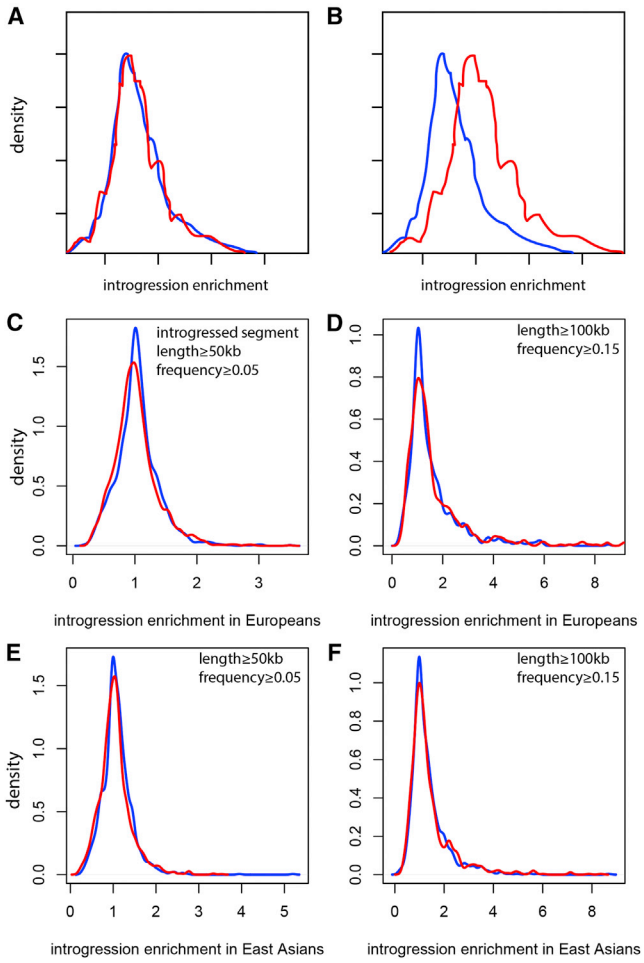



**Figure 7. Enrichments in Adaptive Introgression from Neanderthals to Modern Humans among Gene Ontology Annotations Over-represented or Not Over-represented in VIPs**

We compared the enrichment in adaptive introgression of specific GO annotations. We estimated how much each GO function was enriched in introgressed segments by comparing the number of genes of the GO function within introgressed segments compared to the number of genes expected by chance (STAR Methods). The enrichment is then the ratio of the observed number of genes for a given function within introgressed segments divided by the average random expectation. All the enrichments were measured excluding VIPs because we want to specifically estimate the effect of internal host functions independently of interactions with viruses. Red, distributions of enrichments (x axis) across GO functions found at more than 50 VIPs and that are significantly over-represented among VIPs; blue, distributions of enrichments across GO functions found at more than 50 VIPs but not over-represented among VIPs.

(A) Mock example corresponding to the case where GO functions over-represented in VIPs are not enriched in adaptive introgression (the red and blue distributions overlap each other).

(B) Mock example corresponding to the case where GO functions over-represented in VIPs are enriched in adaptive introgression compared to other GO functions (the red distribution is shifted to the right).

(C) Enrichments in introgressed segments longer than 50 kb and at frequencies higher than 5% in Europe.

(D) Enrichments in introgressed segments longer than 100 kb and at frequencies higher than 15% in Europe.

(E) Enrichments in introgressed segments longer than 50 kb and at frequencies higher than 5% in East Asia.

(F) Enrichments in introgressed segments longer than 100 kb and at frequencies higher than 15% in East Asia.

See also Table S6.

VIPs show any biased signals (Wilcoxon rank sum test $p > 0.05$ in all cases) in their presence in IS.

## Crucial Immune and Proviral VIPs Are Over-represented in Introgressed Segments

Assuming that viruses were responsible for the observed enrichments, we then asked whether VIPs in IS tend to possess any particular functions. We used a permutation test to estimate the over-representation of GO annotations of biological functions at VIPs inside IS compared to VIPs outside IS (STAR Methods). We used IS of at least the length and the frequency where we first started observing highly significant (bootstrap test $p < 10^{-3}$ both in Europe and East Asia) enrichments in IS at VIPs (longer than 50 kb and at frequencies higher than 5%; Figure 3).

In line with a possible causal role of viruses, multiple functions related to immune response or to crucial steps of the viral replication cycle were significantly over-represented among VIPs within IS in Europe, although not in East Asia (Table S6). This pattern was particularly pronounced when using only the manually curated LT-VIPs (Table S6), which may reflect the fact that LT-VIPs are better annotated. For example, the GO annotation "immune effector process" was one of the most strongly over-represented in Europe (all VIPs: 57 VIPs versus 34 expected by chance, permutation test $p = 5 \times 10^{-4}$; LT-VIPs: 49 versus 27.5 expected by chance, $p = 10^{-4}$). Importantly, the "immune effector process" function was not over-represented among non-VIPs within IS (permutation test $p > 0.05$ for all GO annotations in Table S6).

Interestingly, the GO function for a crucial early step of infection, "virion attachment to host cell," was among the most strongly over-represented GO functions among all VIPs, both in Europe (5 VIPs instead of 0.8 expected by chance, permutation test $p < 10^{-7}$), and in East Asia (4 VIPs versus 0.75 expected by chance, $p = 5.10^{-3}$). All the virion attachment VIPs are LT-VIPs. Of the five virion attachment VIPs, four were shown to be involved in RNA virus attachment (ICAM1, CD209/DCSIGN, HSP90AB1, and CLEC4M) and one in DNA virus attachment (PVRL2). Interestingly, of the four virion attachment VIPs interacting with RNA viruses, three interact with HCV (CD209/DCSIGN, HSP90AB1, and CLEC4M; HCV VIPs: 3 versus 0.4 expected virion attachment VIPs, $p = 7.10^{-4}$).

Another GO annotation category that represents an important step in the viral replication cycle, "viral genome replication," was also strongly over-represented particularly at LT-VIPs within IS in Europe (all VIPs: 17 VIPs versus 7.8 expected, $p = 0.029$ and LT-VIPs: 17 LT-VIPs versus 6.9 expected, $p = 1.6 \times 10^{-3}$). Intriguingly, a large number of these "viral genome replication" VIPs were again VIPs that interact with HCV as in the case for virion attachment (LTF, EIF2AK2/PKR, ATG5, MAVS, CD209/DCSIGN, CLEC4M, VAPB; all HCV VIPs: 7 versus 1.7 viral genome replication VIPs, $p = 0.009$;

HCV LT-VIPs: 7 viral genome replication VIPs versus 1.5 expected, $p = 10^{-4}$).

The representation of RNA and DNA viruses within this over-represented function therefore matches our previous finding of a stronger effect of RNA viruses in Europe and strongly suggests a causal role of viruses.

## DISCUSSION

Here, we presented evidence that a substantial proportion of IS from Neanderthals to modern humans and vice versa are strongly enriched for proteins interacting with viruses. We further detected a particularly strong signal for VIPs interacting with RNA viruses in Europeans. The more comprehensive annotations of host-virus interactions we used as well as the controls for multiple genomic factors explain why such signals were not noticed in previous functional analyses of IS (STAR Methods).

Altogether, our results suggest that adaptive introgression in response to viruses might have been more prevalent than previously known based on the small number of published examples (Abi-Rached et al., 2011; Dannemann et al., 2016; Mendez et al., 2012, 2013; Nedelec et al., 2016; Quach et al., 2016; Sams et al., 2016). Indeed, we estimate that out of all long (≥100 kb) and high-frequency (≥15%) IS from Neanderthals to modern humans, 32% (54 of 171) in Asians and 25% (27 of 105) in Europeans might have been positively selected in response to viruses.

Note that we specifically tested the model of positive directional selection of IS soon after interbreeding which is expected to drive long IS of Neanderthal ancestry to high frequency into the population. However, we do not exclude the possibility that the IS were subsequently maintained by balancing selection or frequency-dependent selection of some kind. This is an intriguing possibility especially if the increased genetic variability introduced by introgression was advantageous in variable environments.

Overall, these results provide preliminary support for the ''poison-antidote'' hypothesis under which the interactions between modern humans and Neanderthals exposed each species to novel viruses while gene flow between the species afforded a measure of resistance by allowing VIPs that were already adapted to the presence of specific viruses in the donor species to cross species boundaries and provide adaptive function in the recipient species.

In this respect, the finding that the greater enrichment of Neanderthal ancestry at RNA VIPs is restricted to Europe is particularly interesting. Several authors have concluded that multiple pulses of interbreeding occurred between Neanderthals and modern humans, with at least one pulse before the split of Asian and European modern populations and multiple independent pulses after the split of these populations (Kim and Lohmueller, 2015; Vernot and Akey, 2015; Villanea and Schraiber, 2018). The strong enrichment at RNA VIPs in Europe but not in East Asia then suggests that this difference arose due to the pulses of interbreeding after the split between Asians and Europeans. Whether adaptive introgression at DNA VIPs in East Asians compensated for a previous bias toward RNA VIPs still visible in Europeans, or adaptive introgression at RNA VIPs occurred specifically after an independent pulse of interbreeding in Europe, remains an open question.

Despite the strong statistical signals, we believe that the evidence gathered in favor of the poison-antidote scenario is preliminary. Indeed, although the enrichments we describe are rigorously defined, they represent only statistical associations. We believe that more functional work will be required to establish the causal impact of the virus-host interactions on the detected patterns of adaptive introgression. Intriguingly, such evidence is beginning to accumulate. Quach et al. (2016) recently found that regulatory variants that affect host gene expression during IAV infection are particularly strongly enriched in Neanderthal ancestry in Europeans. Furthermore, Sams et al. (2016) recently found that flaviviruses—a class of RNA viruses that include HCV, Dengue, and other viruses—might have been driving selection at the adaptively introduced Neanderthal haplotype at the *OAS1* locus.

The analysis presented here opens the door for more functional studies of this kind. Our own cursory look at the functional data revealed some tantalizing patterns. For instance, several of the RNA virus VIPs that have a low FDR for adaptive introgression are known to affect specific steps of the viral replication cycle, and in some cases, introgression increased the frequency of specific functional variants that plausibly confer increased resistance against viruses. For example, the IAV VIP PPIE (also known as cyclophilin E) has been shown to inhibit the formation of the viral ribonucleoprotein complex required for IAV RNA replication (Wang et al., 2011b). Introgression at the PPIE locus in Europeans has increased the frequency of allelic variants of expression quantitative trait loci (eQTL) SNPs that are associated with very high expression of PPIE across many tissues, including the lungs (GTEx Consortium, 2015). It is possible that such an increased expression results in greater inhibition of the viral ribonucleoprotein complex and thus increased viral resistance.

Another noteworthy example is Toll-like receptor 2 TLR2, a HIV VIP where introgression increased the frequency of alleles at multiple linked eQTL SNPs associated with higher *TLR2* expression in many tissues (GTEx Consortium, 2015). HIV protein ENV has been shown to bind and inhibit TLR2 activity (Reuven et al., 2014), suggesting that increased TLR2 expression might have at least partially prevented such inhibition for related lentiviruses.

Furthermore, it will be interesting to study whether the presence of Neanderthal ancestry at VIPs still leads to variable susceptibility to modern viruses in modern humans. Interestingly, the C-type lectin receptor CD209/DCSIGN, one of the five virion attachment HCV VIPs found to be highly enriched in IS in Europe (Crucial Immune and Proviral VIPs Are Over-Represented in Introgressed Segments, in Results), interacts with HCV and Dengue virus. Introgression from Neanderthals at this locus has affected the frequency of alleles at a well-known variant (rs4804803) within the promoter of DCSIGN with a documented effect on *CD209/DCSIGN* expression and on HCV and dengue virus infection severity (Ryan et al., 2010; Wang et al., 2011a).

While these scenarios are clearly speculative, they show the power of our enrichment approach to systematically formulate plausible and testable hypotheses for specific adaptive introgression candidates. Many of the identified genes are identified at low enough FDR that it is sensible to carry out functional studies of the introgressed variants.

Finally, these results suggest that the genomes of humans and other species contain signatures of past arms races with diverse viruses and other pathogens, making it possible to use host genomic signatures to study ancient interactions with ever present and ever shifting viral and other pathogens. In this respect, it is worth noting that even though we focused on Neanderthal ancestry, we anticipate that it should also be possible to study the impact of ancient epidemics on introgression from Denisovans to modern humans, especially in populations such as Melanesians with a larger percentage of Denisovan ancestry (Vernot et al., 2016). The results of such studies should provide important insights into the dynamics of past, present, and future epidemics.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Annotation of VIPs
  - Introgressed segments from Neanderthals to modern humans
  - Introgressed segments from modern humans to Altai Neanderthals
  - Genomic factors
  - Identifying important genomic factors
  - Bootstrap test
  - Permutations with a target average
  - Gene Ontology Permutations analysis
  - RNA versus DNA VIP analysis
- DATA AND SOFTWARE AVAILABILITY

### SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.08.034.

### ACKNOWLEDGMENTS

We wish to thank Rajiv McCoy, Jamie Blundell, Kerry Geiler-Samerotte, Sandeep Venkataram, Sharon Greenblum, Emily Ebel, and other current and former members of the Petrov Lab for comments on the manuscript. This work is funded by NIH (1RO1GM10036601 and R35GM118165 to D.P.).

### AUTHOR CONTRIBUTIONS

D.E. and D.A.P. conceived and designed the analyses. D.E. performed the analyses. D.E. and D.A.P. wrote the paper.

### DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 2, 2018
Revised: July 4, 2018
Accepted: August 16, 2018
Published: October 4, 2018

### REFERENCES

Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., et al. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. Science *334*, 89–94.

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Dannemann, M., Andrés, A.M., and Kelso, J. (2016). Introgression of Neandertal- and Denisovan-like haplotypes contributes to adaptive variation in human Toll-like receptors. Am. J. Hum. Genet. *98*, 22–33.

Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.L., Patin, E., and Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. Am. J. Hum. Genet. *98*, 5–21.

Duret, L., and Arndt, P.F. (2008). The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. *4*, e1000071.

Ebel, E.R., Telis, N., Venkataram, S., Petrov, D.A., and Enard, D. (2017). High rate of adaptation of mammalian proteins that interact with Plasmodium and related parasites. PLoS Genet. *13*, e1007023.

Enard, D., Cai, L., Gwennap, C., and Petrov, D.A. (2016). Viruses are a dominant driver of protein adaptation in mammals. eLife *5*, e12469.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X.J., Yip, K.Y., Khurana, E., and Gerstein, M. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. *15*, 480.

Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., et al. (2015). An early modern human from Romania with a recent Neanderthal ancestor. Nature *524*, 216–219.

Geoghegan, J.L., Duchêne, S., and Holmes, E.C. (2017). Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. PLoS Pathog. *13*, e1006215.

Gittelman, R.M., Schraiber, J.G., Vernot, B., Mikacenic, C., Wurfel, M.M., and Akey, J.M. (2016). Archaic hominin admixture facilitated adaptation to out-of-Africa environments. Curr. Biol. *26*, 3375–3382.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Guirimand, T., Delmotte, S., and Navratil, V. (2015). VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic Acids Res. *43*, D583–D587.

Harris, K., and Nielsen, R. (2016). The genetic cost of Neanderthal introgression. Genetics *203*, 881–891.

Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature *476*, 170–175.

Houldcroft, C.J., and Underdown, S.J. (2016). Neanderthal genomics suggests a pleistocene time frame for the first epidemiologic transition. Am. J. Phys. Anthropol. *160*, 379–388.

Jagoda, E., Lawson, D.J., Wall, J.D., Lambert, D., Muller, C., Westaway, M., Leavesley, M., Capellini, T.D., Mirazón Lahr, M., Gerbault, P., et al. (2017). Disentangling immediate adaptive introgression from selection on standing introgressed variation in humans. Mol. Biol. Evol. Published online December 6, 2017. https://doi.org/10.1093/molbev/msx314.

Juric, I., Aeschbacher, S., and Coop, G. (2016). The strength of selection against Neanderthal introgression. PLoS Genet. *12*, e1006340.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., et al.; 1000 Genomes Project Consortium (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. Science *342*, 1235587.

Kim, B.Y., and Lohmueller, K.E. (2015). Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations. Am. J. Hum. Genet. *96*, 454–461.

Kim, B.Y., Huber, C.D., and Lohmueller, K.E. (2017). Deleterious variation mimics signatures of genomic incompatibility and adaptive introgression. bioRxiv. https://doi.org/10.1101/221705.

Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.

Kreuder Johnson, C., Hitchens, P.L., Smiley Evans, T., Goldstein, T., Thomas, K., Clements, A., Joly, D.O., Wolfe, N.D., Daszak, P., Karesh, W.B., and Mazet, J.K. (2015). Spillover and pandemic properties of zoonotic viruses with high host plasticity. Sci. Rep. *5*, 14830.

Kuhlwilm, M., Gronau, I., Hubisz, M.J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H.A., Lalueza-Fox, C., de la Rasilla, M., et al. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. Nature *530*, 429–433.

Luisi, P., Alvarez-Ponce, D., Pybus, M., Fares, M.A., Bertranpetit, J., and Laayouni, H. (2015). Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. Genome Biol. Evol. *7*, 1141–1154.

McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. *5*, e1000471.

Mendez, F.L., Watkins, J.C., and Hammer, M.F. (2012). A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. Am. J. Hum. Genet. *91*, 265–274.

Mendez, F.L., Watkins, J.C., and Hammer, M.F. (2013). Neandertal origin of genetic variation at the cluster of OAS immunity genes. Mol. Biol. Evol. *30*, 798–801.

Nedelec, Y., Sanz, J., Baharian, G., Szpiech, Z.A., Pacis, A., Dumaine, A., Grenier, J.C., Freiman, A., Sams, A.J., Hebert, S., et al. (2016). Genetic ancestry and natural selection drive population differences in immune responses to pathogens. Cell *167*, 657–669.

Pääbo, S. (2015). The diverse origins of the human gene pool. Nat. Rev. Genet. *16*, 313–314.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

Quach, H., Rotival, M., Pothlichet, J., Loh, Y.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic adaptation and Neandertal admixture shaped the immune system of human populations. Cell *167*, 643–656.

Racimo, F., Sankararaman, S., Nielsen, R., and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. Nat. Rev. Genet. *16*, 359–371.

Racimo, F., Marnetto, D., and Huerta-Sánchez, E. (2017). Signatures of archaic adaptive introgression in present-day human populations. Mol. Biol. Evol. *34*, 296–317.

Reuven, E.M., Ali, M., Rotem, E., Schwarzer, R., Gramatica, A., Futerman, A.H., and Shai, Y. (2014). The HIV-1 envelope transmembrane domain binds TLR2 through a distinct dimerization motif and inhibits TLR2-mediated responses. PLoS Pathog. *10*, e1004248.

Ryan, E.J., Dring, M., Ryan, C.M., McNulty, C., Stevenson, N.J., Lawless, M.W., Crowe, J., Nolan, N., Hegarty, J.E., and O'Farrelly, C. (2010). Variant in CD209 promoter is associated with severity of liver disease in chronic hepatitis C virus infection. Hum. Immunol. *71*, 829–832.

Sams, A.J., Dumaine, A., Nédélec, Y., Yotova, V., Alfieri, C., Tanner, J.E., Messer, P.W., and Barreiro, L.B. (2016). Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. Genome Biol. *17*, 246.

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. PLoS Genet. *8*, e1002947.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354–357.

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

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al. (2011). The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. *39*, D698–D704.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

The Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. *45* (D1), D331–D338.

Vernot, B., and Akey, J.M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. Science *343*, 1017–1021.

Vernot, B., and Akey, J.M. (2015). Complex history of admixture between modern humans and Neandertals. Am. J. Hum. Genet. *96*, 448–453.

Vernot, B., Tucci, S., Kelso, J., Schraiber, J.G., Wolf, A.B., Gittelman, R.M., Dannemann, M., Grote, S., McCoy, R.C., Norton, H., et al. (2016). Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science *352*, 235–239.

Villanea, F.A., and Schraiber, J.G. (2018). Spectrum of Neandertal introgression across modern-day humans indicates multiple episodes of human-Neandertal interbreeding. bioRxiv. https://doi.org/10.1101/343087.

Wang, L., Chen, R.F., Liu, J.W., Lee, I.K., Lee, C.P., Kuo, H.C., Huang, S.K., and Yang, K.D. (2011a). DC-SIGN (CD209) Promoter -336 A/G polymorphism is associated with dengue hemorrhagic fever and correlated to DC-SIGN expression and immune augmentation. PLoS Negl. Trop. Dis. *5*, e934.

Wang, Z., Liu, X., Zhao, Z., Xu, C., Zhang, K., Chen, C., Sun, L., Gao, G.F., Ye, X., and Liu, W. (2011b). Cyclophilin E functions as a negative regulator to influenza virus replication by impairing the formation of the viral ribonucleoprotein complex. PLoS ONE *6*, e22625.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Res. *44* (D1), D710–D716.

# STAR★METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited Data | | |
| Analysis scripts | This paper | https://github.com/DavidPierreEnard/Matching_VIPs_nonVIPs |
| Confounding factors table | This paper | https://github.com/DavidPierreEnard/Matching_VIPs_nonVIPs |
| Software and Algorithms | | |
| Analysis scripts | This paper | https://github.com/DavidPierreEnard/Matching_VIPs_nonVIPs |

## CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, David Enard (denard@email.arizona.edu).

## METHOD DETAILS

All the STAR Methods we used are quantifications and statistical analyses. All the details related to these STAR Methods are therefore provided in the following section, QUANTIFICATION AND STATISTICAL ANALYSIS.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Annotation of VIPs

We previously manually annotated 1256 VIPs from a set of 9861 human proteins with orthologs conserved across mammals (Enard et al., 2016). Here, we extended our manual annotation effort to all protein coding genes in the human genome and identified 664 additional VIPs, for a total of 1920 manually curated, high quality VIPs (Table S1). The 664 additional VIPs were all identified with low-throughput STAR Methods and extracted from the virology literature as previously described (Enard et al., 2016). In addition to the 1920 low-throughput VIPs, we also used 2614 other VIPs identified for viruses infecting humans by high throughput STAR Methods and annotated in the VirHostNet2.0 database (Guirimand et al., 2015) or identified in at least one of 14 different recent studies not listed in VirHostNet2.0 (Table S1). We excluded VIPs only identified by yeast two-hybrid because of notoriously high rates of false positives and negatives. The 4534 resulting VIPs are all listed in Table S1 together with their respective viruses. Note that LT-VIPs for specific viruses can also be high-throughput VIPs for other viruses (Table S1). Note that our annotation of VIPs is much more comprehensive than annotations of host-virus interactions provided by Gene Ontology annotations. Indeed we found that only 18% of VIPs are annotated with GO functions related to viruses, defined as GO functions with the words "virus" or "viral" in their name. Together with the fact that Sankararaman et al. (2014) did not control for purifying selection and other confounding factors in their functional analysis and did not use controls far enough from VIPs, this could potentially explain why the enrichments in introgression at VIPs were not previously noticed by these authors even though they conducted a GO enrichment analysis. In Figures S4L–S4Q, we show in particular that not controlling for confounding factors and not choosing control non-VIPs far from VIPs largely eliminates the signal of enrichment at VIPs. The same naive approach not controlling for confounding factors and not choosing control DNA VIPs far from RNA VIPs however still detects the enrichment of Neanderthal ancestry at RNA VIPs in Europe (Figures S6L and S6M).

### Introgressed segments from Neanderthals to modern humans

We used the segments of Neanderthal ancestry in both Asian and European modern humans that were identified and kindly provided by Sankararaman et al. (2014). Sankararaman et al. (2014) estimated for each SNP in East Asian or European populations the population-wide probability that an allele was inherited from Neanderthals (Figure S1A; posterior probability on the y axis, blue curve on the graph). For each SNP either in European or East Asian populations, the CRF approach provides a posterior probability that any given allele at the SNP site was inherited from Neanderthals. To estimate the probability that a specific allele comes from Neanderthals Sankararaman et al., (2014) first define a 100Kb window (Figure S1B). In the example of Figure S1B the 100Kb window contains seven SNPs. The sample studied is made of four individuals for a total of eight phased haplotypes each representing a single chromosome. To estimate the probability that each allele on each different phased haplotype comes from Neanderthals, the CRF uses all other alleles on the same phased haplotype in the 100Kb window. In particular it uses 1) SNP sites where the tested haplotype carries the same allele as Neanderthals but this allele is absent in Africa and 2) SNP sites where the tested haplotype carries the same allele

as African populations but this allele is absent in the Neanderthal diploid individual genome. Compared to a Hidden Markov Model, the emission probability of the CRF at a given SNP site then depends not only on the allelic state at this SNP but also on the allelic states of all the other informative SNPs in the 100Kb window weighted by their genetic distance from the tested SNP. In summary, the CRF incorporates the surrounding haplotype structure to estimate the probability of Neanderthal ancestry separately for each allele in the 100Kb window.

Then, by summing all the weighted probabilities for every allele at a specific SNP site it is possible to get the overall probability that any allele at a particular SNP was introgressed from Neanderthals (Figure S1B). This probability can then be used as a proxy for the frequency of Neanderthal ancestry at a particular SNP site.

We then defined an introgressed segment (blue rectangle in Figure S1A) at a frequency higher than a fixed threshold (for example threshold 0.2 in Figure S1A) as an entire region where the posterior probabilities at consecutive SNPs exceed the fixed threshold. To extend the introgressed segment we tolerated that the posterior probability falls transiently below the fixed threshold for no more than ten consecutive SNPs (~5Kb on average) before going back to values higher than the threshold (small dent below 0.2 in Figure S1A). Note that the specific number of consecutive SNPs allowed below the frequency threshold does not affect our results. Indeed, we estimated very similar enrichments in introgression at VIPs when using ten (~5kb) or 100 consecutive SNPs (~50Kb) (Figure S7A).

Note that in addition to posterior probabilities, Sankararaman et al. (2014) provide predictions of high-confidence Neanderthal segments in specific modern human individuals. Indeed, some stretches of phased haplotypes such as haplotype 6 (SNP sites 1, 2, 3, 4 and 5) in Figure S1B can have very high estimated probabilities of Neanderthal ancestry, in which case they are classified by Sankararaman et al. (2014) as "high confidence" segments of Neanderthal ancestry present in specific individuals in the tested population. Importantly the high CRF posterior probabilities we used to infer long IS (> 100kb) at high frequencies (> 15%) where we found the strongest enrichments at VIPs overlap high-confidence individual segments at 99.5% in Europe and 99.7% in Asia, respectively, showing near-perfect agreement between the two types of annotations of IS by the CRF approach. Overall a high proportion of 70% of the IS we used (irrespective of their length or frequency) overlap high confidence Neanderthal segments found in specific individuals by Sankararaman et al. (2014). Importantly the enrichments in IS observed at VIPs when using only those IS that overlap with high confidence segments are indistinguishable from the enrichments observed when using all IS (Figure S7B). This confirms that using the posterior probabilities provided by Sankararaman et al. (2014) to define IS above a fixed frequency threshold is an appropriate approach. The IS with their coordinates, the genes they contain as well as the information of whether or not they overlap with high confidence Neanderthal segments are available as Table S7.

### Introgressed segments from modern humans to Altai Neanderthals

We use the segments of modern human ancestry in the Altai Neanderthal genome provided by Kuhlwilm et al. (2016) as a supplementary table (Table S18) in their manuscript.

### Genomic factors

All analyses were conducted using hg19 genomic coordinates and protein-coding gene annotations from Ensembl version 83 (Yates et al., 2016). Genomic factors included the densities of coding (Yates et al., 2016), conserved (Siepel et al., 2005), and regulatory elements (ENCODE Project Consortium, 2012). For each protein-coding gene in the human genome, these densities were measured within 50 kb windows at the genomic center of each gene (halfway between the most 5′ transcription start and most 3′ transcription stop sites), ensuring that all genes were treated equally irrespective of their genomic structure. To measure coding sequence density (CDS), we used coding sequences annotated in Ensembl version 83 (Yates et al., 2016). The density of conserved elements was the density of segments conserved across mammals identified by PhastCons (Siepel et al., 2005) applied to alignments of 46 mammalian genomes, and available at the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/). The density of regulatory elements was the density of all the Encode DNase I segments cumulated across all ENCODE cell types (ENCODE Project Consortium, 2012), available at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/.

In addition to these densities, we also controlled for various functional aspects of genes such as mRNA expression and the number of protein–protein physical interactions. For the former, we used mRNA expression (measured in RPKM) for Ensembl protein-coding genes across 53 different tissues from GTEx version 6 (GTEx Consortium, 2015), available at https://www.gtexportal.org/home/. For the number of protein–protein interactions (known as 'degree' in the protein–protein interaction network), we used a version of the BioGrid database, curated and made available by Luisi et al. (2015) (Stark et al., 2011).

In addition to these functional factors, GC content is well known to correlate with the long-term recombination rate (Duret and Arndt, 2008). Because recombination rate strongly affects the strength of background selection against IS, we controlled for both GC content and direct estimates of the local recombination rates measured within 200-kb windows centered on genes, as described above. In particular, we used the fine-scale genetic maps measured by Hinch et al. (2011) in African Americans. We further showed that our results are robust to the specific recombination map being used (Figures S4D, S4E, S6G, and S6H). Note also that all analyses in the manuscript were conducted using only genes with a recombination rate greater than 0.0005 cM/Mb to avoid confusion between genes where the recombination rate is null and genes located within gaps in the recombination map (null versus unknown recombination rate).

Finally, to study introgression at VIPs compared to non-VIPs it is crucial to control for the amount of deleterious variants, defined as the amounts of segregating deleterious mutations within different regions of the genome. We used two statistics that are both expected to correlate with the amount of deleterious variants. First, we used Tajima's D (Tajima, 1989), measured using variants from the 1000 Genomes Project (Auton et al., 2015), as an estimator of the excess of rare alleles within 50 kb windows centered on Ensembl version 83 protein-coding genes. Deleterious alleles are expected to segregate at lower frequencies than neutral ones, and although Tajima's D is often used to detect complete selective sweeps, it was initially created to detect an excess of rare deleterious alleles. Because Tajima's D is also sensitive to selective sweeps, we can also control for those sweeps that occurred later in evolution. Indeed, in addition to accounting for deleterious mutations, controlling for Tajima's D is also likely to partially account for the fact that IS from Neanderthals to modern humans may have been eliminated at locations in the genome where adaptive *de novo* mutations that occurred after interbreeding resulted in selective sweeps that reached high frequencies.

As a second statistic also expected to correlate with the amount of deleterious variants, we used the scores of deleteriousness attributed to non-coding variants (which represent the vast majority of variants) from the 1000 Genomes Project (Auton et al., 2015) by the annotation tool FUNSEQ (Fu et al., 2014; Khurana et al., 2013). More specifically, we measure the average deleteriousness score within 50 kb windows centered on Ensembl version 83 protein-coding genes. Tajima's D and the average FUNSEQ score represent two very different ways to estimate the amount of deleterious variants, and are therefore complementary.

Overall 68% of genes are shorter than 50kb, and 80% are shorter than 100kb. So the 50 kb windows we used are sufficient for a majority of the genes in the analysis. For very large genes well over 100kb, it is possible that the values of the factors measured in 50kb windows do not always correlate well with the values of these factors if they had been measured over the whole length of these long genes. To address this limitation we repeated the comparison of VIPs and non-VIPs and the comparison of RNA VIPs and DNA VIPs using only genes shorter than 50kb. The results are largely unaffected (Figures S4J, S4K, S6N, and S6O), thus showing that the 50kb windows used for measuring confounding factors are sufficient.

### Identifying important genomic factors

In order to estimate the effect of viruses on adaptive introgression, it is crucial to first eliminate factors intrinsic to the host that also affect the occurrence of IS along the genome. This can be achieved by comparing VIPs and non-VIPs that are matched for genomic factors that affect the occurrence of IS; measures of such factors are expected to be significantly different inside versus outside IS. For example, IS from Neanderthals occur more frequently in regions of modern human genomes with higher recombination rates (Sankararaman et al., 2014) because more intense background selection eliminated more IS in regions of low recombination. In agreement with these previous findings, we found that recombination is significantly higher inside versus outside IS from Neanderthals to both Asian and European modern humans (Figures S2A and S2B). More specifically, we measured recombination in 200 kb windows centered on genes (see "Genomic factors" in STAR Methods) inside and outside IS. Genes with their genomic center within an IS were considered to be 'inside,' whereas all other genes are considered 'outside.' We then counted the number of genes inside IS and calculated the average recombination rate across these genes. To determine how different this average differs from that of genes outside IS, we randomly sampled the same number of genes outside IS 1,000 times to obtain an empirical null distribution for the average.

The comparison of genes inside versus outside IS can be performed not only for recombination, but also for any possible genomic factor. Figures 2 and S2 show the genomic factors that differ significantly inside versus outside IS from Neanderthals to modern humans, as well as for IS in the other direction. For all factors other than recombination, we did not use a simple permutation test, but instead used a permutation test with a target average (see below) that makes it possible to compare genes inside and outside IS with similar recombination rates.

### Bootstrap test

We created a bootstrap test to compare VIPs and non-VIPs matched for all important host genomic factors that affect the occurrence of IS (Figures 2 and S2). If a given genomic factor had the value *a* at a specific VIP, we looked for all non-VIPs around the value *a* within the range of values from *a*-*a*x to *a*+*a*y, where x and y are always above zero. At this point, we selected for further analysis only those VIPs with more than three matching non-VIPs. For each VIP with at least three matching non-VIPs, we then randomly chose one of the matched non-VIPs as its control. By doing the same for all VIPs, we obtained a control set of non-VIPs with the same important genomic properties, i.e., those properties that differ inside versus outside IS and also between VIPs and non-VIPs. This matching process was a bootstrap because the same non-VIP could serve as the control for several VIPs. By repeating the matching process many times, we could create many random sets of control non-VIPs from which empirical null distributions (i.e., genes with the same genomic properties except for the interactions with viruses) could be estimated for any possible genomic factor, including those that we tried to match between VIPs and non-VIPs. This means we could use the bootstrap test itself to adjust the values of x and y that define the range of a given factor in matched non-VIPs. In practice, we manually adjusted the values of x and y through trial and error for each genomic factor separately until all factors had non-significantly different averages of all the genomic factors included between VIPs and matched non-VIPs (bootstrap test $p > 0.05$ after 200 iterations of the matching process). Table S3 lists all the values of x and y for all the main bootstrap tests performed for this manuscript, together with the number of VIPs passing the minimum requirement of at least three matched non-VIPs, as well as the total number of matched non-VIPs used as controls. Because several genomic factors were correlated with each other, changing x and y for a specific factor often affected the match

between the averages of another genomic factor between VIPs and non-VIPs. This interdependence of several genomic factors made the matching process complicated to automate, and explains the use of manual trials and errors. Once all important genomic factors were properly matched between VIPs and non-VIPs, we ran 5,000 iterations of the matching process to test for an excess of IS at VIPs compared to matched non-VIPs. To measure the excess at VIPs, we counted the number of IS that overlap VIPs (meaning the genomic center at equal distance from the genomic start and end of a VIP gene overlaps an introgressed segment), divided by the number of IS that overlap matched, control non-VIPs. For those IS that contained one or more non-VIPs that were matched with multiple VIPs, we first randomly chose one non-VIP to represent the whole segment (if there were several of them), and then added to the overall count of segments overlapping non-VIPs the number of times the chosen non-VIP was matched with distinct VIPs. In our case, counting the number of IS overlapping VIPs or non-VIPs instead of counting the number of VIPs and non-VIPs within IS was conservative. Indeed, VIPs retained in the bootstrap test tended to be clustered together more closely than the matched non-VIPs (Table S4).

Because the IS could be very large, and therefore included both VIPs and potential non-VIP controls, we only matched VIPs with non-VIPs that were at least 500 kb away from any VIP, and in parallel only counted IS from Neanderthals to modern humans that were smaller than 500 kb. We chose a minimal distance of 500 kb between VIPs and control non-VIPs as a good compromise between having a wide enough representation of sizes of IS, and keeping a sufficient number of non-VIPs that could still be used as controls. For IS from modern humans to Neanderthals, the largest introgressed segment found in the Neanderthal Altai genome was 310 kb, so as potential controls we used all non-VIPs at least 310 kb away from any VIP.

### Permutations with a target average

Genes inside and outside IS have very different recombination rates, with genes inside IS having much higher recombination rates than genes outside (Figures S2A and S2B). This is because purifying selection eliminated more IS in low recombination regions. Many genomic factors such as coding or regulatory density are well known to correlate with the rate of recombination. To avoid confusing the effect of a specific genomic factor on the occurrence of IS with the correlated effect of recombination, we compared genomic factors inside and outside IS using a permutation test with a target average that was previously introduced in Enard et al. (2016). In brief, the permutation test with a target average makes it possible to build random control sets of genes outside IS with the same overall average recombination rate as genes inside IS. This way we could isolate the specific effect of a genomic factor while eliminating the potential confounding effect of recombination. To test different genomic factors and get empirical observed p values (Figure S2), we built 1,000 random control sets of genes outside IS and compared them with genes inside IS. Genes inside IS are genes with their genomic center –the coordinate half way between a gene start and end's sites—overlapping an introgressed segment. Genes outside IS are genes with their genomic center outside any introgressed segment. We repeated the building of 1,000 random control sets of for each frequency threshold and for each length threshold in Figure S2. For each threshold, the genes inside IS are the genes inside IS with frequency or length beyond the fixed threshold, but genes outside IS are genes outside all IS regardless of frequency or length.

### Gene Ontology Permutations analysis

In order to test the over-representation of IS within specific GO functions, we shuffled GO annotations between genes. However the shuffling was not a simple random shuffling. We first ordered genes based on their order on chromosomes and then separated the ordered genes into ten groups of equal size each containing only neighboring genes and finally randomly shuffled the order of the ten groups. This shuffling preserves the clustering structure of GO annotations between neighboring genes that is expected to affect the variance of the null distribution in the over-representation test.

### RNA versus DNA VIP analysis

We used VIPs that interact with only one RNA virus, and VIPs that interact with only one DNA virus, for two reasons. First, by comparing VIPs that interact with the same number of viruses (one in this case), we avoid confusing an effect of the type of virus (RNA versus DNA) with an effect of the number of viruses with which VIPs interact. Second, VIPs already known to interact with multiple viruses might be more likely to interact with as-yet-unknown viruses than VIPs known to interact with only one virus. Thus, VIPs currently only known to interact with multiple RNA viruses may nonetheless be more likely to be involved in as-yet-unknown interactions with DNA viruses and vice versa. Consistent with this, the VIPs in our dataset that interact with two or more RNA viruses are more likely to also interact with at least one DNA virus than VIPs that interact with only one RNA virus (62.6% versus 31.8%, respectively, proportion comparison test $p < 10^{-16}$). Reciprocally, VIPs that interact with two or more DNA viruses are more likely to also interact with at least one RNA virus than VIPs that interact with only one DNA virus (64.8% versus 35.4%, $p < 10^{-16}$).

## DATA AND SOFTWARE AVAILABILITY

The scripts required to carry out enrichment analyses are available at https://github.com/DavidPierreEnard/Matching_VIPs_nonVIPs together with the necessary explanations.

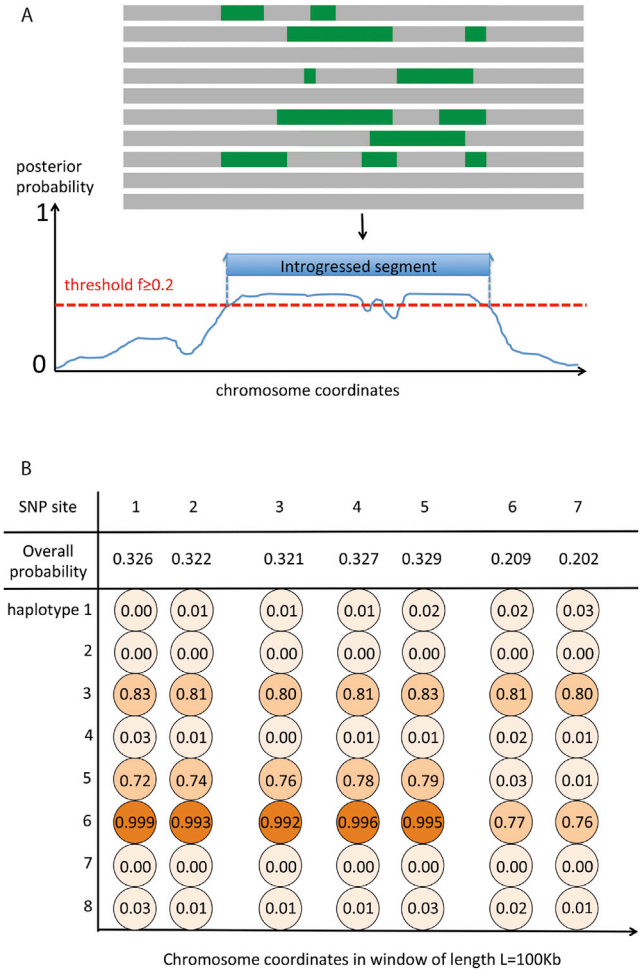



**Figure S1. Definition of Introgressed Segments, Related to Figure 1**

(A) Green areas depict regions that were inherited from Neanderthals in different individuals from the same population. The population-wide posterior probability of an allele being inherited from Neanderthals (y axis) is depicted by the blue curve on the graph. The introgressed segment in the figure (blue rectangle) is defined as a genomic region where the posterior probabilities at SNPs exceed the fixed threshold of 0.2. We tolerated that the posterior probability falls transiently below the fixed threshold for no more than ten consecutive SNPs (small dent below 0.2 in the figure).

(B) Allele-specific estimates of probabilities of Neanderthal ancestry in a genomic window. Light orange. Low probability. Orange: moderate probability. Dark orange: high probability. Each round represents a specific allele and the corresponding probability of Neanderthal ancestry estimated by the CRF.

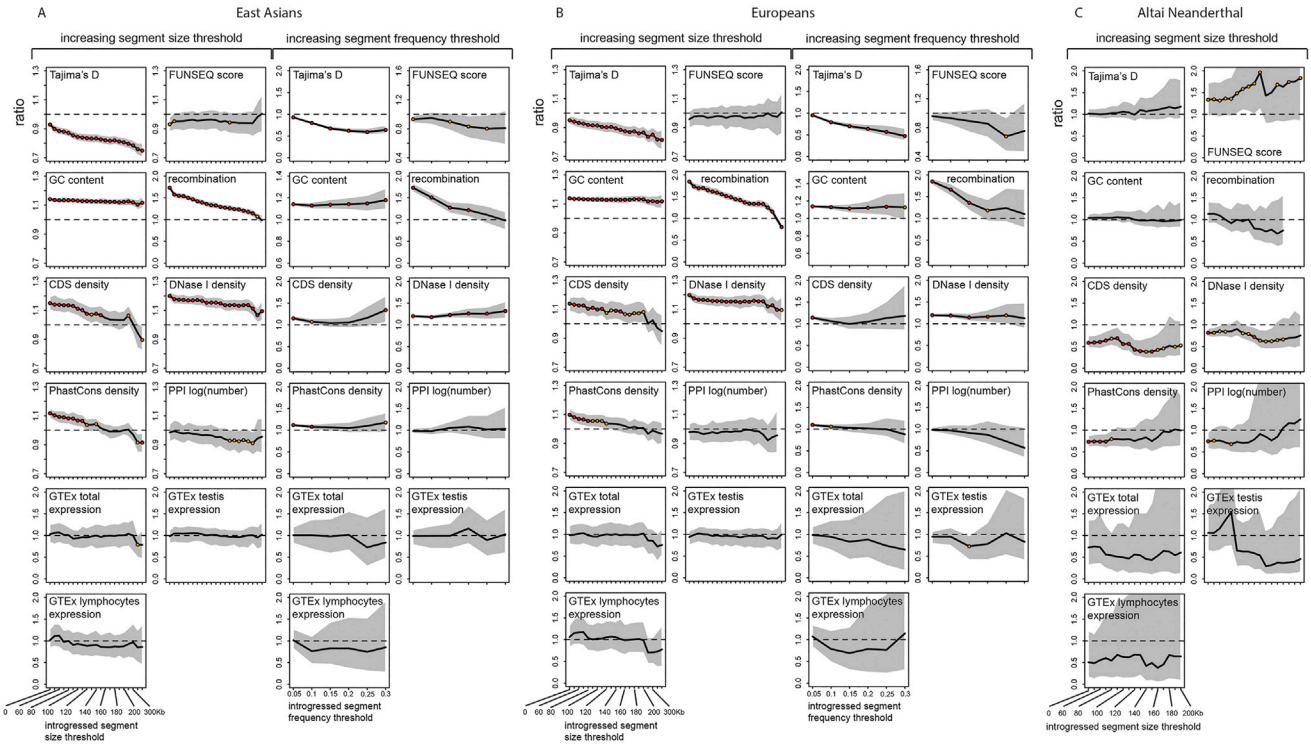



**Figure S2. Genomic Factors inside and outside Introgressed Segments, Related to Figure 2**

(A) In East Asians.

(B) In Europeans.

(C) In the Altai Neanderthal individual genome. The y axis represents the ratio of the average of the statistic for genes inside introgressed segments to the average of the statistic for control genes outside introgressed segments. Control genes outside introgressed segments were matched with those inside introgressed segments for recombination using permutations with a target average ($10^4$ iterations, STAR Methods). The x axis represents either increasing introgressed segment size threshold or increasing introgressed segment frequency threshold. Ratios greater than 1 (dashed lines) indicate that the tested statistic was inside than outside introgressed segments. Black line: observed ratio. Grey area: 95% confidence interval for the ratio. Orange dots: permutation test $p < 0.05$. Red dots: $p < 0.001$. In addition to the total GTEx expression, we also specifically controlled for testis and lymphocyte expression because these tissues often experience elevated rates of adaptation. Moreover, in modern Asian humans, the number of protein–protein interactions is slightly lower within large segments of Neanderthal ancestry than in the rest of the genome. However, we did not add this factor to the bootstrap test because this difference was subtle and in the conservative direction (not accounting for it makes it harder to detect an excess of introgressions), with VIPs having far more protein–protein interactions than non-VIPs.

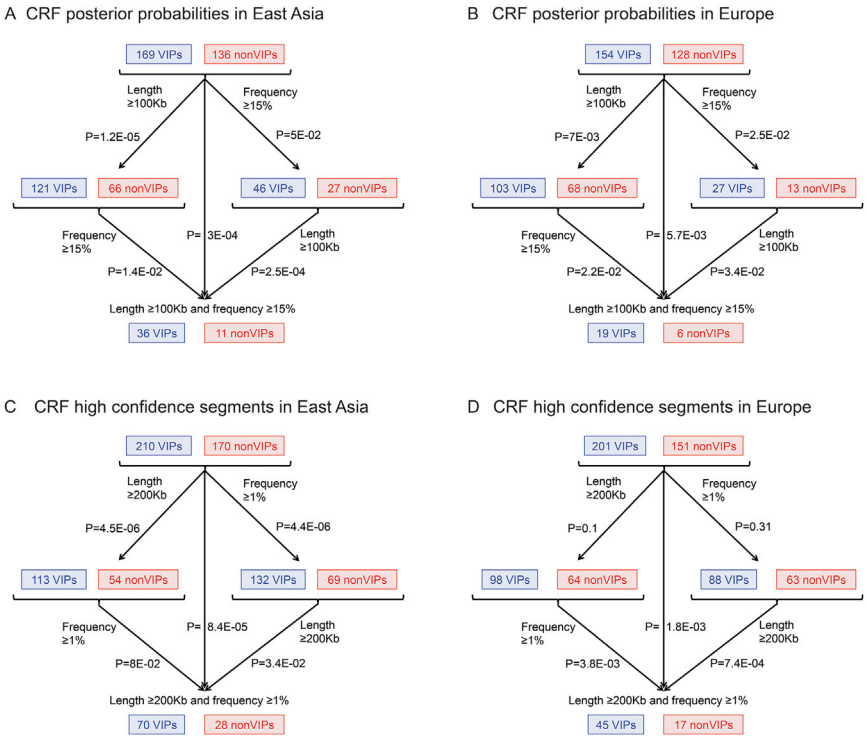



**Figure S3. Hypergeometric Test Results for the Excess of Long and Frequent Neanderthal Segments, Related to Figure 3**

The p values represented are the p values of the distinct hypergeometric tests conducted

(A) Introgressed segments in East Asia defined using the population-wide CRF posterior probability of Neanderthal ancestry

(B) Introgressed segments in Europe defined using the population-wide CRF posterior probability of Neanderthal ancestry.

(C) Introgressed segments in East Asia defined using only the high-confidence CRF posterior probability ($\geq$0.99) individual segments of Neanderthal ancestry.

(D) Introgressed segments in Europe defined using only the high-confidence CRF posterior probability ($\geq$0.99) individual segments of Neanderthal ancestry. The figure reads as follows. As an example, in (A) in East Asia we start with a total of 169 introgressed segments at VIPs and 136 introgressed segments at control non-VIPs. Of these 169 and 136 introgressed segments, 121 at VIPs and 66 at non-VIPs are longer than 100kb (arrow going down to the left). This sample of 121 long segments at VIPs and 66 long segments at non-VIPs is highly skewed toward long segments at VIPs compared to random expectations given the initial population of 169 segments at VIPs and 136 segments at non-VIPs. As a result, the hypergeometric test is highly significant ($p = 1.2x10^{-5}$). This p value for the hypergeometric test is represented next to the left arrow that connects the initial population of segments from which the sample of segments longer than 100kb is taken from. The left arrow further down connects the sample of segments longer than 100kb and the subset of those segments that in addition to being longer than 100kb, are also at frequencies higher than 15%. There are 36 such segments at VIPs and 11 at non-VIPs, which given the original sampling population of 121 long segments at VIPs and 66 at non-VIPs is again unexpected according to the hypergeometric test ($p = 1.4x10^{-2}$). Note that even though high-confidence segments and the CRF posterior probability segments largely overlap, their estimated frequencies are very different and the frequency of the high confidence segments is typically lower than the frequency of the corresponding overlapping CRF posterior probability segments. This is because the high confidence Neanderthal haplotype fragments only represent a limited subset of all the Neanderthal haplotype fragments at any Neanderthal introgressed segment. We therefore use two very different frequency thresholds for the CRF posterior probability segments (15%) and the high confidence segments (1%). This means that we do not expect the same numbers of segments overlapping VIPs and non-VIPs when using the two different types of segments. Note also that the overall number of segments at VIPs and non-VIPs is higher when using high confidence segments (for example in C. 210 and 170 versus 169 and 136 in A.) because we only used CRF posterior probability segments at frequencies higher than 5% and multiple high confidence segments are associated with CRF posterior probability segments at frequencies lower than 5%.

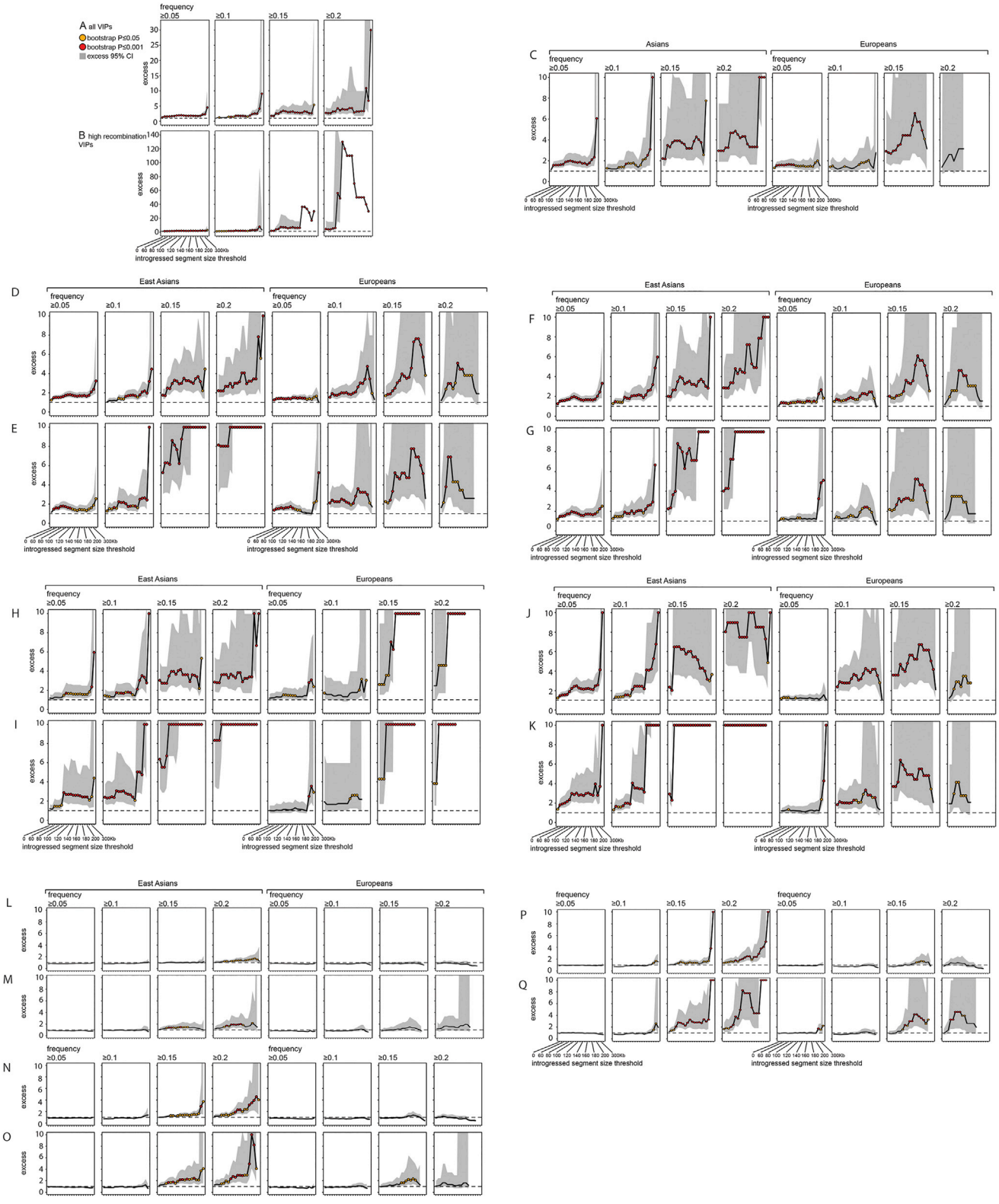



**Figure S4. Additional Controls for the VIPs versus Non-VIPs Comparison, Related to Figure 3**

(A and B) same as Figures 3A and 3B but showing full-scale enrichments of Neanderthal ancestry at VIPs in East Asia. In Figure 3 the enrichments of Neanderthal ancestry at VIPs were represented using a maximum of ten fold, meaning that enrichments beyond ten fold appeared at the ten fold plateau on the figure. This was

*(legend continued on next page)*

done to make the important trends at lower enrichment values visible to the reader. This figure represents the enrichment in Neanderthal ancestry in East Asia, but this time without imposing a ten fold maximum enrichment representation. The enrichments were obtained using a shrinkage parameter of 0.1 in the cases where zero segment overlapped control non-VIPs. For example in high recombination regions there are 13 Neanderthal segments at frequencies higher than 20% and longer than 120kb that overlap VIPs, versus zero segments that overlap control non-VIPs on average. In this case we replace zero for the non-VIPs with 0.1 and the excess is therefore 13/0.1 = 130. In addition, in this case because all the random sets of control non-VIPs have zero overlapping segments we are not able to measure a confidence interval.

(C) same as Figure 3A but using only LT-VIPs.

(D and E) same as Figures 3A and 3B but using the deCODE recombination map.

(F and G) same as Figures 3A and 3B but with an additional control for McVicker's B in 50kb windows centered on genes. H and I same as Figures 3A and 3B but using only adaptive introgressed loci.

(J and K) same as Figures 3A and 3B but using only genes with genomic spans less than 50kb.

(L) All VIPs compared to non-VIPs including VIPs very close to VIPs (> 0kb). No control for any confounding factor.

(M) Same as A but using only high recombination regions of the genome.

(N) All VIPs compared to non-VIPs including VIPs at 250kb or further from VIPs. No control for any confounding factor.

(O) Same as (N) but using only high recombination regions of the genome.

(P) All VIPs compared to non-VIPs including VIPs at 500kb or further from VIPs (distance used for Figure 3, see STAR Methods). No control for any confounding factor.

(Q) Same as (P) but using only high recombination regions of the genome.

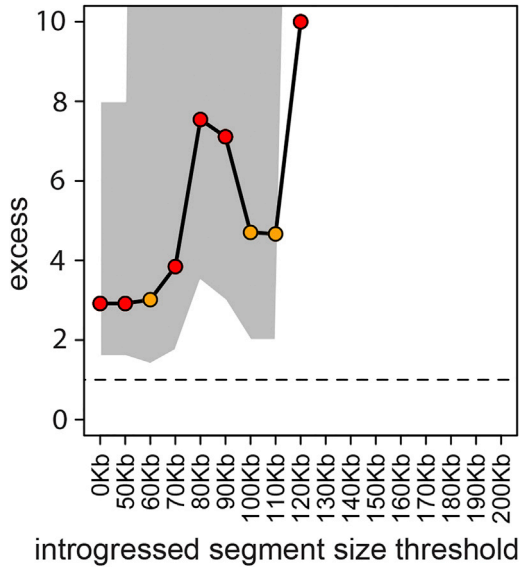
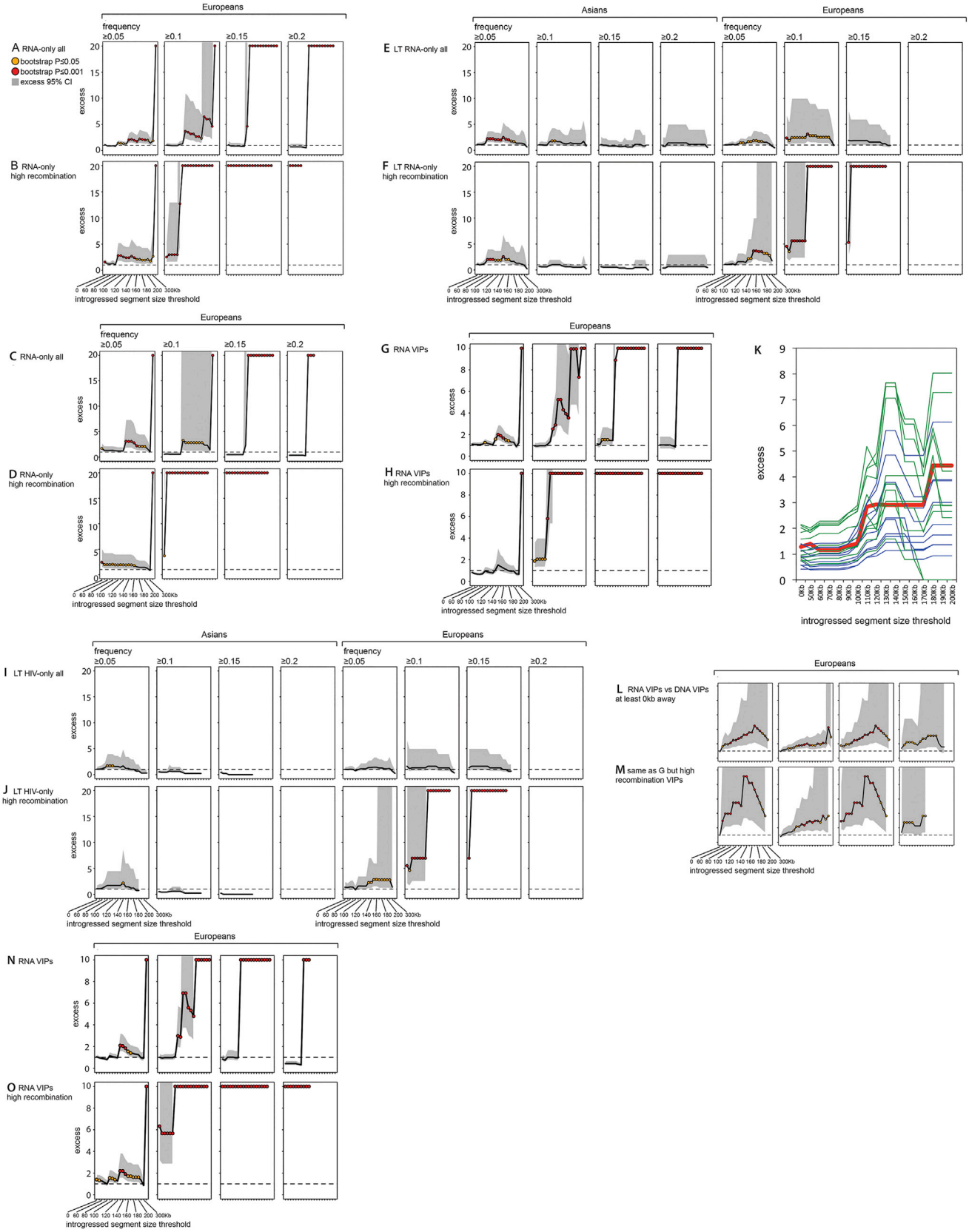



**Figure S5. Excess of Introgression from Modern Humans to Neanderthals at LT-VIPs, Related to Figure 4**

Legend as in Figure 4.




*(legend on next page)*

**Figure S6. Additional Controls for the RNA versus DNA VIPs and Specific Virus Comparisons, Related to Figures 5 and 6**

(A and B) same as Figures 5A and 5B but without VIPs interacting with pathogens other than viruses and without immune genes.

(C and D) same as Figures 5A and 5B but using only adaptive introgressed loci.

(E and F) same as Figures 5A and 5B but using only LT RNA and DNA VIPs.

(G and H) same as Figures 5A and 5B but using the deCODE recombination map.

(I and J) same as Figures 6C and 6D but using only LT HIV VIPs.

(K) Insufficient power to detect a significant excess of Neanderthal introgressions in European modern humans at HCV-only VIPs. In contrast to HIV-only and influenza virus–only VIPs, we did not detect a significant excess (bootstrap test $p > 0.05$) of introgressions at HCV-only VIPs (Figures 6E and 6F). However, this could simply reflect insufficient power to detect an excess due to the fact that there are far fewer HCV-only VIPs than HIV-only or influenza virus–only VIPs (108 versus 320 and 374, respectively, that can be used in the bootstrap test). To evaluate the power to detect a significant excess of introgressions with only 108 VIPs, we sub-sampled ten random sets of 108 HIV-only VIPs and 108 influenza virus–only VIPs. We then ran the bootstrap test to compare each of these random sets with DNA-only VIPs, just as we did when comparing HCV-only VIPs with DNA-only VIPs. We then compared the observed excess for the random sets (blue curves for HIV, and green curves for influenza virus) with the actual excess measured for the 108 HCV-only VIPs (red curve). We used the bootstrap test with introgressions at frequencies higher than 10%, which corresponds to the frequency threshold where we measured the highest excess for HCV-only VIPs. The graph shows that the excess at HCV-only VIPs is within the range of excess for sub-sampled HIV-only and influenza virus–only VIPs, demonstrating that in the case of HCV, we did not have enough statistical power to draw a conclusion.

(L) RNA VIPs compared to DNA VIPs, including DNA VIPs very close to RNA VIPs (> 0kb). No control for any confounding factor.

(M) Same as (L) but using only high recombination regions of the genome.

(N and O) same as Figures 5A and 5B but using only genes shorter than 50kb.

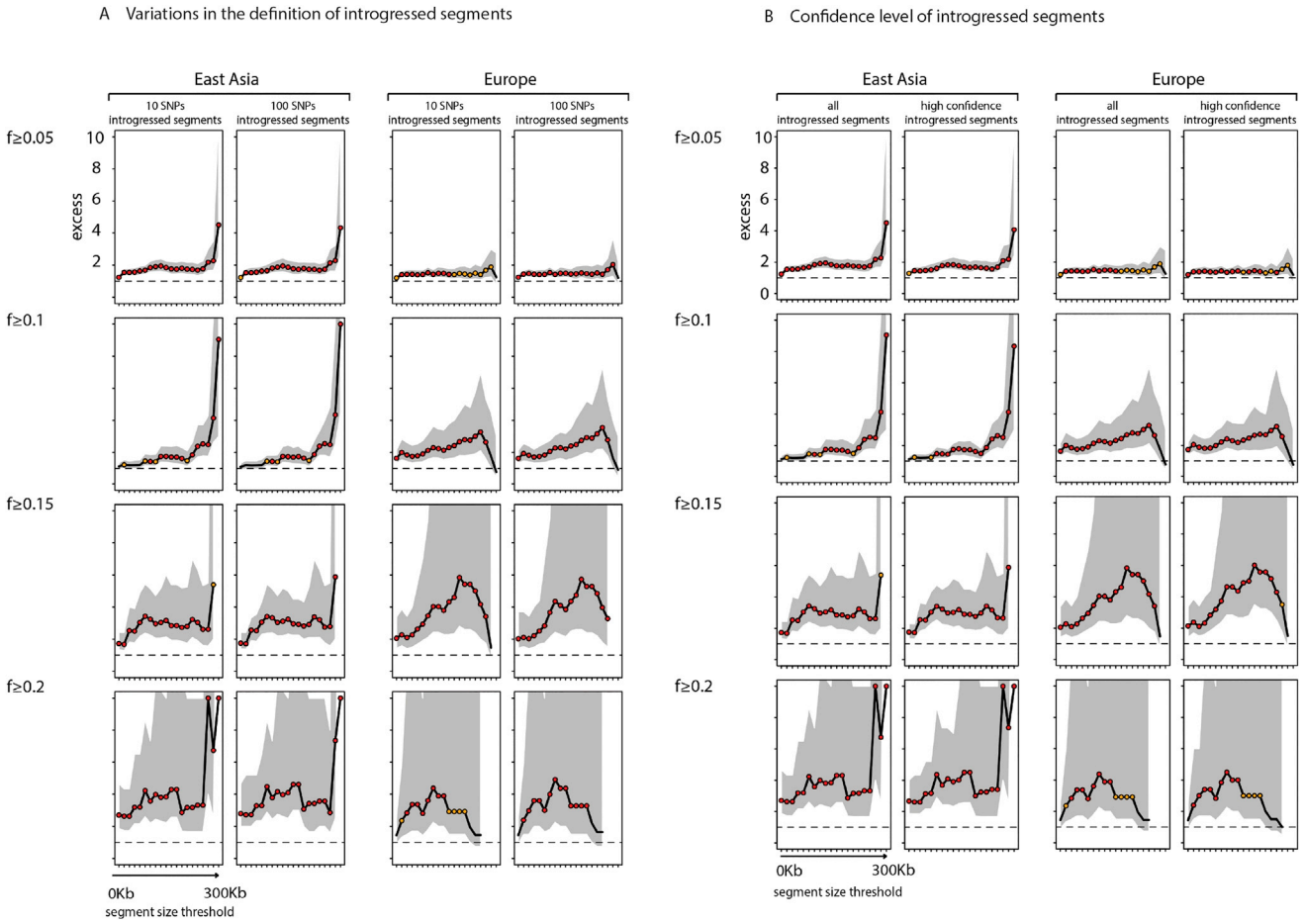



**Figure S7. Robustness of the Results to Variations in the Definition of Introgressed Segments, Related to STAR Methods**

Here high confidence introgressed segments means all CRF segments that happen to overlap with high confidence segments.