Cell Genomics Transparent Peer Review Record

Genetic adaptation to pathogens and increased risk of inflammatory disorders in post-Neolithic Europe

Authors:

Gaspard Kerner,[1*] Anna-Lena Neehus,[2,3] Quentin Philippot,[2,3] Jonathan Bohlen,[2,3] Darawan Rinchai,[4] Nacim Kerrouche,[4] Anne Puel,[2,3] Shen-Ying Zhang,[2,3,4] Stéphanie Boisson-Dupuis,[2,3,4] Laurent Abel,[2,3,4] Jean-Laurent Casanova,[2,3,4,5,6] Etienne Patin,[1,8] Guillaume Laval,[1,8] and Lluis Quintana-Murci,[1,7,8,9,*]

Summary

Initial submission: Received : 6/20/22

Scientific editor: Laura Zahn

First round of review: Number of reviewers: 2
Revision invited: 8/9/22
Revision received : 10/24/22

Second round of review: Number of reviewers: 1
Accepted : 12/14/22

Data freely available: Yes

Code freely available: Yes

Referees' reports, first round of review

Reviewer #1: My review is rather short since the manuscript is already in an excellent shape. The authors used a computationally costly but straightforward ABC approach to estimate the time and strength of selection based on frequencies reconstructions from ancient DNA from Europe. The main strength is that the authors clearly understand that it is ok if their arguably arbitrary cutoff of 1% from neutral simulations results in including false positives, because false positive occur randomly and will never create on their own the very impressive immune functional enrichments discovered by the authors. In this respect this work must be praised because it reveals the true extent of recent, pervasive pathogen-driven adaptation. Previous approaches with ancient DNA that tried to reduce false positives at all cost did it at the expense of genome-wide statistical power and forgot the fact that false positives will not occur at specific functions in the genome. This is an extremely common mistake that is made repeatedly and has led some to erroneously conclude that recent positive selection was rare. This manuscript provides very strong, biological, empirical evidence that it is not the case and that very clear functional immune patterns appear once reaching a better statistical compromise between specificity AND power. It might be the case that other reviewers will complain about false positives as they often do. Again, false positives do not magically know where loci with immune functions are located in the genome, and thus cannot create the very strong immune enrichments discovered by the authors. Saying otherwise is to make a very severe, basic statistics error.

I have a few technical concerns that I believe the authors will be able to address.

P6L4. Please mention that you check statistical significance of enrichment using a single gene per locus to control for gene clustering.

P6L28. "This result cannot be explained by differences in detection power, as our approach has a higher power for variants with selection beginning at earlier timepoints". Is this true if the selection started at earlier time points but then remained only for a few thousand years? Here I am wondering if the apparent higher number of more recent selection candidates could instead reflect the fact that pathogen selective pressures are overall short-lived and that results in less power to detect older events that were not sustained all the way to close to the present. The authors should try if possible to test it, and if not possible to discuss the fact that they cannot exclude the possibility. One way to look at it would be to first use the population simulations to estimate the power to detect selection events that started early but where selection lasted only for a few thousand, maybe around 3,000 years since it seems to be the average age of a lot of selection events found by the authors. Second, the authors could look at actual frequency trajectories of older selection events that they detected to see if some start with an increase early but then plateau after a few thousand years.

P31L7. "Taking LD and derived allele frequency (DAF) (r 2 32 < 0.6) into account, we found that candidate variants were enriched in cis-eQTLs in whole blood, particularly for strong eQTL associations". This is a very interesting and compelling result, but given that eQTLs tend to be clustered together even in rather large regions in LD, I would like to know more about what the authors mean by taking LD into account, and I would like to know if this means or not that clustering of eQTLs was taken into account since it will increase the null expected variance for the encrihment. Can the authors do as they did when they used a single gene per locus to estimate gene clustering-aware enrichments?

In the Methods you need to give explicitely the population sizes simulated at different times.

The authors need to give more details about what recombination patterns were simulated for the ABC. Uniform recombination? To what level? Recombination patterns reproducing the known recombination patterns at a given tested locus? This requires much more detail about what was done and about how having simulated recombination patterns that differ from the real recombination map could have created biases.

Reviewer #2: M&M

Empirical p value computation

A) You mention that the significance thresholds are influenced by the frequency of the variants and, therefore, you group the markers according to certain frequency bins. However, despite your sentence "We identified the bin to which a variant belonged by calculating, for each variant, the CI for allele frequency estimation at each epoch, according to an approximation to the normal distribution of the 95% binomial proportion CI."; it is still unclear how the boundaries of these frequency bins were defined and why the bins are not consistent between positive ([0.025-0.2]; [0.2-0.6] and [0.6-0.8]) and negative ([0.025-0.05]; [0.05-0.1]; [0.1-0.2] and [0.2-0,8]) selection analyses. Furthermore, according to the explanation provided, one would expect 1) the existence of different frequency bins in each epoch (apart from those used). 2) It is possible that for a given frequency bin different sizes (number of members) might be expected for each epoch... and, consequently, different null distributions that might be appropriate for normalization purposes.....
How were these issues addressed to avoid biasing the results?

B) Based on your text: "We ended up with 21,129 candidate variants for positive selection, and 27,591 for negative selection (psel <0.01)", it appears that your results showed an inflated number of significant markers. Could this also be in part due to the lack of multiple testing correction?, because it appears that it has not been applied.

C)You say that "approximated the empirical null distribution with a known theoretical distribution, to improve discrimination between very small p values.". Why did you not use this approach to create the frequency intervals or their quantiles to identify possible significant markers?

Time of selection onset for positively selected loci
A) One of your goals here was, as you put it, "We investigated whether the frequency trajectories based on both ancient and modern DNA samples resulted in biased T estimations, due to differences in genotype calling between datasets" However, further on you say: "We thus repeated the ABC estimation for frequency trajectories, but we excluded the last epoch corresponding to current frequencies." Given all this, it is not clear how you treated the modern data set, since, according to the text, the last epoch was removed: "current" frequencies, i.e. those from modern DNA samples. I would appreciate an extended explanation of the treatment of the modern data and why the use of pseudo-haplotypes instead of full SNP-calls was not investigated as potential T estimate bias.

Results

Genetic adaptation has occurred principally since the Neolithic period
B) Typically, for Figure 4 (A) one would expect to see significant points above a certain p-value threshold (ideally marked as a line on the graph). However, the points marked as significant are at the bottom with a -log10(p) close to 0 (meaning a p-value ≈ 1). The same unexpected effect is observed in the selection coefficient estimates. Therefore, this graph is counterintuitive, suggesting that the significant markers are not significant and are not the ones with the highest selection coefficient (in this case it should be placed as the lowest since it is negative selection). Could you explain why this issue and/or find a way to better display your results.

Figure 3B, the timing and the ancestral components here cannot fit. In the period 10000-7500 BP there are

no individuals north of the Alps in Western Europe that carry an Anatolian genetic component. There seems to be a general problem with the time slices presented here.

Authors' response to the first round of review

Reviewer #1:

My review is rather short since the manuscript is already in an excellent shape. The authors used a computationally costly but straightforward ABC approach to estimate the time and strength of selection based on frequencies reconstructions from ancient DNA from Europe. The main strength is that the authors clearly understand that it is ok if their arguably arbitrary cutoff of 1% from neutral simulations results in including false positives, because false positive occur randomly and will never create on their own the very impressive immune functional enrichments discovered by the authors. In this respect this work must be praised because it reveals the true extent of recent, pervasive pathogen-driven adaptation. Previous approaches with ancient DNA that tried to reduce false positives at all cost did it at the expense of genome-wide statistical power and forgot the fact that false positives will not occur at specific functions in the genome. This is an extremely common mistake that is made repeatedly and has led some to erroneously conclude that recent positive selection was rare. This manuscript provides very strong, biological, empirical evidence that it is not the case and that very clear functional immune patterns appear once reaching a better statistical compromise between specificity AND power. It might be the case that other reviewers will complain about false positives as they often do. Again, false positives do not magically know where loci with immune functions are located in the genome, and thus cannot create the very strong immune enrichments discovered by the authors. Saying otherwise is to make a very severe, basic statistics error.

We thank the Reviewer for their positive appraisal and their comments, which have been all considered during the revision and have considerably improved the clarity of the manuscript. Point by point responses are provided here below.

I have a few technical concerns that I believe the authors will be able to address.
P6L4. Please mention that you check statistical significance of enrichment using a single gene per locus to control for gene clustering.

RESPONSE: As suggested by the reviewer we have now rephrased the sentence as follows: "These 89 loci were also enriched in a curated list of immunity genes whether we considered all candidate genes (OR = 1.6, p = 8.0 Å~ 10^{-3}) or, to account for gene clustering, a single gene per locus (28/89 loci; OR = 1.6, p = 0.049)" (p. 6 l. 8-9 of the revised manuscript).

P6L28. "This result cannot be explained by differences in detection power, as our approach has a higher power for variants with selection beginning at earlier timepoints". Is this true if the selection started at earlier time points but then remained only for a few thousand years? Here I am wondering if the apparent higher number of more recent selection candidates could instead reflect the fact that pathogen selective pressures are overall short-lived and that results in less power to detect older events that were not sustained all the way to close to the present. The authors should try if possible to test it, and if not possible to discuss the fact that they

cannot exclude the possibility. One way to look at it would be to first use the population simulations to estimate the power to detect selection events that started early but where selection lasted only for a few thousand, maybe around 3,000 years since it seems to be the average age of a lot of selection events found by the authors. Second, the authors could look at actual frequency trajectories of older selection events that they detected to see if some start with an increase early but then plateau after a few thousand years.

Response to Reviewers

2

RESPONSE: We agree with the Reviewer that transient selection is a very likely scenario in the context of infectious diseases. Following their advice, we have now assessed whether such transitory episodes of selection may partly explain the observed higher number of recent selection events (p. 7, l. 5-9, **Figures S3C-D**). To this end, we first simulated 80,000 positively-selected variants ($0.01 < s < 0.05$) for which selection started at time T_{start}, sampled from a uniform distribution bounded by 0 and 10,000 years, but lasting for only 3,000 years (i.e., selection ends at $T_{end} = \max(0, T_{start} - 3{,}000)$). We then used these simulations as pseudoempirical data and tested if short-lived events of selection were detectable using our ABC approach assuming continuous, ongoing selection. This analysis showed that 49.8% of the detected selective events have an estimated onset of selection before the start of the Bronze Age, and the remaining 50.2%, after the start of the Bronze Age (**new Figure S3D**). Since the Bronze Age spans half of the study timeframe, these results indicate that we have the same power to detect transient selection events starting before or after the beginning of the Bronze Age. **Therefore, our new analysis indicates that the excess of post Bronze Age events of positive selection is not due to reduced power to detect short-lived events of selection starting in the Neolithic (e.g., in the case of epidemics caused by extinct pathogens).**

Furthermore, to assess whether old short-lived events of selection have T estimates biased downwards, we simulated 30,000 positively-selected variants ($0.01 < s < 0.05$) for which selection started at T_{start}, randomly chosen between 6,000 and 10,000 years ago, and, once more, has lasted for only 3,000 years. We show that the estimated onset of selection is not biased downwards and is consistent with the simulated time of onset of selection (**new Figure S3C**). Finally, we show that most detected variants under positive selection are not plateauing after a first increase in frequency in the study timeframe. On average, our top 89 variants (**Table S2**) show constant increases in frequency with time (mean and 95% CI shown in the left panel of the **new Figure S3B**), even though for the few events of selection dating to the Neolithic (including for example LCT and SLC45A2), the relative increase seems to slow down slightly over the last two millennia (mean and 95% CI shown in the right panel of the **new Figure S3B**). This suggests that most of these variants have increased in frequency until recently, and that their frequency trajectory is consistent with the estimated time of onset of selection.

In the revised Results section 'Genetic adaptation has occurred principally since the Neolithic period' (p. 7, l. 5-9), we now report that the higher number of selection candidates after the Bronze Age is not due to a lower detection rate of old short-lived events of selection, and briefly discuss this topic in the Discussion (p. 14 l. 1-2). We have also added the new supplementary panels B-D to Figure S3 and changed the legend accordingly.

P31L7. "Taking LD and derived allele frequency (DAF) (rÇ < 0.6) into account, we found that candidate variants were enriched in cis-eQTLs in whole blood, particularly for strong eQTL associations". This is a very interesting and compelling result, but given that eQTLs tend to be clustered together even in rather large regions in LD, I would like to know more about what the authors mean by taking LD into account, and I would like to know if this means or not that clustering of eQTLs was taken into account since it will increase the null expected variance for the enrichment. Can the authors do as they did when they used a single gene per locus to estimate gene clustering-aware enrichments?

RESPONSE: We agree with the Reviewer that a clarification is needed regarding the approach used to account for LD, when testing for enrichments. As stated in the methods section 'Enrichment analyses for positively selected loci': "We used independent variants to determine enrichment, by pruning variants in LD with the plink command --indep-pairwise 100 10 0.6 --maf 0.01, on our aDNA dataset, thus removing variants with $r^2 > 0.6$ in 100 kb windows, using sliding windows of 10 variants. For the HLA region, considered to lie between hg19 coordinates 27,298,200 and 34,036,446 of chromosome 6, we used a more conservative LD pruning method considering 1,000 kb rather than 100 kb windows (plink command --indep-pairwise 1000 100 0.6 --maf 0.01), for variants with a minor allele frequency (MAF) >1%. Where indicated, we also matched the DAF distribution of the pruned dataset to that of the studied group of variants (e.g., eQTLs or GWAS variants), using 5% frequency bins." **Therefore, we tested enrichment by determining, among independent variants in the aDNA data defined either as candidates for selection (p_{sel} <0.01) or not (p_{sel} ≥ 0.01), the proportion of these independent variants that are also significantly associated with gene expression.** As a result, the contingency table used to test for enrichment does not include eQTL variants in strong LD.

In addition to this section in the Methods, we have now clarified in the main text how we accounted for LD among eQTL variants (p. 8 l. 5-7). Furthermore, we have added to the revised Methods section 'Enrichment analyses for positively selected loci' details about the eQTLGen consortium used to retrieve whole blood cis-eQTLs for this analysis (p. 26 l. 14-16).

In the Methods you need to give explicitly the population sizes simulated at different times.

RESPONSE: We thank the Reviewer for pointing this out. Although we stated that simulated sample sizes per epoch are equal to those of the observed aDNA data (Methods section 'Allele frequency trajectories across epochs'), we did not explicitly mention that our simulations reproduced the same sample sizes per epoch.

In the revised text, we have now clarified in the Methods section 'Forward-in-time simulations' (p. 22 l. 11-13) that we have simulated the same sample sizes per epoch as in the observed data, i.e., 729 samples for the Neolithic, 893 for the Bronze Age, 319 for the Iron Age, 453 for the Middle Ages and 503 for the present epoch.

More generally, we also agree that we should report all the parameters of the simulated demographic model, including effective population sizes. Thus, we have now included a new

supplementary table reporting all simulated demographic parameters, including effective population sizes (new Table S10), that we cite p. 21 l. 26.

The authors need to give more details about what recombination patterns were simulated for the ABC. Uniform recombination? To what level? Recombination patterns reproducing the known recombination patterns at a given tested locus? This requires much more detail about what was done and about how having simulated recombination patterns that differ from the real recombination map could have created biases.

RESPONSE: We estimated selection parameters and tested for selection at each variant by using only the temporal trajectory of allele frequencies for the corresponding variant. Therefore, our ABC approach does not require information from nearby sites. Since it is unnecessary in our analysis and computationally costly to simulate linked variation forward in time, variants were simulated one at a time.

In the revised text, we have now clarified in the Methods section 'Forward-in-time simulations' (p. 22 l. 16-18) that variants were simulated one at a time.

Reviewer #2:

We thank the Reviewer for their comments and suggestions, which have been all considered during the revision and have considerably improved the quality and clarity of the manuscript. Point by point responses are provided here below.

M&M

Empirical p value computation

A) You mention that the significance thresholds are influenced by the frequency of the variants and, therefore, you group the markers according to certain frequency bins. However, despite your sentence "We identified the bin to which a variant belonged by calculating, for each variant, the CI for allele frequency estimation at each epoch, according to an approximation to the normal distribution of the 95% binomial proportion CI."; it is still unclear how the boundaries of these frequency bins were defined and why the bins are not consistent between positive ([0.025-0.2]; [0.2-0.6] and [0.6-0.8]) and negative ([0.025-0.05]; [0.05-0.1]; [0.1-0.2] and [0.2-0,8]) selection analyses. Furthermore, according to the explanation provided, one would expect 1) the existence of different frequency bins in each epoch (apart from those used). 2) It is possible that for a given frequency bin different sizes (number of members) might be expected for each epoch... and, consequently, different null distributions that might be appropriate for normalization purposes.....
How were these issues addressed to avoid biasing the results?

RESPONSE: We agree with the Reviewer that this section can benefit from more detailed explanations. In brief, we grouped variants in different bins of derived allele frequency and obtained, for each bin, the null distribution (without selection) of s_l, i.e., the lower bound of the confidence interval of the positive (and negative) selection coefficient s. We used these null distributions to test for positive (and negative) selection. The rationale for using bins of frequency that differ between negative and positive selection is that the estimation accuracy for the s parameter and, therefore, the power to detect selected alleles depends on the mode of selection (see differences between negative and positive selection in **new Figure S2B**) and on

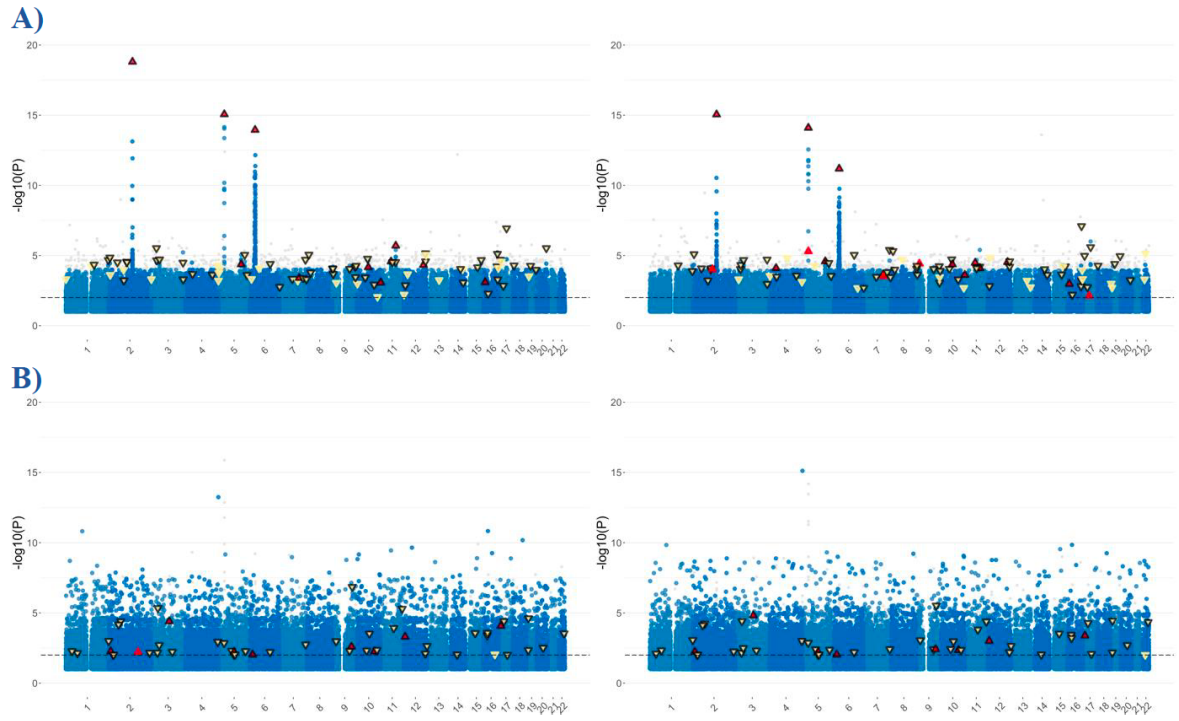
allele frequency (see differences between left and right panels of **Figures S1 and S2B-C**). For example, low-frequency variants under negative selection are hard to identify as they can be confused with low-frequency neutral variants dropping in frequency by chance. To calibrate bin boundaries, we therefore used cross-validation to assess the accuracy of the estimation (**Figure S1**) and verified that accuracy was good (high correlation between true and estimated values, $r^2 > 0.8$). We also excluded variants in the [0-0.025] bin, because estimation accuracy was poor for this bin, in both selection scenarios. Then, for each frequency bin, we computed the distribution of s_I from simulated neutral variants. To obtain the null distribution from which to derive significance thresholds, the simulated neutral variants were matched to the allele frequency spectrum of the aDNA dataset at each bin, to closely reproduce the empirical data (null distributions were obtained from a similar [same order of magnitude] number of simulations across frequency bins). **We have now clarified further our approach in the Methods section.**

Furthermore, we have conducted new analyses to justify further our approach. These analyses show that (i) using less frequency bins reduces power to detect selection and (ii) using the same number of bins with different boundaries does not affect results. Regarding the aforementioned point (i), for the analysis of negative selection, we have now computed type I errors using either a single significance threshold for all tested variants (without using frequency bins) or various thresholds of significance computed from each frequency bin, as done in this work (see new **Figure S2D). We show that, when using a single significance threshold, type I errors are strongly inflated for low frequency bins, whereas they are well calibrated at different nominal values when using significance thresholds by frequency bins.**

Regarding the aforementioned point (ii), we have estimated power by shifting the boundaries of our frequency bins. For positive selection, instead of using [0.025-0.2]; [0.2-0.6] and [0.6-0.8], we have now used [0.025-0.1]; [0.1-0.5] and [0.5-0.8]. We found no differences in power to detect positive selection between the two analyses (**new Figure S2C**). Furthermore, we obtained very similar Manhattan plots when using different frequency bin boundaries on the empirical data (**Figure 1A below**). The overlap of positive selection candidates between the two analyses (highlighted loci in the Manhattan plots with a black contour) is high, ~80% (78/102). Finally, we performed a similar analysis for negative selection, by shifting bin boundaries from [0.025-0.05]; [0.05-0.1]; [0.1-0.2]; [0.2-0.8] to [0.035-0.065]; [0.065-0.115]; [0.115-0.3]; [0.3-0.8]. By doing so, we found 49 out of the 50 highlighted variants in **Figure 4** and only one additional variant (**see Figure 1B below**). **These results demonstrate the insensitivity of our method to the definition of frequency bin boundaries.**

In the revised Methods section 'Empirical p value computation', we have now added several clarifications regarding our rationale and how frequency bins were defined (p. 23-24 l. 32-33;1-18), as well as included the new panels C and D to Figure S2 (cited in p.5 l. 13-15). These new panels show that power remains the same when shifting the boundaries of frequency bins and that type I errors are well calibrated for the frequency bin boundaries used, but not when reducing the number of bins. In addition to statistical power for positive selection, we have now included the power for negative selection in the updated **Figure S2B**.

Figure 1: Comparison of Manhattan plots for different frequency bin boundaries



Manhattan plots showing the genome-wide analysis of (A) positive and (B) negative selection obtained either from the frequency bins from the initial version of the manuscript (left; [0.025-0.2]; [0.2-0.6] and [0.6-0.8]; and [0.025-0.05]; [0.05-0.1]; [0.1-0.2]; [0.2-0.8], respectively for positive and negative selection) or from new frequency bins (right; [0.025-0.1]; [0.1-0.5] and [0.5-0.8]; and [0.035-0.065]; [0.065-0.115]; [0.115-0.3]; [0.3-0.8], respectively for positive and negative selection). Only bins of frequency used to determine thresholds of significance differ between left and right panels. Overlapping loci or variants are highlighted with a black contour in both panels.

B) Based on your text: "We ended up with 21,129 candidate variants for positive selection, and 27,591 for negative selection (psel <0.01)", it appears that your results showed an inflated number of significant markers. Could this also be in part due to the lack of multiple testing correction?, because it appears that it has not been applied.

RESPONSE: The reviewer is correct in that we found a larger number of significant markers for both positive and negative selection than expected under the null hypothesis of no selection. Indeed, if all variants were neutral, 1% of the tested variants would be expected to have a p value <1% ($p_{sel} < 0.01$), while we find ~ 3% (21,129/712,344) of SNPs with a significant signal of positive selection. These results suggest either an uncontrolled bias (i.e., unaccounted demographic changes) or the presence of true signals. In the latter case, we expect that ~70% of our significant signals are likely due to selection, a scenario supported by the fact that we replicate previously detected loci under selection (**Table S1**).

To exclude that the demographic model used to obtain null distributions is causing this signal, we conducted a battery of enrichment analyses (carefully accounting for LD). We found an enrichment of positive selection signals among missense mutations, whole blood cis-eQTLs (**Figures 2A-B**) and several immune-related traits (**Figures 2C-D**), an unlikely observation if

the 3% of significant markers were only false positives. We have also included a new supplementary table showing enrichment of missense and low-frequency variants among candidate negatively-selected variants (new Table S6). Importantly, the loci detected to be under positive selection by our method include the 12 loci previously shown to be subject to positive selection in Europe (Mathieson et al., Nature 2015). **Together, these results provide empirical evidence that variants with $p_{sel} < 0.01$ are enriched in true positives.**

To further address the Reviewer's concern, we now show QQ-plots for the neutral and observed distributions of the lower bound of the confidence interval of the positive (and negative) selection coefficient (s_l), used to detect significant markers (**see new Figures S2EF**). Similarly to QQ-plots of p values, these plots enable the detection of biases (i.e., systematically larger observed s_l values, relative to neutral values) or the presence of true positives (i.e., s_l observed values larger than expected in the right tail of the distribution only). This analysis shows the absence of systematic biases and indicates instead the presence of true positives. Finally, we have now determined how many variants show a significant signal of positive selection, when applying a Bonferroni correction (SNPs with $p_{sel} < 10^{-7}$). Again, we observed an excess of significant SNPs, even after excluding SNPs in LD and keeping the lowest p value in each detected regions: 3 SNPs out of 712,344 remain significant after multiple testing correction (the red dots with $p_{sel} < 10^{-7}$ in the Figure 1 above), whereas 712,344 Å~ $10^{-7} = 0.07 \approx 0$ is expected.

In the revised Results section 'Searching for the footprints of time-dependent negative selection', we have now included a new supplementary table showing the enrichment of negative selection signals in missense and low-frequent variants (**Table S6**; p. 11 l. 17). Furthermore, we have included additional panels to **Figure S2** (E and F) showing neutral and observed s_l distributions for both positive and negative selection. These panels are now cited in revised main text (p. 5 l. 27 and p. 11 l. 14).

C) You say that "approximated the empirical null distribution with a known theoretical distribution, to improve discrimination between very small p values.". Why did you not use this approach to create the frequency intervals or their quantiles to identify possible significant markers?

RESPONSE: We agree with the Reviewer that using a known theoretical null distribution to create the frequency intervals or their quantiles could have been an alternative approach to test for significance. However, in this study, we used an ABC approach that allows both parameter estimation and testing for selection, throughout the use of null distributions of the parameter estimates (here s_l) obtained by simulations assuming no selection. Our approach has three main advantages:
- First, ABC enables the estimation of the selection parameters, which is a major aim of this study. Using ABC, we systematically estimate the age of selection and show an enrichment of positive selection events postdating the beginning of the Bronze Age. We also provide formal estimations of the intensity and the age of negative selection at many variants across the genome. These novel results would not have been achieved using the statistical approach proposed by the Reviewer.
- Second, ABC enables to account for uncertainty in the parameter estimation, a key step to

properly assess selection significance. Indeed, powered by extensive simulations and the use of nuisance evolutionary parameters, our null distribution also accounts for the uncertainty on the recent European demographic history.
- Finally, this ABC approach uses five time points across time, leading to a more informative analysis than any using only two time points.

For all these reasons, we favored ABC, even though this approach is known to be computationally intensive. Of note, we did fit a theoretical distribution on the distribution of s_I, estimated by ABC, but only for practical considerations. Indeed, the empirical p values derived from s_I (p_{selI}) are bounded by a minimum value dependent on the number of simulations conducted to estimate the empirical null distribution. Conversely, fitting a theoretical distribution to the empirical null distribution allows to discriminate between small p values and enables more direct comparisons of our results with previous approaches using theoretical distributions (Mathieson et al., Nature 2015) (**Figure 1A**).

As the Reviewer can appreciate in **Figure 1A** and **Figure 4A**, p values might slightly differ when using the approximate theoretical and empirical null distributions. Thus, as suggested by the Reviewer in a following comment, we have now updated these two figures in the revised version of the manuscript (see answer to the corresponding comment for details).

Time of selection onset for positively selected loci
A) One of your goals here was, as you put it, "We investigated whether the frequency trajectories based on both ancient and modern DNA samples resulted in biased T estimations, due to differences in genotype calling between datasets" However, further on you say: "We thus repeated the ABC estimation for frequency trajectories, but we excluded the last epoch corresponding to current frequencies." Given all this, it is not clear how you treated the modern data set, since, according to the text, the last epoch was removed: "current" frequencies, i.e. those from modern DNA samples. I would appreciate an extended explanation of the treatment of the modern data and why the use of pseudo-haplotypes instead of full SNP-calls was not investigated as potential T estimate bias.

RESPONSE: We thank the reviewer for pointing this out, and we realize that our previous explanations might have been unclear. We used the 503 unrelated European samples of the 1000 Genomes project to obtain current frequency estimates for each variant of the aDNA dataset. For the present epoch, allele frequency estimates were thus obtained from diploid data. Consistently, to obtain current frequency estimates in our simulations, we simulated 503 diploid European individuals at the last generation of the simulated population and obtained frequency estimations from them. In contrast, as frequency estimations for past epochs were obtained from publicly available pseudo-haploid aDNA data (https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypespresent-day-and-ancient-dna-data), frequency estimations for the simulated aDNA samples were obtained by computing allele frequencies from the random sampling of one chromosome per individual, mimicking pseudo-haploid data. This being said, the rationale for testing whether the combination of modern and aDNA can lead to T estimation biases is not due to the combination of allele frequencies obtained from pseudo-haploid and diploid data, as explained below, but to the fact that the 1000 Genomes project dataset and the aDNA dataset

were processed with different pipelines, which may induce biases in the estimated allele frequency trajectories. One example is how variants near indels are called (see Methods section 'Variant filtering' and **Table S9**). To test whether such differences in genotype calling pipelines lead to T estimation biases, we verified that removing the present epoch (i.e., current frequencies from 1000 Genomes project) in our ABC approach does not affect the estimation of T (see **new Figure S3E**). As we find no significant differences, we conclude that onsets of selection are estimated similarly using either the aDNA dataset alone or in combination with modern DNA data.

With respect to the second part of the Reviewer's comment, we would like to clarify that we have actually tested the use of pseudo-haploid data as a potential source of bias in T and s estimation, but we agree with the Reviewer that we should have explicitly discussed these results in the manuscript. Indeed, we have compared T and s ABC estimates from pseudoempirical data with their true values, by cross-validation (**Figure S1**). We simulated the pseudo-empirical data by sampling diploid individuals at their radiocarbon dates and generated pseudo-haploid data by sampling one allele at each locus, to closely reproduce the observed pseudo-haploid data used in this study (see Methods section 'Forward-in-time simulations'). If pseudo-haploid data would bias T or s estimates, we would expect the ABC estimation to be biased when comparing true to estimated values. However, cross-validation tests (**Figure S1**), power analyses to detect selection (**new Figure S2B-C**) and type I error estimates (**new Figure S2D**) based on these pseudo-empirical data show that parameter estimation is accurate and unbiased and that our power to distinguish selection from neutral processes is substantial. In addition, when simulating 100 sets of 89 independent variants (pseudo-empirical data) using a uniform T distribution, and matching the allele frequencies and selection strengths of each of the 89 positively selected variants detected in our analysis, we show that the resulting 100 distributions of 89 T ABC estimates did not exhibit an excess of low T estimates (**Figure S3**). **Collectively, our results clearly show that the use of ancient pseudo-haploid data did not bias T or s estimation** and thus, they cannot explain the observed excess of selection events postdating the beginning of the Bronze Age.

Following the reviewer's comment, we have now explicitly mentioned in the revised section 'Genetic adaptation has occurred principally since the Neolithic period' of the main text that no intrinsic methodological bias is observed due to the nature of the data used (p. 7 l. 4-5).

We have also added a new section in the revised methods called 'Simulation-based evaluation of the ABC approach' in which we carefully detail the analyses that were conducted assessing the performance of our ABC approach (p. 22-23 l. 33;1-9).

Results

Genetic adaptation has occurred principally since the Neolithic period

B) Typically, for Figure 4 (A) one would expect to see significant points above a certain pvalue threshold (ideally marked as a line on the graph). However, the points marked as significant are at the bottom with a -log10(p) close to 0 (meaning a p-value ≈ 1). The same unexpected effect is observed in the selection coefficient estimates. Therefore, this graph is counterintuitive, suggesting that the significant markers are not significant and are not the ones with the highest selection coefficient (in this case it should be placed as the lowest since it is negative selection). Could you explain why this issue and/or find a way to better display

your results.

RESPONSE: We thank the Reviewer for pointing this out and agree that **Figure 4A** can be confusing, as it is not meant to highlight the most negatively-selected variants among all variants. Instead, **Figure 4A** highlights negatively-selected variants that (i) have a $p_{sel} < 0.01$, (ii) are missense variants, and (iii) are located at conserved positions of the genome (GERP > 4), which represent only 0.16% of the tested variants. The rationale of this conservative approach was to find candidate pathogenic variants, so that we restricted our analysis to variants predicted to affect proteins. This being said, the reason why some p values have $-\log_{10}(p) < 2$ (i.e., $p > 0.01$) is that we have used a theoretical approximation of the empirical null distribution to estimate p values. While this approximation allows to discriminate well between low p values, it does a worse job for less significant p values as the fit is better at the tail of the distribution. Thus, an empirical p value of ~ 1% can be falsely transformed into a p value slightly above 1% using the theoretical distribution. To address the Reviewer's comment, we have now modified **Figures 1A and 4A** to represent the minimum between the p value obtained from the empirical null the distribution and that obtained from the theoretical approximation. We have also added a dashed line at $-\log_{10}(p) = 2$, indicating that variants above that line have a p value < 0.01.

In addition to changes in **Figure 1A** and **Figure 4A**, we have also added new experimental data to our work. As the Reviewer mentioned, our candidate pathogenic negatively-selected variants are not among the most significant ones so one might wonder about their clinical relevance. We have thus tested, in addition to the tested TYK2 P1104A and LBP D283G, the IL23R R381Q and TLR3 L412F variants, which we already speculated to be associated to immune phenotypes in the Discussion section. Now, 4 out of the 6 detected candidate pathogenic variants in this work have been tested experimentally. By combining overexpression system models, the use of patient cells and RNA-sequencing techniques, we tested the impact of these variants on immune-related pathways and confirmed that they impair either the expression and/or the function of the targeted genes, suggesting an overall impact on immunity and infection.

Following the Reviewer's suggestions, we have now modified Figures 1A and 4A as explained above. Furthermore, we have included new experimental data for variants IL23R R381Q and TLR3 L412F in the Results section (p. 12-13 l. 15-32;1-15), and modified the corresponding Methods (p. 30-31 l. 6-34;1-28) and Discussion sections (p. 14 l. 22-32) accordingly, as well as updated Figure 5 and the corresponding legend.

Figure 3B, the timing and the ancestral components here cannot fit. In the period 10000-7500 BP there are no individuals north of the Alps in Western Europe that carry an Anatolian genetic component. There seems to be a general problem with the time slices presented here.

RESPONSE: We thank the Reviewer for noticing this issue. Indeed, there are no individuals north of the Alps in Western Europe who carry the Anatolian genetic component more than 7.5 kya. The only individuals with such a component and within such a timeframe are from Turkey (n=39), Bulgaria (n=12), Croatia (n=3), Greece (n=2), Italy (n=4), Hungary (n=8), North Macedonia (n=1), Romania (n=1) and Serbia (n=2), as can be seen from **Table S8**. This

figure was made with the 'bleiglas' package of R version 3.6.2, which provides smooth estimates (in this case of the PRS) and accounts for uncertainty in the age and location of the samples by considering the age and geographic location of each individual as normal random variables with means given by the provided values and standard deviations estimated from the data. As the algorithm iterates over different values of such distributions, it obtains smooth estimates of the Crohn's Disease polygenic score over time and space. This approach is useful as it accounts for the inherent uncertainty in the estimates of the parameters of the aDNA samples. However, we made a mistake and we thank again the reviewer for pointing this out. Indeed, we took the individual's age and geographic location of one of these iterations instead of taking the actual observed values. For example, one of the samples displayed in France in the 7,500-10,000 ya panel is actually subject 'CHA001~merged' which is from Spain and dates back to 7,127 ya (**Table S8**).

In the revised Figure 3B, we now display the estimated age and known geographic location of each sample, so that the figure is consistent with the observed data.

Editor's comments complementing those of Reviewer 2:

I also want to provide a slightly edited version of reviewer's 2 comments for you to address as I don't think all of these aspects were relayed in their review:

There is the problem that the authors work with pseudo-haplotype calls, which allow an estimation of trends but do not necessarily reflect the true frequencies.

RESPONSE: We respectfully disagree with this comment, as explained in the specific answer to Reviewer 2. We have implemented our ABC approach to efficiently manage ancient pseudo-haploid data, a required procedure since the ancient data consist of pseudohaploid calls. Briefly, we simulated ancient samples by sampling diploid individuals at their estimated radiocarbon dates and generating pseudo-haploid data by randomly sampling one allele at each locus, closely reproducing the pseudo-haploid empirical aDNA data used in this study (see Methods section 'Forward-in-time simulations'). We then used a simulation-based procedure to evaluate the performance of our ABC approach, and compared parameter estimates to simulated values. Importantly, simulated pseudo-empirical data closely match real empirical data used in this study: they reproduce pseudo-haploid calls, the genetic drift intensity and admixture events characterizing recent European history (from Anatolian and Yamnaya-related migration waves; see **new Figure S2A**), without which the evaluation of our method would have been inaccurate and/or biased. For example, we show that both the simulations used for our ABC estimations and the pseudo-empirical data match closely ancestry proportions estimated from empirical data, suggesting no demographic biases in terms of admixture proportions, and, thus, that our simulations account for the different admixture events occurring in Neolithic and Bronze Age Europe (**new Figure S2A**).

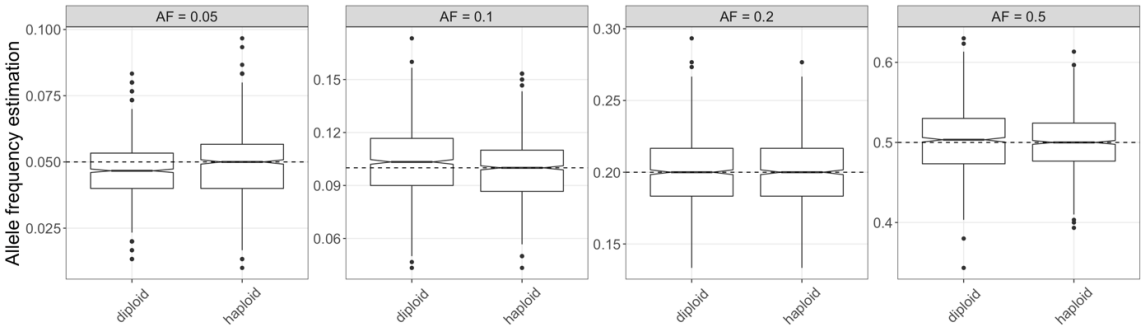
Based on these pseudo-empirical datasets, we showed by cross-validation tests (**Figure S1**), power analyses to detect selection (**new Figures S2B-C**) and type I error assessments (new Figure S2D) that both T and s estimates are accurate and unbiased, and that power to

discriminate selected SNPs is substantial. Together, these results clearly show that the use of ancient pseudo-haploid data is not problematic. Finally, it is important to highlight that pseudo-haploid calls do much better than "an estimation of trends". At equal number of chromosomes, estimations of allele frequencies based on diploid or haploid data are expected to be equivalent (**see Figure 2A below**). Moreover, at equal number of individuals (i.e., the diploid data includes twice more chromosomes than the haploid data), allele frequencies tend to show small variances when estimated from either the haploid or diploid data, as sample sizes increase towards those used in this work (**Figure 2B below).**

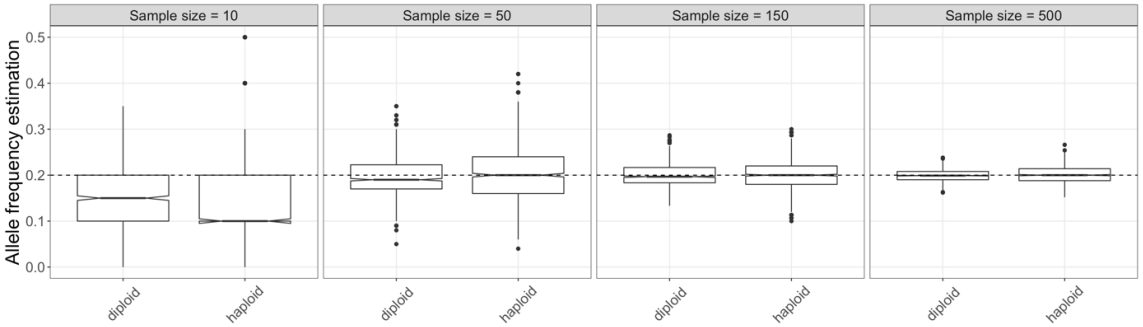
All these elements demonstrate that our approach efficiently manages the use of pseudohaploid calls, a common practice in evolutionary studies based on ancient DNA data, thus, preventing from biases in the inference of selection parameters.

Figure 2: Comparison of allele frequency estimations using haploid vs diploid data

A)



B)



A) Allele frequency estimates from diploid and haploid datasets, at equal number of chromosomes. A sample size of n = 150 diploid or equivalently n = 300 haploid samples were used to estimate population allele frequencies of 0.05, 0.1, 0.2 or 0.5 (from left to right). Homozygotes for either allele and heterozygotes were sampled using a binomial distribution, assuming Hardy-Weinberg equilibrium. For haploid samples, we then randomly sampled one chromosome from each sampled diploid individual. This process was repeated N =1,000 times. The horizontal dashed line indicates the expected allele frequency in each case.
B) Allele frequency estimates as in A) but for different sample sizes and an expected allele frequency of 0.2. In this case, 'sample size' is the number of simulated individuals for the diploid case and the number of simulated chromosomes for the haploid case.

In addition, the grouping or dating of the sample groups is problematic. Due to these two factors, it is difficult to determine the reason for the change in frequency of variants, it is difficult to distinguish whether it is really a matter of selection, drift or simply admixture

processes during the Neolithic.

RESPONSE: As mentioned above, by analyzing pseudo-empirical data, which closely reproduces genetic drift and admixture (**new Figure S2A**), we found that our ABC approach has no inflated type I errors (**new Figure S2D**), high accuracy (**Figure S1**) and substantial power to discriminate neutrality from selection (**new Figures S2B,C**). This simulation-based evaluation of our ABC approach was conducted by grouping simulated ancient individuals across the same time periods (Neolithic, Bronze Age, Iron Age, Middle Ages and present) used to analyze empirical data. As explained below, these analyses show that the sample groups used in this study are not problematic and that our ABC approach can efficiently distinguish whether the increase (or decrease) in frequency of a given variant across the predefined time periods is consistent with selection, drift or admixture.

Regarding the grouping or dating of sample groups: In the empirical data, and also in the simulated pseudo-empirical datasets used to evaluate our ABC approach, the allele frequency trajectories were computed by grouping ancient individuals according to the widely-used time periods defined by cultural and/or demographic changes in European history (Neolithic, Bronze Age, Iron Age and Middle Ages, see e.g. Skoglund and Mathieson, 2018; Childebayeva, A. et al., 2022; Mathieson and Terhorst, 2022; Feldman et al., 2019), and using the corresponding estimated radiocarbon date of each sample (https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypespresent-day-and-ancient-dna-data). Across simulations, ancient individuals were simulated according to these estimated radiocarbon dates. **Finally, the same time periods were used to compute the allele frequency trajectories in the ABC simulations, subsequently used to analyze real and pseudo-empirical data.**

Regarding how our approach discriminates between selection, drift and admixture: We have used an ancestry-aware approach, where ancestries were indeed accounted for and tracked within our simulation method. This is done by matching ancestry proportions of the simulated aDNA samples to the ancestry proportions of the real aDNA samples (we have now included, as mentioned to Reviewer 2, a new figure explicitly showing this match; see **new Figure S2A**). Therefore, the simulated allele frequency trajectories account for (i) the ancestry of the observed samples, (ii) the intensity of genetic drift (we simulated effective population sizes expected in Europe at different epochs using previously published parameters [Gravel et al., PNAS 2011]) and (iii) the fluctuation of ancestry due to admixture of local Europeans with individuals carrying Anatolian and Yamnaya-related ancestry. In other words, our simulations approximate the expectation and variance of an allele frequency trajectory conditioned on many evolutionary parameters, such as its selection coefficient s, its time of onset of selection T, ancestry, genetic drift and admixture. **Consequently, our estimation of the selection parameters s and T (and thus our method to detect selection) well-accounts for the intensity of genetic drift and admixture expected in Europe since the Neolithic**. Of note, if our method was not able to discriminate between selection, drift and admixture, our s estimation would be in turn inaccurate, i.e., true and estimated s values would strongly differ, and type I errors would exceed expectations at nominal values. However, once more, we showed by cross-validation that true and estimated s values are strongly correlated, and we show now that type I errors are controlled at different nominal

values. Lastly, note that the loci detected to be under positive selection by our method include the 12 loci previously shown to be subject to positive selection in Europe, using a different admixture-aware approach (Mathieson et al., Nature 2015).

Collectively, these results demonstrate that our approach can distinguish selection from drift in the presence of recent admixture.

The transfer of polygenic risk scores (PRS) to prehistoric societies is also problematic. These are established for modern European populations, for example, and are therefore difficult to transfer to other populations. The population genetic analyses and the authors clearly show that prehistoric populations differ from modern populations. Therefore, there is no evidence that a PRS is transferable to them. These are exciting findings but need more data to support them.

RESPONSE: We agree with this comment, reason by which we already acknowledged, in the Methods section 'Calculations of polygenic scores', that the transfer of PRS from modern to prehistoric populations is not completely straightforward (*"We acknowledge that the PRS obtained are proxies for the actual PRS across time because associated variants may have changed in frequency due to drift or admixture"*). Furthermore, PRS might be slightly over or underestimated because effect sizes can also change following environmental changes or shifts in the genetic background of past versus modern individuals. Having said that, we can notice that the variants used here to construct PRS are lead variants (genome-wide significant) in GWAS analyses of heritable traits. Thus, their effect size on disease is often less dependent on the genetic background of the individual than other classically studied traits such as height.

More importantly, in the context of the present study, we did not seek to obtain accurate estimates of the PRS of ancient individuals. Our main objective was instead to show that variants that increase genetic susceptibility to infectious diseases and autoimmune disorders nowadays have, on average, significantly decreased and increased in frequency, respectively, over the last 10,000 years. In other words, we use PRS as a methodological tool to show that disease-associated variants in the current generation have significantly fluctuated in frequency over time. More specifically, we tested the hypothesis of directional selection acting on variants affecting such traits, for which PRS are well admitted, convenient statistics as they weight variants by their effect sizes. As a complementary analysis to the test of this hypothesis, we now show that the larger the odds-ratio of GWAS variants associated with autoimmune disease risk, the larger their median selection coefficient (see **new Figure S5B**). Conversely, the larger the odds-ratio of GWAS variants associated with infectious disease risk, the lower their median s. Of note, the absolute median selection coefficient of the variants that most contribute to the PRS of either disease is higher than the 1% significant threshold used in this work. **This new analysis thus confirms the feature already captured by PRS, that is that variants contributing to the PRS of autoimmune and infectious diseases have, in average, significantly increased and decreased in frequency due to selection, with a magnitude of selection intensity (s) correlated with the GWAS effect sizes of the studied variants.**

In the revised Methods section 'Calculations of polygenic scores' (p. 27 l. 10-16) and in the

section ‘Limitations of the study’ (p. 15 l. 13-16), we now clarify further the use of PRS for our analysis and caution that PRS estimated on ancient genomes should not be interpreted as suitable estimates of disease genetic risk per se, since PRS are not directly transferable from modern to prehistoric populations. Yet, we state explicitly that variation of PRS here is only used as a tool to measure changes in allele frequency of variants associated with some immune related traits, to address questions relating to directional selection on these traits. Finally, we formally show with new Figure S5B that variants contributing to the PRS have been, on average, under natural selection, consistent with the direction proposed (increase in risk for autoimmune disorders and decrease in risk for infectious traits). We also discuss new Figure S5B in the revised Results section of the main text (p. 10 l. 7-9).

Referees’ report, second round of review

Reviewer #1: The authors have addressed my concerns. This is a very valuable contribution to better understand human genomic adaptation.

Authors’ response to the second round of review

none