Genome graphs detect human polymorphisms in active epigenomic states during influenza infection

Cristian Groza[1], Xun Chen[2], Alain Pacis[3], Marie-Michelle Simon[4], Albena Pramatarova[4], Katherine A. Aracena[5], Tomi Pastinen[6], Luis B. Barreiro[7,8,9], Guillaume Bourque[2,3,4,10,*]

1. Quantitative Life Sciences, McGill University, Montréal, QC, Canada
2. Institute for the Advanced Study of Human Biology, Kyoto University, Kyoto, Japan
3. Canadian Centre for Computational Genomics, McGill University, Montréal, QC, Canada
4. McGill Genome Centre, Montréal, QC, Canada
5. Human Genetics, University of Chicago, Chicago, IL, USA
6. Genomic Medicine Center, Children's Mercy Hospital and Research Institute, KC, MO, USA
7. Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, IL, USA
8. Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA
9. Committee on Immunology, University of Chicago, Chicago, IL, USA
10. Human Genetics, McGill University, Montréal, QC, Canada

*Corresponding author: guil.bourque@mcgill.ca

Abstract

Background

Epigenomic experiments can be used to survey the chromatin state of the human genome and find functionally relevant sequences in given cells. However, the reference genome that is typically used to interpret these data does not account for SNPs, indels, and other structural variants present in the individual being profiled. Fortunately, population studies and whole genome sequencing can assemble tens of thousands of sequences that are not in the reference [18], including mobile element insertions (MEIs), which are known to influence the epigenome [66, 60, 1]. We hypothesized that the use of a genome graph, which can capture this genetic diversity, could help identify more peaks and reveal notable regulatory sequences hidden by the use of a biased reference.

Results

Given the contributions of MEIs to the evolution of human innate immunity, we wanted to test this hypothesis in macrophages derived from 35 individuals of African and European ancestry before and after in-vitro *Influenza* infection. We used local assembly to resolve non-reference MEIs based on linked reads obtained from these individuals and reconstructed over five thousand Alu, over three hundred L1, and tens of SVA and ERV insertions. Next, we built a genome graph representing SNPs, indels and MEIs in these genomes and demonstrated improved read

mapping sensitivity and specificity. Aligning H3K27ac and H3K4me1 ChIP-seq and ATAC-seq data on this genome graph revealed between 2 to 6 thousand novel peaks per sample. Notably, we observed hundreds of polymorphic MEIs that were marked by active histone modifications or accessible chromatin, of which 12 were associated with differential gene expression. Lastly, we found a MEI polymorphism in an active epigenomic state that is associated with the expression of TRIM25, a gene that restricts influenza RNA synthesis [46].

Conclusion

Our results demonstrate that the use of graph genomes capturing genetic variability can reveal notable regulatory regions that would have been missed by standard analytical approaches.

Introduction

We have previously shown that genome graphs can recover epigenomic signal that would have been missed in genetically variable regions of the genome [29]. However, this only considered the impact of small variants on peak calls in a single genome, in an undefined biological context and did not explore the effects of such an approach on putatively functional regulatory regions. Meanwhile, structural variants involve the largest number of variable nucleotides in a population, have larger effect sizes on gene expression than SNPs or indels [56, 7] and contribute to functionally relevant epigenetic differences between humans and chimpanzees [66]. Currently, the epigenetic features that occur on structural variants are not immediately accessible when mapping to a linear reference genome [13] but could be accessed using a graph genome reference. For example, pan-genome approaches have been used to find differential CpG methylation within structural variants in twelve medaka fish genomes [41].

Obtaining accurate maps of structural variation can be challenging with short read libraries for several reasons. First, repeats are abundant in eukaryotic genomes and often prevent mapping short reads to unique locations in the genome. Therefore, resolving variation in repeats is more difficult because of decreased coverage and mapping quality. Second, mapping short reads reveals only the break points of insertions that exceed the read length. Thus, reconstructing the sequence of larger insertions requires assembling short reads into larger contigs. Third, short read assembly algorithms cannot distinguish between highly similar sequences and collapse copy number variation into one sequence. To mitigate these shortcomings, paired-end and linked read libraries have been developed. For instance, paired-end reads have already been used to assemble MEI polymorphisms [62]. Linked-read libraries [38] go further than paired-end libraries by labeling each read with a barcode that represents a small number of DNA fragments from which the read may originate. Reads that share the same barcode are clustered together into a read cloud, providing long range positional information in regions of the human genome that cannot be reached by short reads alone. Read clouds have been successfully used to genotype structural variants [8, 53, 45, 23, 5, 44] and to assemble genomes [61, 47].

To better explore the potential epigenetic role of human polymorphisms, including structural variants, we wanted to expand genome graphs to represent SNPs, indels and a select set of structural variants. In particular, we wanted to include in this graph the polymorphic mobile element insertions (MEIs) by resolving the sequence of individual TE instances by locally assembling the read clouds tagged by neighboring barcodes (Fig S1, see Methods). The true sequence of these MEIs will be incorporated into the graph genome, will reflect their evolutionary history and could reveal functional elements such as enhancers [55] and binding sites for transcription factors [9, 48].

To test whether such an approach would reveal enhancers in sequences that are absent from the reference genome, we wanted apply our graph-based analysis to a biologically relevant system. Given that endogenous retrovirus MEIs have been found to regulate innate immunity [9], this

prompts searches for other MEIs with immune function. Specifically, we were interested in the response to infection in macrophages derived from 35 individuals of african- or european-descent. In all cases, we obtained whole genome sequencing and 10X linked read data to characterized genetic variants and H3K4me1, H3K27ac ChIP-seq, ATAC-seq and RNA-seq data, before and after infection, to characterize changes in chromatin. In particular, using this data we can link the genotype of MEIs with their epigenomic state in a sample and explore their role in immunity. By looking for MEIs marked with histone modifications common to enhancers [10, 63, 49, 66] in open chromatin [14], we can identify potential regulatory sequences that are associated with the expression of genes that participate in immunity. Graph genomes are uniquely suited for this task because they can map epigenomic signal directly onto the sequences of polymorphic MEIs of a population while preserving the genomic context of the insertion. To date, similar analyses have been limited to transposable elements that are already present in the reference genome [57]. Here, we introduce non-reference transposable elements first from a single benchmark genome and then from a cohort of 35 genomes. Lastly, we identify non-reference insertions with potential regulatory function based on their epigenomic profile.

Results

Adding MEIs to the NA1878 genome graph improves read mapping

First, we wanted to test whether representing MEIs in a genome graph would noticeably improve read mapping and allow us to recover epigenomic signal on MEIs. We chose the NA12878 genome to develop and benchmark the approach to call and locally reassemble non-reference MEIs because whole genome sequencing, linked read whole genome sequencing and a haplotype resolved assembly are available [25]. We ran MELT and ERVcaller on paired-end whole genome sequencing data and identified and genotyped 2175 Alu, 351 LINE1, 106 SVA and 6 ERV insertions (Fig 1A). Of these calls, 66% (1738) were previously listed in the set compiled by Ebert et al [18]. The larger number of MEIs in our list could be due to using a very deeply sequenced library for this sample. Using our local read cloud assembly tool (`BarcodeAsm`), we managed to assemble the sequence of 1054 Alu, 117 LINE1, 17 SVA and 35 ERV copies, sometimes recovering both copies in homozygous loci (Fig 1B). We note an of excess ERV annotations due to a set of misannotated short sequences that appear with low frequency. Next, to validate the assembled contigs, we aligned them against the haplotype resolved assembly of NA12878 and calculated the proportion of their length that is spanned by the match (Fig 1C). Despite expecting errors in our local assembly and the de novo assembly, we obtain perfect hits for 784 of the 963 loci (81.4%) while 925 loci (96.1%) have hits that span more than 95% of the contig length. The hits are equally distributed between the two haplotypes of the de novo assembly. We tested our ability to map reads on these MEIs by adding them to a genome graph, aligning WGS data and re-genotyping the insertions after removing non-specific alignments (Methods). Indeed, we were able to recall 826 of the 963 (85.8%) incorporated MEIs.

Previously, we introduced an axis that orders reference genomes according to how similar they are to the genome they are meant to represent [29]. Our axis ranged from the least similar sequence (the reference genome) to the most similar (the de novo assembly). We wanted to place the resulting graph on this axis through the following read simulation and alignment experiment. We created a +SNVs graph containing only SNVs, a +indels graph containing SNVs and indels, and a +MEIs graph containing SNVs, indels and MEIs. Lastly, we created a de novo graph with all the variants called in the NA12878 haplotype resolved de novo assembly (Methods). Since reads are simulated from the de novo graph, this genome is true by construction. The 600 million reads were aligned to the increasingly complete subgraphs, starting from the reference

graph to the de novo graph. Then we compared the alignments on each graph to the ground truth of the simulation. As we align to increasingly complete genome graphs, the number of true positive alignments increases (Fig 1D). The +SNVs graph (with 3.5×10^6 SNVs) correctly finds around 2.2×10^5 more true alignments (+0.062 per SNV) compared to the reference graph. The +indels graph (with 5.2×10^5 indels), adds 3.3×10^4 true alignments (+0.063 per indel) over SNVs alone. The impact of SNVs and indels is similar because most indels are short and allelic bias in indels overcomes SNVs only at longer lengths [26]. The +MEIs graph (with 963 MEIs) adds another 1.4×10^4 true alignments (+14.5 per MEI) on top of SNVs and indels, a much larger impact than SNVs and short indels. Finally, the de novo graph outperforms our best genome by 1.4×10^6 true alignments since it represents even more structural variants. For example, we expect the de novo assembly to contain most Alu, L1 and SVA polymorphisms, while we reconstructed only a subset of MEI insertions from read clouds. Overall, the MEIs bring the graph closer to the true NA12878 diploid genome, even if they do not complete it.

Having established that our genome graph recovers more true mappings, we look for MEIs that support active histone and chromatin accessibility marks. We mapped three replicates for each H3K4me1, H3K27ac ChIP-seq and ATAC-seq library to the MEI graph and called peaks. Among all the replicates, we count 19 polymorphisms that overlap peaks in ATAC-seq, 22 in H3K27ac and 58 in H3K4me1 in at least one of the replicates (Fig 1E), with roughly half the loci observed in all three replicates. Most loci are covered on both the TE and the reference allele and a smaller subset are covered on only one of the alleles (Fig 1F). While the number of events is small in a single genome, they show that we can profile the chromatin in these regions.

Read clouds recover MEIs in a population

Next, we look to extend the method and apply it to a cohort of individuals. Similar to NA12878, we called and genotyped 7362 Alu, 1344 LINE1, 649 SVA and 19 ERV insertions among 35 individuals (Fig 2A) and ran BarcodeAsm to assemble the sequences in these regions. When checking against the MEIs compiled by Ebert et al. [18], we find that 40% (3758) of MEIs are unique to our cohort.

Next, we introduced a population consensus approach before attempting to identify the final MEI at each locus (Methods). Using this approach we were able to assemble and annotate the insertions for 5140 Alu, 316 LINE1, 94 SVA and 48 ERV distinct loci (Fig 2A). The population consensus approach allowed us to recover a larger fraction of events because the number of attempts to reassemble a locus is equal to the frequency of the MEI allele in the cohort. Consequently, while singleton MEIs are the most numerous, they are the least likely to be assembled (Fig S3). The resulting multiple sequence alignments show few ambiguous nucleotides, suggesting that the consensus insertions are representative of most samples (Fig S2A). Moreover, the length distributions of the consensus insertions include the Alu peak at 300 bp and a long tail associated with longer truncated and full length TEs such as the LINE1 (Fig S2B-C).

Again, we ranked each TE subfamily by abundance to check if known features of TEs emerge in this population (Fig 2B). For example, the AluY sub-family is responsible for the majority of Alu amplification in humans [15] while the L1HS sub-family is the only active LINE1 [32]. Therefore, the most prevalent polymorphisms should belong to these known active sub-families. Accordingly, looking at the annotations of each assembled insertion, the AluY sub-families are foremost among Alus and L1HS is ranked ahead of L1P1 and L1PA2. Similarly, the human specific SVA F sub-family [4] is the most abundant in the SVA family.

Given that some of these MEIs are frequent in the population, they might indeed be the source of enhancer [55] or other functional activity. Therefore, we detailed the genomic distribution of the assembled insertions relative to both functional and repetitive parts of the genome. To this end, we looked at the distances of each insertion to the nearest exon, enhancer and repetitive

sequences in the reference genome (Methods). We see that 92 insertions are within the bodies of exons (Fig 2C), with many more being in close proximity to exons. Notably, Alu integration in genes is involved in alternative transcription events, where an Alu becomes an exon or causes an exon to be skipped or an intron to be retained [37]. At the same time, 147 insertions are located within enhancers (Fig 2D) and more than half of insertions (2941) are nested within a repetitive sequence (Fig 2E).

To find if these patterns are unusual, we generated the same distributions with MEIs found in the 1000 Genomes Project and with a set of positions uniformly sampled from the genome. In fact, we reproduced the same patterns in the 1000 Genomes Project [58]. In the random positions, the distributions showed similar modes but had much longer tails. Therefore, the patterns are created by the distribution of exons, enhancers and repeats in the genome interacting with the distribution of MEIs in the genome.

Lastly, we checked if the insertion genotypes that were confirmed by successful assembly recapitulate the known African and European population structure of the data set (Fig 2E). As expected, we see two clusters in the principal component analysis that are consistent with genetic ancestry and SNV calls from whole genome sequencing (Fig S4).

Cohort genome graphs increase the number of peak calls

Next, we wanted to explore the extent to which the cohort genome graph impacts alignment and peak calling in epigenomic datasets relative to the reference genome. Since this graph contains roughly 3.0×10^5 deletions, 2.9×10^5 insertions and 4.1×10^6 SNPs per sample (Fig 3A) and increases the mapping rate [26] (Fig 3B), we expect new peaks to appear and some peaks to disappear, with most peaks remaining unchanged. To achieve this, we peak called H3K4me1, H3K27ac ChIP-seq and ATAC-seq alignments in the reference and the cohort genome graphs. Next, we summarized the peaks called only in the cohort graph (personal-only), those called only in the reference graph (ref-only) and those called with both the reference and the cohort graph (common peaks).

Among H3K4me1 samples, we observed an average of 4700 (2.5%) personal-only peaks and 2200 (1.2%) ref-only peaks per sample (Fig 3C, S5A). In H3K27ac, we counted 1800 (3.0%) personal-only peaks and 1100 (1.9%) ref-only peaks (Fig S6A, S5D) on average. We found that both flu-infected and non-infected ChIP-seq samples show similar numbers and do not present any condition specific patterns. However, among ATAC-seq samples, we detect considerably more altered peaks in flu-infected samples than in non-infected samples (Fig S6B, S5C). In flu-infected samples, personal-only events average 4000 (2.3%) peaks per sample and ref-only events average 2400 (1.4%) peaks per sample. In non-infected samples, the same numbers and proportions are roughly halved. We suspect that this is linked to cell death in flu-infected samples introducing cell-free DNA and more background in the ATAC-seq library preparation. Consistent with this hypothesis, infected samples show an excess of low quality peaks relative to non-infected samples (Fig S6E-F).

We asked whether altered peaks were associated with sequence variants as would have been expected. We estimated the influence of genotype on common and personal-only peaks by logistic regression on SNPs and indels within the peak while controlling for peak width (Table 1). Depending on the dataset, the log-odds for a peak to be personal-only increase by 0.11 to 0.19 when SNPs were present and by 0.45 to 0.56 when indels were present. The log-odds decrease by 0.72 to 0.87 per 100 bp of peak width. This logistical regression model has a cross-validation AUC of 0.90 to 0.92.

The extra peaks are influenced by rescued unmapped reads and by the shift of peaks across the statistical significance threshold of the peak caller [29]. This multi-sample epigenomic dataset is an opportunity to better understand how reliable the altered peaks are compared to common peaks. To do so, we contrasted the population frequency of personal-only, ref-only and common

peaks. We created a peak replication curve (inverse cumulative distribution) by counting what proportion of peaks is present in at least a set number of samples (Fig 3D, S6C-D). We also computed the curves that are expected by chance by randomly resampling the peaks associated with each sample (Methods). H3K4me1, H3K27ac and ATAC peak sets generate very similar curves, with common peaks having the longest tails followed by personal-only and ref-only peaks, with the curves expected by chance decaying the fastest. Under the resampling simulation, none of the peaks are observed in more than 20 samples, but a proportion of personal-only peaks are replicated in more than 40 samples. Therefore, mapping epigenomic data to a cohort genome graph yields more peaks that are seen across many samples.

Our dataset features flu-infected and non-infected macrophages and should contain active pathways related to immunity. To understand the relevance of the newly identified peaks in understanding the immune response, we retrieved the ontological descriptions [3, 59] of genes within 10 kbp of a personal-only peak and looked for terms related to immunity and viral infection (Fig 3E). Indeed, we find thousands of personal-only peaks that are in the vicinity of genes described by such terms, including "positive regulation of NF-kappaB", a family of transcription factors involved in regulating immunity [31]. Further, we found that the genes nearby personal-only peaks are functionally enriched in immune biological processes, similarly to genes near common peaks (Fig S9). Therefore, the new peaks refine the picture of the chromatin state surrounding immune genes, especially in regions such as the MHC where genetic variants and altered peaks are denser (Fig S8).

Often, it is useful to identify quantitative trait loci that affect epigenomic features when interpreting the functional impact of variants. A cohort genome graph could affect downstream analyses that estimate read counts to find chromatin accessibility (caQTL) or histone modification (hQTL) quantitative trait loci. To find the extent of this effect, we mapped caQTLs (chromatin accessibility) and hQTLs (H3K4me1 and H3K27ac histone modifications) using reference-based and graph-based read count estimates. Then we checked by how much the model inferences (effect size and p-value) change between the two sets of read count estimates for tens of millions of SNP and peak pairs. While most QTLs remain the same, some do change (Fig 3F, S10). In particular, the estimated effect size changes by 1.4 or more for one QTL in 1000. Similarly, the observed p-value (as $-log(p)$) changes by 6.6 or more for one QTL in 1000. This means that tens of thousands of associations between SNPs and peaks are affected. Therefore, removing reference bias from read count estimates using cohort graphs could improve QTL discovery.

Genome graphs measure epigenomic signal on MEIs

Next, we focus on the MEIs that were assembled and introduced in the cohort genome graph. We were interested in quantifying the number of reads that were assigned to the alternative MEI allele versus the reference allele. We took advantage of the known population genotypes to see how specific read mapping was in these repetitive and polymorphic sequences. We did this by aligning whole genome sequencing reads to the genome graph and regenotyping the MEIs in each sample (Methods). Unlike the single genome benchmark, the cohort graph contains MEIs from more than one sample. If the graph is not able to accurately assign reads to the correct TE copy, we expect many polymorphisms to be covered by false positive alignments. Therefore, each sample should show a large excess of genotyped MEIs from other samples relative to ERVcaller and MELT. Instead, we find that the graph genotypes are highly consistent with the calls made by these tools (Fig 4A). When genotyping, `vg` recapitulates between 1274 to 1563 genotypes per sample and only misses 60 to 95 insertions. It also gains between 110 and 196 insertions per sample. The gained genotypes are not necessarily false positives because some could be explained by an increase in sensitivity. After all, the exact location and sequence of these polymorphisms are known *a priori* by the graph genotyping algorithm but need to be found *de novo* by ERVcaller and MELT.

Having confirmed that read mapping was sufficiently specific, we listed H3K4me1, H3K27ac ChIP-seq and ATAC-seq peaks (Fig 4B, S11A-B) overlapping MEIs for which the samples were heterozygous or homozygous for the insertion. Then, we categorized events depending on how the reads in the peak were partitioned between the reference and the MEI allele (binomial test, Methods). The peaks are labeled either as reference peaks that lie only on the reference allele, biallelic peaks that lie on both alleles or MEI peaks that lie only on the insertion allele. In the entire cohort, we found 692 biallelic peaks in H3K4me1, 229 events in H3K27ac and 502 events in ATAC-seq. Reference peaks also exist, with 1141 events in H3K4me1, 357 events in H3K27ac and 809 events in ATAC-seq. Most exciting, we found 714 MEI peaks in H3K4me1, 191 MEI peaks in H3K27ac and 316 MEI peaks in ATAC-seq. As expected, MEIs that support peaks are mostly AluY or other Alu elements (Fig 4C). Furthermore, some SVA (A, E and F), L1 (L1HS, L1PA2) and ERV (LTR5_Hs) insertions also support peaks.

As an additional negative control, we repeated the same with peaks in MEI loci for which the samples were homozygous reference. Such instances should show little coverage on the MEI allele. Indeed, we find that reads in these peaks are overwhelmingly assigned to the reference allele (Fig S11C-E). This is further evidence that the mapping on the MEI allele is reliable after removing multi-mapped reads.

Finally, we counted how many of these peaks can only be detected with the graph genome and would have been missed by a traditional approach. In total, we tallied 366 H3K4me1, 161 H3k27ac, and 320 ATAC personal-only peaks (Fig 4D) on MEIs. All together, 22.0% of H3K4me1 MEI peaks, 44.5% of H3K27ac MEI peaks and 44.0% of ATAC MEI peaks are personal-only, meaning that they were not detected with the reference genome. On the other hand, MEIs rarely disrupt peaks in the reference graph (Fig S11F). We show one Alu insertion in the graph that supports a H3K4me1 personal-only peak (Fig 4E), and its linear surjection (Fig S12). We predict that the frequency of similar events can only increase with structural variant size.

Population data reveals MEIs that are potential enhancers

So far, we focused on detecting single MEI alleles that support chromatin marks in individual genomes. Next, we want to obtain an overview of MEIs at the level of the entire cohort. This will reveal patterns across multiple samples and between conditions and allow us search for MEIs with chromatin states that are specific to flu-infection. Thus, we found hundreds of instances that support combinations of marks across one or more samples (Fig 5A). From this picture, we summarized the population frequency of peaks and distinguish between those observed in several samples and singletons (Fig S13). In total, we have identified 218, 90 and 210 MEI polymorphisms that support H3K4me1, H3K27ac and ATAC marks respectively. Of these, only 71, 48 and 102 are singletons for each mark respectively. Therefore, we detect a considerable number of MEIs that support marks in more than one sample.

We were also interested to relate the allele frequency of MEIs to how often they are occupied by peaks. Topmost (Fig 5A), we see MEIs with high allele frequency that almost always support a peak. Ranking in the middle are MEIs with high allele frequency that support peaks less often or MEIs with intermediate allele frequency that often support peaks. In the tail, we find common MEIs that rarely support peaks or very rare MEIs that are occupied by peaks. In short, common MEIs are not necessarily always occupied by peaks.

Among H3K4me1 MEI peaks, 58 are unique to flu-infected samples and 41 are unique to non-infected samples (Fig 5B). In H3K27ac, 30 MEI peaks belong only to flu-infected samples and 33 belong only to non-infected samples. And in ATAC, 76 MEI peaks are observed only in flu-infected samples and 75 are observed only in non-infected samples.

By interpreting this population level data, we identify MEIs that carry the epigenomic marks that are characteristic of regions in an active epigenomic state or of enhancer sequences [10, 63, 49, 14] in flu-infected or non-infected conditions (Fig 5C). The most promising examples

would carry a combination of active marks, in addition to the ATAC mark for chromatin accessibility. For example, 56 MEIs support both H3K4me1 and ATAC, a combination which suggests poised enhancers in open chromatin. Similarly, 16 instances show H3K27ac and ATAC, which indicates active enhancers in open chromatin. Even more encouragingly, 26 insertions are marked by H3K4me1, H3K27ac and ATAC, which is stronger evidence for active enhancers in open chromatin. In addition, a sizeable number of loci show a combination of the active histone marks in closed chromatin. Interestingly, there is a noticeable number of SVA elements showing ATAC peaks, second only to Alu elements.

If these MEI peaks have the same regulatory functions as genome-wide peaks of the same mark, they should be similarly located in the genome. When looking at the positions of peaks relative to exons, introns, genes and enhancers, the distributions of MEI and genome-wide peaks are alike (Fig 5D). A large part fall in the introns of genes, with a much smaller proportion in exons. Some of the insertions live in regions already labelled as enhancers in the GeneHancer annotation. Another significant portion are located within 10 kbp of a gene, with the remaining part being inserted in intergenic regions that are farther away. Compared to random positions, genome-wide and MEI peaks skew much closer to genes than expected by chance. Therefore, the MEI peaks we identified are enriched in genes when compared to uniformly sampled positions in the genome.

Again, we checked if any of these genes are related to the immune and viral activity that is expected in this data set. We find dozens of MEI loci that are in the vicinity of genes associated with "viral processes", "immune system processes", the "inflammatory response" and the "positive regulation of NF-kappaB" (Fig S7). For example, an Alu insertion that supports H3K27ac peaks (Fig S14A) in 21 samples is immediately upstream of CD300E, an immune-activating receptor gene [36] associated with "immune system process", "regulation of immune response" and "innate immune response". Importantly, this frequent MEI peak is located within DNase and transcription factor clusters (Fig S14B), which is further evidence for active chromatin. Therefore, as previously found by Chuong et al. [9], some MEIs could be involved in immune response. However, the number of genes is too small to obtain functional enrichment that is statistically significant.

Furthermore, we asked if any MEIs are eQTLs for expressed genes. To this end, we mapped MEI-eQTLs and found 18 MEIs in the flu-infected condition and 34 MEIs in the non-infected condition that are associated with gene expression at a false discovery rate lower than 5×10^{-2}. Of these, 3 MEIs support marks in the flu-infected condition and 9 MEIs support marks in the non-infected condition (Table S1). In the flu-infected condition, we detected an AluYh3 MEI that is positively associated (effect size 0.239, FDR 2.131×10^{-8}) with TRIM25 (Table 2), a gene that restricts influenza RNA synthesis [46] and is involved in immune processes such as "defense response to virus", "interferon-gamma-mediated signaling", "positive regulation of NF-kappaB transcription factor activity", "RIG-I binding" and other terms related to viral infection (see Supplements). This MEI supports H3K27ac peaks in 24 samples, of which 23 are in flu-infected samples and ATAC peaks in 22 samples, of which 16 are in flu-infected samples. It also supports H3K4me1 peaks in 3 flu-infected samples. When viewed in the UCSC Genome Browser, this locus (chr17:54947569) contains DNase and transcription factor clusters.

We show the average read depth at this locus in flu-infected samples that carry the MEI (Fig 6A). To demonstrate the quality of our calls in this locus, we also include the average read depth in flu-infected samples that do not carry the MEI (6B), in non-infected samples that carry the MEI (6C) and in non-infected samples that do not carry the MEI (6D). The more traditional and easier to understand linear projection of the average read depth at this locus confirms that signal is higher in flu-infected samples that carry the MEI when using an accurate genome graph (Fig 6D). Overall, this MEI exists in a flu-specific active chromatin state that could not be detected with the linear reference genome and is a flu-specific eQTL that is associated with TRIM25 and DGKE gene expression 6F).

Discussion

We devised a method to retrieve reads around a MEI from a linked-read library and assemble them into a contig. Depending on the size and family of the transposable element, this method successfully recovers up to a third of mobile elements in a single genome, and up to two thirds when scaling to a population of genomes. We view these rates as lower bounds, since they do not account for the number of false positive calls of ERVcaller and MELT. The resulting sequences were sufficient to create genome graphs that represent non-reference MEIs in the NA12878 genome and in a cohort of 35 ancestrally diverse individuals. Overall, these insertions show the annotations predicted by ERVcaller and MELT. However, we annotate more ERV elements than expected in both NA12878 and the cohort of genomes. Looking closer, the LTR5_Hs ERV subfamily follows the frequency predicted by ERVcaller and MELT. There are also several other ERV subfamilies that occur with very low frequency. These instances are likely due to the miss-annotation or miss-assembly of difficult regions and account for the surplus of ERV annotations. Therefore, manual curation may be needed to remove unreliable annotations.

Simulations in the NA12878 genome demonstrated that adding large structural variants increases the number of true positive mappings. However, the gains depend on the mappability of inserted sequences and the design of paired-end libraries. Structural variants featuring unique sequences are more mappable and should be accessible to short reads. Meanwhile, longer library fragment length improves alignment by resolving ambiguous mappings in larger repeats. Within the repeatome, an excess of low mappability insertions combined with a short library fragment length will lead to poorer mapping compared to a simpler genome graph. The NA12878 MEI genome graph and read libraries are well within these limitations, since we observed multiple ChIP-seq and ATAC-seq peaks on MEIs with good mappability in multiple replicates.

For the cohort of genomes, we successfully recovered a subset of structural variants and encoded it in a genome graph. Since we recovered mostly short Alu mobile element insertions, we avoided the bulk of penalties and still assigned reads to the correct alleles of each genome. Indeed, the sample regenotyping analysis and the low read depth on the MEI allele in homozygous reference samples indicated good sensitivity and specificity. As such, we are confident that the cohort genome graph is able to accurately measure chromatin state on most structural variants. However, estimating read depth in full length L1 insertions remains difficult since they exceed the fragment length of the read library. Our results showed hundreds of loci that support ChIP-seq or ATAC-seq peaks on the MEI allele, often in multiple samples.

More, the cohort genome graph calls thousands of additional peaks that are enriched in variants in each sample. Despite their smaller size, these personal-only peaks are replicated in other samples at rates that are above what is expected by chance. This suggests that cohort genome graphs improve peak calling on loci that often support epigenomic events but that are sensitive to reference bias.

This analysis detected a few hundred MEIs that exist in an active epigenomic state and a smaller subset that are also eQTLs. Of course, the epigenome is different in every cell and tissue type. Indeed, Alu elements that are similarly marked by active histone marks in open chromatin regions were found to be transcribed in various tissues where they act as cell-type specific enhancers [64]. We expect further explorations of other cell types to highlight additional MEIs that are not covered in the current set. In total, there may be many other functional sequences that are hidden by a haploid reference but that could be discovered through a pan-genomic reference [12].

Overall, this approach represents sequence variation and measures epigenomic signals within the same harmonized framework. In the future, genome graphs of complete genomes will give rise to complete epigenomes [27]. Then, chromatin profiling algorithms such as Segway [34] and ChromHMM [19] could automatically find structural variants with regulatory activity. Therefore, we anticipate that genome graphs will become a useful tool in the study of comparative

epigenomics between cell types and individuals in a population.

Methods

Locally assembling read clouds with BarcodeAsm

For the purpose of locally assembling read clouds, we wrote BarcodeAsm [11]. The inputs to BarcodeAsm are a BED file describing the regions to be locally assembled, a position sorted BAM file that was aligned with `lariat` [5] and the same BAM file again, but sorted and indexed by the read barcodes.

BarcodeAsm uses the position sorted BAM file to identify barcodes that are present in the target assembly window. Then it moves to the barcode sorted BAM file to retrieve all the reads that are tagged by the previously identified barcodes.

BarcodeAsm also allows filtering reads recovered from outside the local window by mapping quality. We introduced this feature because we expect reads that belong to novel transposable element insertions to be unmapped or mapped to the wrong copy with very low mapping quality. We found that selecting for such reads decreases the size of the assembly and reduces noise.

Next, the `fermi-lite` [42] library assembles the resulting collection of reads and creates a unitig graph, from which the contigs associated with the assembly window are extracted. Finally, BarcodeAsm aligns the contigs to the local window with `minimap2` [43]. A select set of fermi-lite assembly parameters and `minimap2` alignment parameters are exposed via the command line and can be adjusted to each application.

Reassembling transposable element insertions

ERVcaller [6] and MELT [24] were used to genotype novel insertions of Alu, LINE1, SVA, and ERV transposable elements. To further recover potential insertions within the same type of reference repeats (nested TE insertions), candidate nested insertions that were detected in other public datasets were not removed. For each genotype, local genomic windows were centered on the insertion site. These windows are 800 bp in length for the short Alu elements and 8 kbp for the longer transposable elements. To optimize the outcomes of the assembly, BarcodeAsm was run separately on the short and long insertions with different parameters. For Alus, a minimum read overlap of 30 bp, a maximum mapping quality of 10, and a minimum k-mer frequency of 2 were required. For the larger MEIs, a minimum read overlap of 35, a maximum mapping quality of 20, and a minimum k-mer frequency of 8 were used instead. These parameters were found through a grid search approach and are expected to vary with different data sets.

For the NA12878 benchmark, the SRA accession numbers ERR174324, ERR174325 to ERR174341 were merged in order to genotype MEIs. MEIs were assembled from the public NA12878 linked reads hosted at `https://support.10xgenomics.com/genome-exome/datasets/2.0.0/NA12878_WGS`. Insertions were extracted directly from the BarcodeAsm output using `scripts/alignment_to_vcf.py` to create a VCF file.

For the larger cohort, a consensus sequence approach was used to take advantage of multiple MEI copies in the population. Here, contigs that contain insertions are selected but not immediately used to recover an insertion. Instead, `scripts/msa.py` generates a multiple alignment and calculates a consensus contig for a particular locus. This consensus contig is aligned back to the local window to call the consensus insertion and to create a multi-sample VCF file using `scripts/extract_consensus.py`. For NA12878, the assembled contigs were validated by matching them against its haplotype resolved de novo assembly [25] using `minimap2 -H` [43] and selecting the haplotype with the best mapping quality.

Annotating assembled insertions

To annotate the insertions, RepeatMasker was run using the Dfam [35] database and the longest annotation was selected. For each assembled polymorphism, the distance to the nearest enhancer in the GeneHancer annotation [20] and the distance to the nearest exon in the GENCODE annotation [21] were calculated. The same was done with the MEIs from the 1000 Genome Project [58] and with an equal number of random positions sampled uniformly from the genome. The population structure of the MEI genotypes were compared to the population structure of WGS variants by running principal component analysis with SNPRelate [65].

Creating and benchmarking genome graphs

The genomes graphs were generated with `vg construct` [26] and the corresponding VCF file. For the benchmark genome graph, the NA12878 Platinum callset [17] was merged with the MEI VCF file. For the cohort genome graph, SNPs and indels were called using the LongRanger pipeline [47] independently for each sample were merged with the population MEIs to create a multi-sample VCF file. A Nextflow script to generate the genome graphs from these inputs and the b37 reference genome is found in `pop_graph.nf`. The sensitivity and specificity of the resulting genome graphs was checked by aligning matching WGS data for every sample (downsampling the merged read set by 5x in the case of NA12878) and removing non-specific alignments with `vg filter -r 0.90 -fu -m 1 -q 10 -D 999`. After, the assembled MEI snarls were genotyped with `vg call -m2,4` [33] and the graph genotypes were compared to ERV and MELT genotypes.

To evaluate the impact of the genome graphs on alignment, a read set was simulated from the diploid assembly of NA12878. To achieve this, a genome graph was created from hg19 using the structural variant sequences called by the authors directly from the chromosome scale assembly. Paired-end reads were simulated using `vg sim` from this graph with a fragment size of 2000 bp, to ensure that we can access at least some of the long copy number variants that have low mappability. Then, this read set was aligned to the hg19 reference graph. The alignment was repeated on increasingly complete NA12878 genome graphs, first by including only SNVs and indels, then adding MEIs. Finally, the simulated reads were aligned to their source genome graph, which is the true genome by construction. The previous alignments are compared against this last alignment using `vg gamcompare`.

Evaluating the impact of genome graphs on peaks

In parallel, ChIP-seq and ATAC-seq data was aligned to the b37 reference graph to obtain reference peaks. The reference peaks were intersected with the graph peaks, and categorized into common peaks, personal-only and ref-only peaks. Common peaks are unaltered peaks that are found with both the reference and cohort genome graph. Personal-only peaks and ref-only peaks are altered peaks that are only found in the cohort graph or the reference graph respectively. A logistic regression model including peak width, the presence of an indel and the presence of a SNP was fitted using `cv.glnmnet` [22]. The number of common and altered peaks were balanced by subsampling common peaks. We report the average cross-validation mean coefficients and median AUC. To summarize the population properties of peaks, curves were generated for common, personal-only and ref-only peaks. The width of each peak was fixed to 200 bp and the proportion of samples that have a peak at the same location was calculated. The resulting curve is the inverse cumulative distribution of peak frequencies. A permutation simulation was performed to obtain the inverse cumulative distribution that is expected from random overlaps. In the simulation, a new peak set of the same size is randomly sampled for each individual from the set of all peaks and the overlaps are used to recompute the inverse cumulative distribution. The simulation was run 100 times and the average inverse cumulative distribution is reported.

QTL mapping

We filtered to exclude non-autosomal and non-biallelic variants. Additionally, we removed SNPs and MEIs that had a call rate of <90% across all samples, that deviated from Hardy–Weinberg equilibrium at $p < 10^{-5}$, and with minor allele frequency less than 5%. This resulted in 7,383,243 SNPs and 1222 MEIs used for QTL mapping. We used the R package `MatrixeQTL` [51] to examine the associations between SNP genotypes and chromatin accessibility, H3K4me1 and H3K27ac histone marks using both graph and reference read counts. We also mapped MEI-eQTLs to find association between mobile element insertion genotypes and gene expression counts (derived from alignments to the reference genome with `STAR` [16]). In each case, we calculated age and batch corrected expression matrices. We performed the following analysis in NI and flu samples separately as batch effects may be stronger in flu-infected samples. We calculated normalization factors to scale the raw library sizes using `calcNormFactors` in `edgeR (v 3.28.1)` [50] and used the `voom` function in `limma (v 3.42.2)` [52] to apply these factors to estimate the mean-variance relationship and convert raw read counts to $logCPM$ values. We then fit a model using mean-centered age and admixture, removing batch effects using `ComBat` from the `sva` Bioconductor package [40]. We then regressed out age effects, resulting in the age and batch corrected expression matrices used as inputs for `MatrixeQTL`. To increase the power to detect cis-QTL, we accounted for unmeasured-surrogate confounders by performing principal component analysis (PCA) on the age and batch corrected expression matrices. The number of PCs chosen for each data type empirically led to the identification of the largest QTL in each condition and are reported in the table below.

Analysis	Condition	Regressed PCs
caQTL	Non-infected	1 to 3
	Flu-infected	1 to 3
H3K27acQTL	Non-infected	1
	Flu-infected	1
H3K4me1QTL	Non-infected	1 to 2
	Flu-infected	1 to 4
eQTL	Non-infected	4
	Flu-infected	4

Mapping was performed combining individuals in order to increase power, thus, we included the first eigenvector obtained from a PCA on the SNP and MEI genotype data as a covariate in our linear model. Local associations (i.e., putative cis QTL) were tested against all SNPs and MEIs located within the peak or gene or 100 kbp upstream and downstream of each peak or gene. We recorded the strongest association (minimum p-value) for each peak/gene, which we used as statistical evidence for the presence of at least one QTL for that peak/gene. We permuted the genotypes ten times, re-performed the linear regressions, and recorded the minimum p-value for each peak/gene for each permutation. We used the R package `qvalue` [54] to estimate FDR. In all cases, we assume that alleles affect phenotype in an additive manner.

Functional enrichment analysis

To find if peaks are enriched in functional pathways, the names of genes that are within 10 kbp of a peak were compiled into gene sets. A gene may appear only once in a gene set, even if it is nearby several peaks. This was done separately for the flu-infected and non-infected conditions and for common and personal-only H3K4me1, H3K27ac and ATAC peaks. Gene sets were checked for functional enrichment in Biological Process Gene Ontology (GO:BP) terms

with `gprofiler2` [39]. We also calculated the gene ratio of enriched terms, which is the share of genes in the list associated with a GO term.

Evaluating peaks on genome graphs

H3K27ac, H3K4me1 ChIP-seq and ATAC-seq libraries were aligned to the genome graphs and peaks were called with Graph Peak Caller [30] to obtain graph peaks. All the snarls in the genome graph were genotyped with `vg call -m2,4` in order to partition the reads between the reference and alternative alleles in peaks that overlap a polymorphism. This information is available in the AD (alternate allele depth) and DP (site depth) tags in the VCF output. As before, non-specific alignments were removed. This approach was validated by listing peaks that overlap polymorphic loci in homozygous reference samples to create a negative control peak set, which should not show any coverage on the alternate allele.

Since the MEI allele is longer than the reference allele, we scaled down read counts on the longer allele according to:

$$C_{a,adjusted} = C_a \frac{2L_R}{2L_R + L_a}$$

where C_a denotes the read count on the long allele, L_R the read length and L_a the length of the long allele.

Further, peaks were binned according to the partitioning of reads between the reference and alternative alleles using a two sided binomial test with $\alpha = 0.05$. Peaks that were skewed towards the MEI allele (p-value $\leq \alpha/2$) are placed in the MEI support bin. Peaks that show roughly equal read coverage on both alleles (p-value $\geq \alpha/2$) are labeled as biallelic. When reads are skewed towards the reference allele (p-value $\leq \alpha/2$), the peak belongs to the reference support bin.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

- The source code of BarcodeAsm is available on the Zenodo repository [11].
- Additional code and processed data to reproduce the analysis, figures and manuscript are available on the Zenodo repository [28].
- Code for QTL mapping is available on the Zenodo repository [2].

Competing interests

The authors do not declare any competing interests.

Funding

This work was supported by a Canada Institute of Health Research (CIHR) program grant (CEE-151618) for the McGill Epigenomics Mapping Center, which is part of the Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) Network. GB is supported by a Canada Research Chair Tier 1 award, a FRQ-S, Distinguished Research Scholar award and the Canadian Center for Computational Genomics (C3G) is supported by a Genome Canada Genome Technology Platform grant. This research was enabled in part by support provided by Calcul Quebec and Compute Canada.

Additional files

Figures

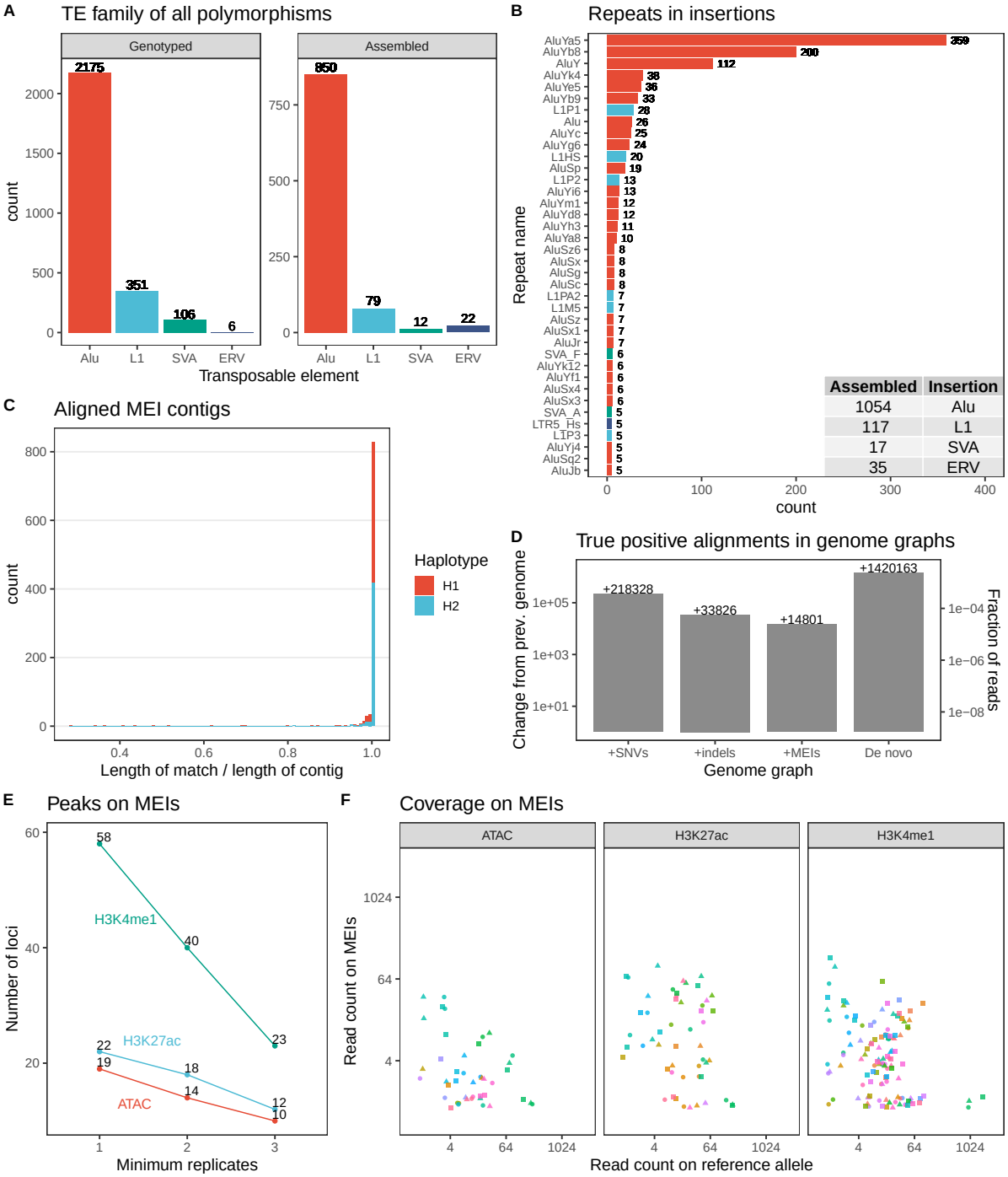

Figure 1: Benchmark in the **NA12878** genome. A) The number and family of MEIs genotyped from short read sequencing data using ERVcaller and MELT compared to the number of loci for which an insertion was recovered. B) The reassembled insertions binned by the annotated repeat name and summarized by repeat family (table). C) The spans of MEI contigs that could be matched and confirmed in the haplotype resolved assembly of NA12878. D) Change in the number of true positive alignments relative to the previous genome in increasingly complete genomes of NA12878, starting with the reference as the baseline. E) Number of MEIs supporting peaks that were called at least once, twice and three times in the replicates. F) Allele coverage at the peak calls that overlap MEIs, stratified by locus (color) and replicate (shape).

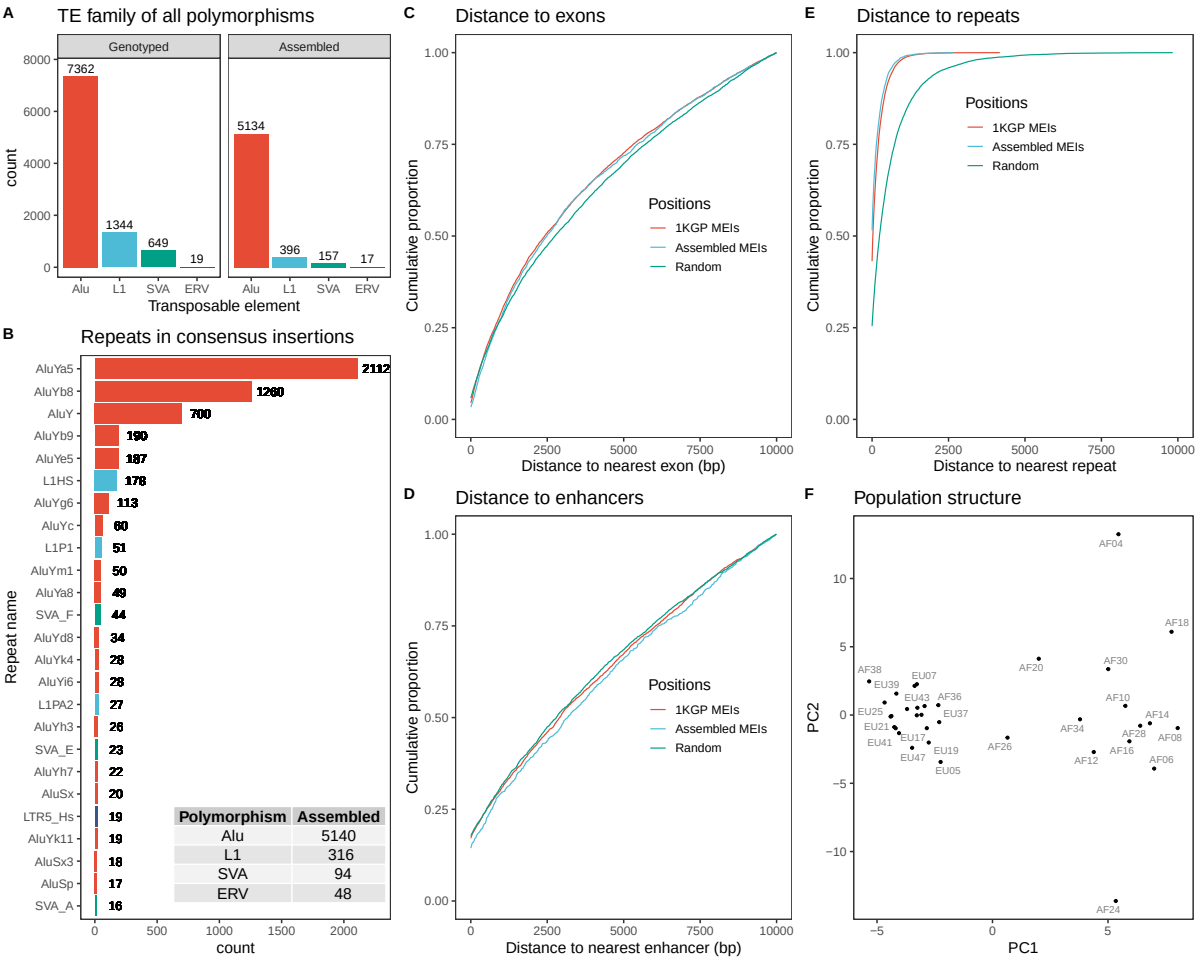

Figure 2: A) The number and family of MEIs genotyped from short read sequencing data using ERVcaller and MELT in the entire cohort. B) The reassembled insertions binned by the annotated repeat name and summarized by repeat family (table) across the cohort. C) The genomic distribution of assembled insertions relative to annotated exons in GENCODE. D) The genomic distribution of the same insertions relative to the annotated enhancers in the GeneHancer annotation. E) Similarly, the genomic distribution relative to repeats annotated in RepeatMasker. F) The observed population structure in MEI genotypes as projected by principal component analysis.

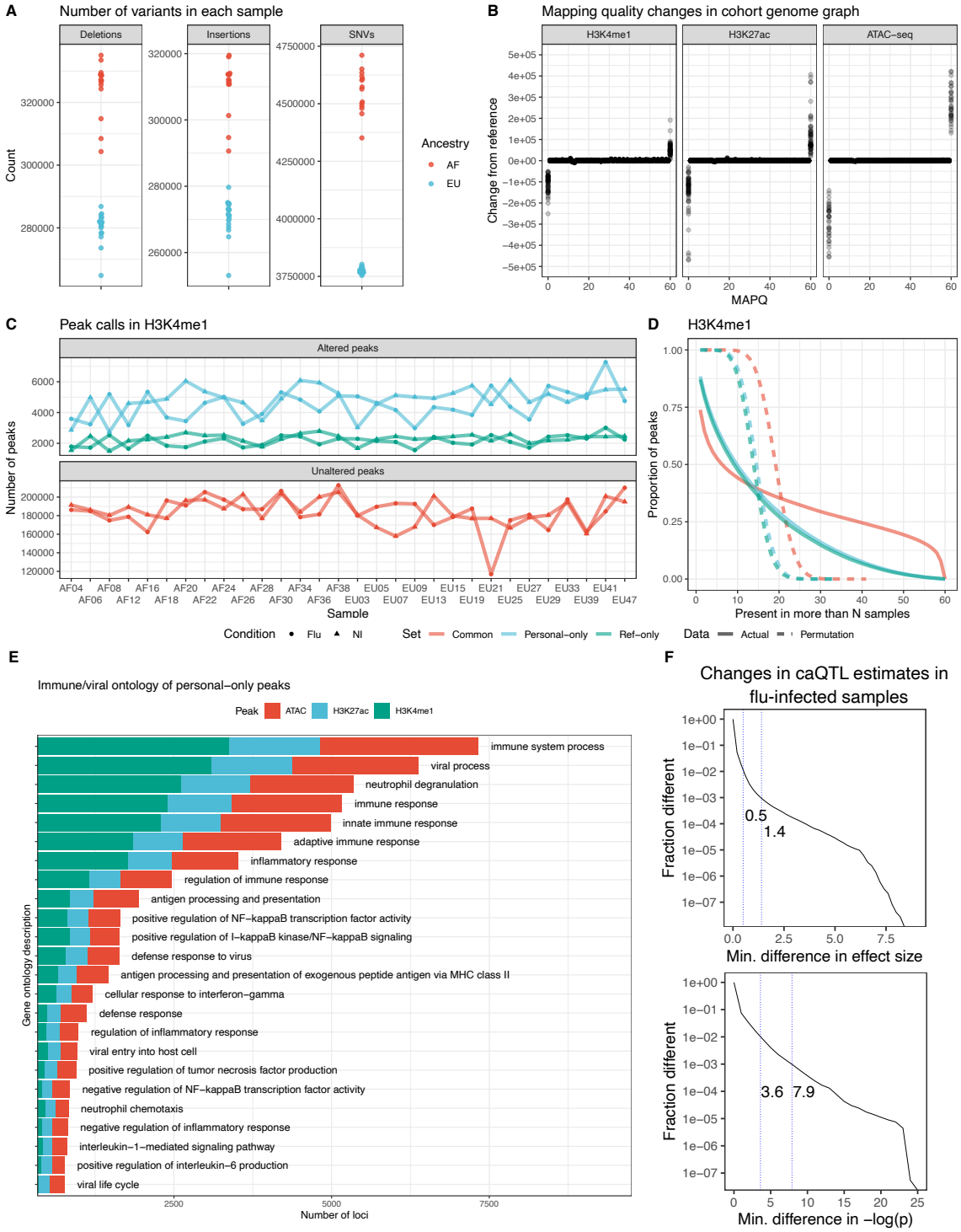

Figure 3: A) Number of deletions, insertions and SNVs in the cohort graph for each sample. B) Number of alignments that change mapping quality between the reference and the cohort genome graph. C) The number of H3K4me1 altered (personal-only, ref-only) and unaltered (common) peaks between the cohort and the reference genome graphs, stratified between flu-infected and non-infected (NI) read sets. D) Inverse cumulative distributions describing how many peaks are observed in more than a number of samples. Curves that are expected by chance are also shown (dashed lines). E) Immune related gene ontology descriptions of genes within 10 kbp of personal-only peaks. F) Inverse cumulative distributions showing the fraction of caQTL effect size and p-value estimates that changed by a minimum amount between the genome graph and the reference.

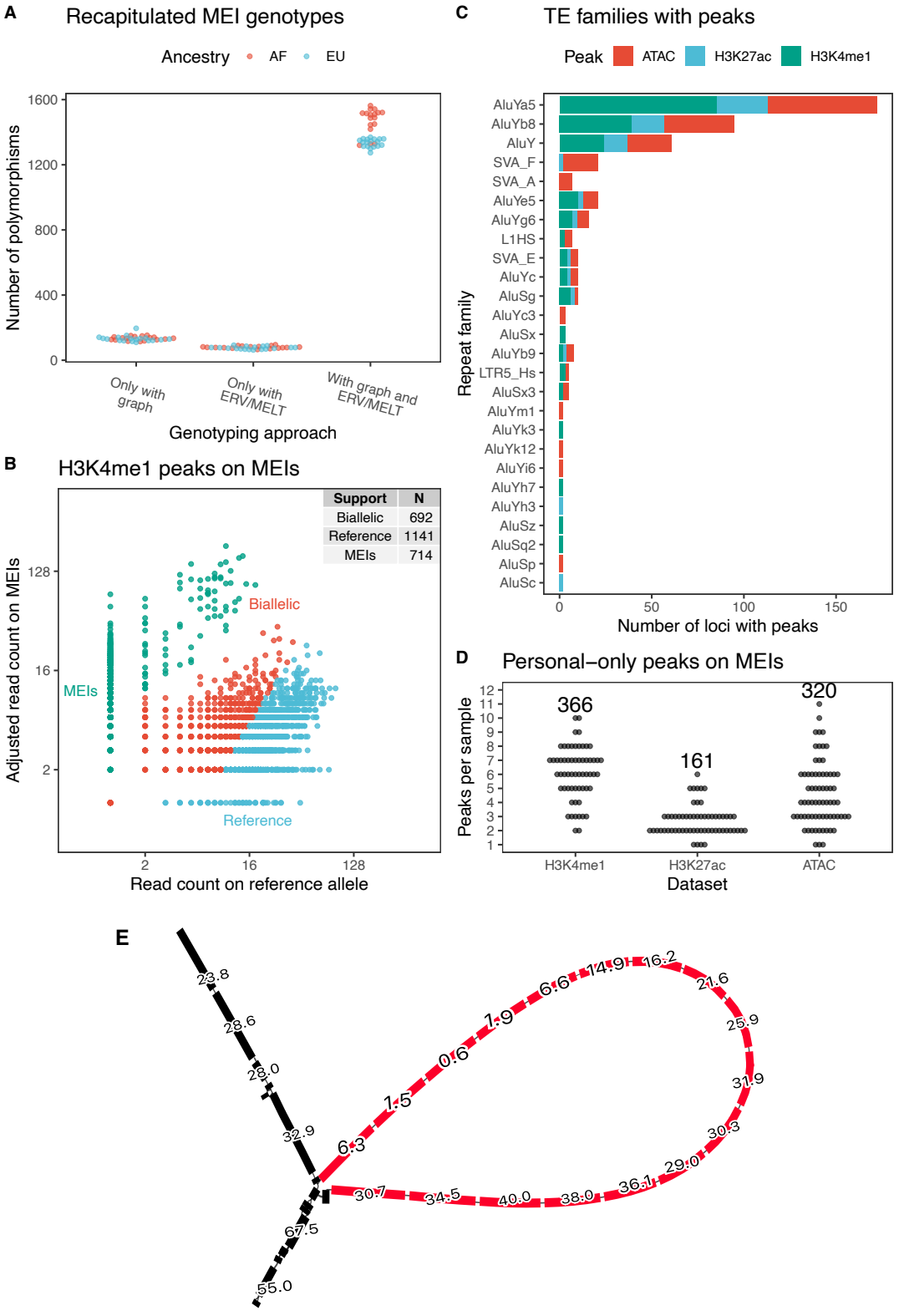

Figure 4: A) Assembled MEIs that were regenotyped using the cohort genome graph. Graph genotypes are compared to the previous genotypes that were called with ERVcaller and MELT. B) Partitioning of reads between the reference and alternative allele in peaks that overlap heterozygous or homozygous MEIs. C) TE families that support peaks in at least one sample. Singletons not shown. D) Number of personal-only peaks that overlap MEIs in each sample. E) A graph region that represents an Alu insertion, annotated by the average read depth on each node within a H3K4me1 peak.

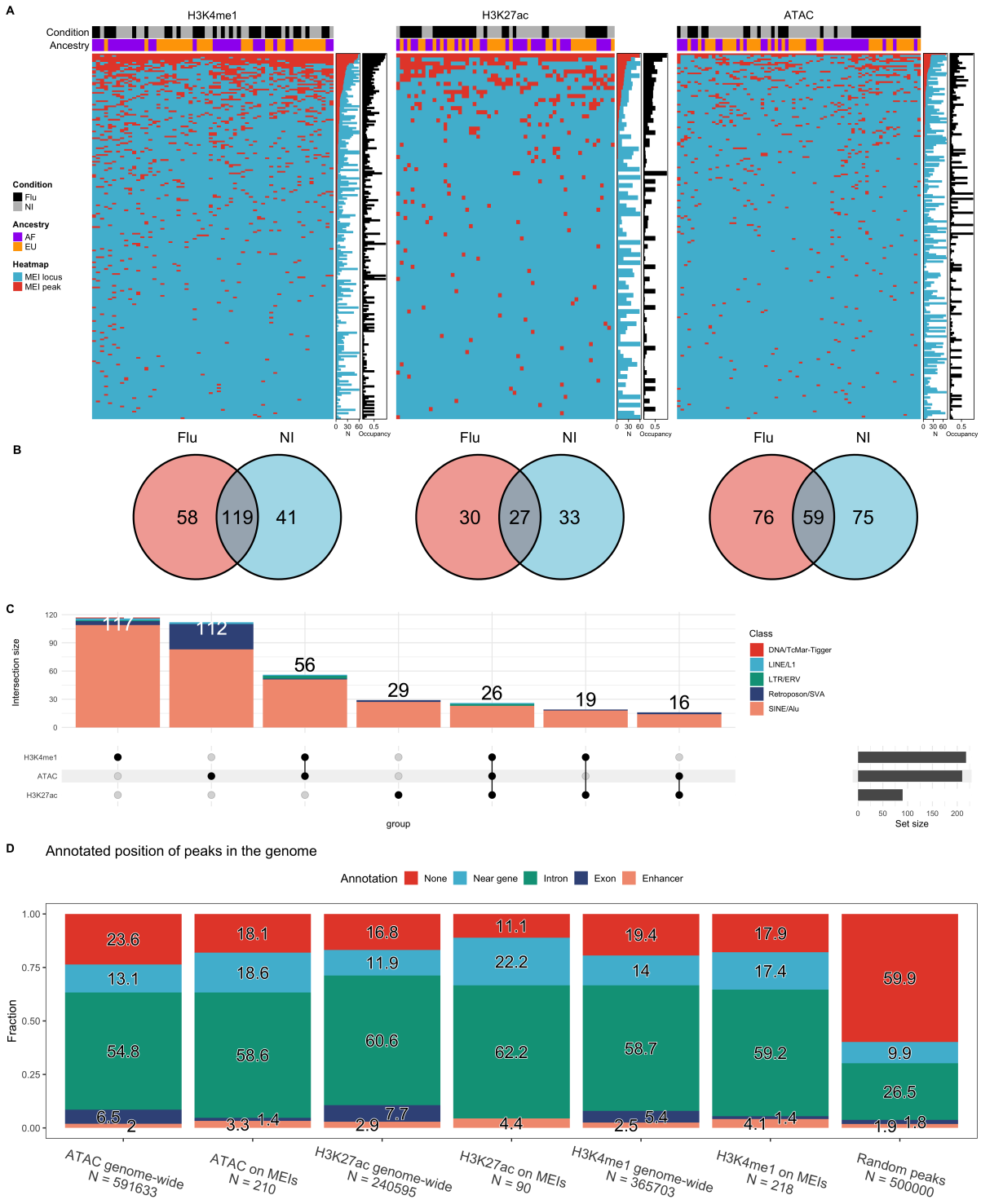

Figure 5: A) Summary of MEIs (rows) that support H3K4me1, H3K27ac and ATAC peaks in the cohort samples (columns). Occupancy is the ratio between samples that support peaks on the MEI (N - red) and those that carry the MEI (N - blue). B) Venn diagrams showing MEI peaks that are shared between flu-infected and non-infected conditions. C) Upset plot describing the number of MEIs that support a combination of H3K4me1, H3K27ac and ATAC peaks, annotated by transposable element class. D) The annotated positions of genome-wide peaks and MEI peaks in the genome, with uniformly and randomly sampled genome positions for comparison. Peaks near genes are within 10 kbp of a gene boundary.

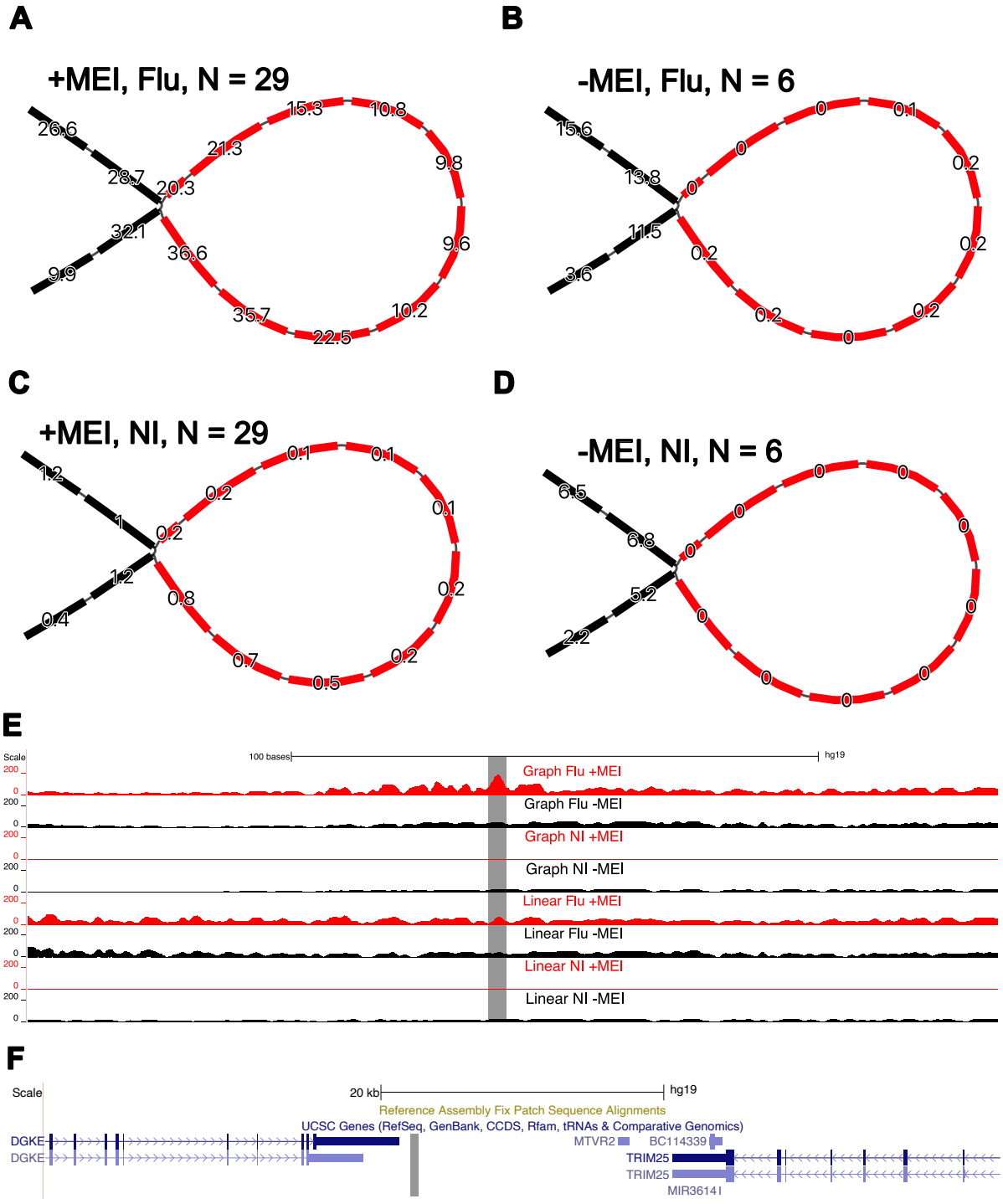

Figure 6: A) Average read depth in the locus of an AluYh3 MEI-eQTL in flu-infected samples that carry the insertion, B) in flu-infected samples that do not carry the MEI, C) in non-infected samples that carry the MEI and D) in non-infected samples that do not carry the MEI. The read depths of homozygous nodes were halved before averaging. Reads below a MAPQ of 10 were not counted. E) A genome browser view of the read depth after projecting alignments onto the linear genome, contrasting the alignments to the graph genome and the reference genome. D) Nearby genes that are associated with this MEI-eQTL (DGKE, TRIM25). The grey strips denote the position of the MEI.

Tables

Peak	intercept	width	indel	snp	AUC
H3K4me1	2.44	−0.72	0.55	0.11	0.91
H3K27ac	2.72	−0.78	0.56	0.12	0.92
ATAC	2.70	−0.87	0.45	0.19	0.90

Table 1: Relative influence of peak width (100bp), SNPs and indels on the log-odds of a peak being found only with the cohort graph genome.

Gene	Effect size	FDR
chr11:47806640 — H3K27ac — AluYh7		
NUP160 - NI	0.278	3.990×10^{-7}
chr17:54947569 — H3K4me1, H3K27ac, ATAC — AluYh3		
DGKE - Flu	0.912	9.559×10^{-16}
TRIM25 - Flu	0.239	2.131×10^{-8}

Table 2: Summary of MEI-eQTLs listing the regressed gene, the biological condition, the effect size, and the false discovery rate. Each entry describes the insertion coordinate, the marks supported by the MEI and the TE family.

References

[1] Alexandre de Andrade, Min Wang, Maria F. Bonaldo, Hehuang Xie, and Marcelo B. Soares. "Genetic and epigenetic variations contributed by Alu retrotransposition". In: *BMC Genomics* 12.1 (Dec. 2011), p. 617. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-617. URL: https://doi.org/10.1186/1471-2164-12-617.

[2] Katie Aracena. *katiearacena/Groza_et_al_mapping: for_preprint*. Sept. 2021. DOI: 10.5281/zenodo.5519627. URL: https://doi.org/10.5281/zenodo.5519627.

[3] M Ashburner et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". eng. In: *Nature genetics* 25.1 (May 2000), pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. URL: https://pubmed.ncbi.nlm.nih.gov/10802651.

[4] O. B. Bantysh and A. A. Buzdin. "Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA". In: *Biochemistry (Moscow)* 74.12 (Dec. 2009), pp. 1393–1399. ISSN: 1608-3040. DOI: 10.1134/S0006297909120153. URL: https://doi.org/10.1134/S0006297909120153.

[5] Alex Bishara, Yuling Liu, Ziming Weng, Dorna Kashef-Haghighi, Daniel E. Neuburger, Robert West, Arend Sidow, and Serafim Batzoglou. "Read clouds uncover variation in complex regions of the human genome". en. In: *Genome Research* (Aug. 2015), gr.191189.115. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.191189.115. URL: http://genome.cshlp.org/content/early/2015/08/18/gr.191189.115 (visited on 07/13/2018).

[6] Xun Chen and Dawei Li. "ERVcaller: identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data". In: *Bioinformatics* 35.20 (Mar. 2019), pp. 3913–3922. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz205. URL: https://doi.org/10.1093/bioinformatics/btz205 (visited on 04/09/2020).

[7] Colby Chiang et al. "The impact of structural variation on human gene expression". In: *Nature Genetics* 49.5 (May 2017), pp. 692–699. ISSN: 1546-1718. DOI: 10.1038/ng.3834. URL: https://doi.org/10.1038/ng.3834.

[8] Chong Chu, Rebeca Borges-Monroy, Vinayak V. Viswanadham, Soohyun Lee, Heng Li, Eunjung Alice Lee, and Peter J. Park. "Comprehensive identification of transposable element insertions using multiple sequencing technologies". In: *Nature Communications* 12.1 (June 2021), p. 3836. ISSN: 2041-1723. DOI: 10.1038/s41467-021-24041-8. URL: https://doi.org/10.1038/s41467-021-24041-8.

[9] Edward B Chuong, Nels C Elde, and Cédric Feschotte. "Regulatory evolution of innate immunity through co-option of endogenous retroviruses". eng. In: *Science (New York, N.Y.)* 351.6277 (Mar. 2016), pp. 1083–1087. ISSN: 1095-9203. DOI: 10.1126/science.aad5497. URL: https://pubmed.ncbi.nlm.nih.gov/26941318.

[10] Menno P. Creyghton et al. "Histone H3K27ac separates active from poised enhancers and predicts developmental state". In: *Proceedings of the National Academy of Sciences* 107.50 (Dec. 2010), p. 21931. DOI: 10.1073/pnas.1016071107. URL: http://www.pnas.org/content/107/50/21931.abstract.

[11] Groza Cristian. *cgroza/BarcodeAsm: publication version.* Sept. 2021. DOI: 10.5281/zenodo.5510086. URL: https://doi.org/10.5281/zenodo.5510086.

[12] Danang Crysnanto, Alexander S. Leonard, Zih-Hua Fang, and Hubert Pausch. "Novel functional sequences uncovered through a bovine multiassembly graph". In: *Proceedings of the National Academy of Sciences* 118.20 (May 2021), e2101056118. DOI: 10.1073/pnas.2101056118. URL: http://www.pnas.org/content/118/20/e2101056118.abstract.

[13] Josquin Daron and R. Keith Slotkin. "EpiTEome: Simultaneous detection of transposable element insertion sites and their DNA methylation levels". In: *Genome Biology* 18.1 (May 2017), p. 91. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1232-0. URL: https://doi.org/10.1186/s13059-017-1232-0.

[14] Aaron C. Daugherty, Robin W. Yeo, Jason D. Buenrostro, William J. Greenleaf, Anshul Kundaje, and Anne Brunet. "Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans". In: *Genome Research* 27.12 (Dec. 2017), pp. 2096–2107. DOI: 10.1101/gr.226233.117. URL: http://genome.cshlp.org/content/27/12/2096.abstract.

[15] Prescott Deininger. "Alu elements: know the SINEs". In: *Genome Biology* 12.12 (Dec. 2011), p. 236. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-12-236. URL: https://doi.org/10.1186/gb-2011-12-12-236.

[16] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. "STAR: ultrafast universal RNA-seq aligner". eng. In: *Bioinformatics (Oxford, England)* 29.1 (Jan. 2013). Edition: 2012/10/25 Publisher: Oxford University Press, pp. 15–21. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts635. URL: https://pubmed.ncbi.nlm.nih.gov/23104886.

[17] Michael A. Eberle et al. "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". In: *Genome Research* 27.1 (Jan. 2017), pp. 157–164. URL: http://genome.cshlp.org/content/27/1/157.abstract.

[18] Peter Ebert et al. "Haplotype-resolved diverse human genomes and integrated analysis of structural variation". In: *Science* (Feb. 2021), eabf7117. DOI: 10.1126/science.abf7117. URL: http://science.sciencemag.org/content/early/2021/02/24/science.abf7117.abstract.

[19] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature Methods* 9.3 (Mar. 2012), pp. 215–216. ISSN: 1548-7105. DOI: 10.1038/nmeth.1906. URL: https://doi.org/10.1038/nmeth.1906.

[20] Simon Fishilevich et al. "GeneHancer: genome-wide integration of enhancers and target genes in GeneCards". eng. In: *Database : the journal of biological databases and curation* 2017 (Jan. 2017). Publisher: Oxford University Press, bax028. ISSN: 1758-0463. DOI: 10.1093/database/bax028. URL: https://pubmed.ncbi.nlm.nih.gov/28605766.

[21] Adam Frankish et al. "GENCODE reference annotation for the human and mouse genomes". In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D766–D773. ISSN: 0305-1048. DOI: 10.1093/nar/gky955. URL: https://doi.org/10.1093/nar/gky955 (visited on 01/03/2021).

[22] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

[23] Sarah Garcia, Stephen Williams, Andrew Wei Xu, Jill Herschleb, Patrick Marks, David Stafford, and Deanna M. Church. "Linked-Read sequencing resolves complex structural variants". en. In: (Dec. 2017). DOI: 10.1101/231662. URL: http://biorxiv.org/lookup/doi/10.1101/231662 (visited on 07/13/2018).

[24] Eugene J. Gardner, Vincent K. Lam, Daniel N. Harris, Nelson T. Chuang, Emma C. Scott, W. Stephen Pittard, Ryan E. Mills, The 1000 Genomes Project Consortium, and Scott E. Devine. "The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology". In: *Genome Research* 27.11 (Nov. 2017), pp. 1916–1929. URL: http://genome.cshlp.org/content/27/11/1916.abstract.

[25] Shilpa Garg et al. "Chromosome-scale, haplotype-resolved assembly of human genomes". In: *Nature Biotechnology* (Dec. 2020). ISSN: 1546-1696. DOI: 10.1038/s41587-020-0711-0. URL: https://doi.org/10.1038/s41587-020-0711-0.

[26] Erik Garrison et al. "Variation graph toolkit improves read mapping by representing genetic variation in the reference". In: *Nature Biotechnology* 36 (Aug. 2018), p. 875. URL: https://doi.org/10.1038/nbt.4227.

[27] Ariel Gershman et al. "Epigenetic Patterns in a Complete Human Genome". In: *bioRxiv* (Jan. 2021), p. 2021.05.26.443420. DOI: 10.1101/2021.05.26.443420. URL: http://biorxiv.org/content/early/2021/05/27/2021.05.26.443420.abstract.

[28] Cristian Groza. *Genome graphs detect human polymorphisms in active epigenomic states during influenza infection: code and processed data.* tex.referencetype: dataset tex.version: 1. Sept. 2021. DOI: 10.5281/zenodo.5534716. URL: https://doi.org/10.5281/zenodo.5534716.

[29] Cristian Groza, Tony Kwan, Nicole Soranzo, Tomi Pastinen, and Guillaume Bourque. "Personalized and graph genomes reveal missing signal in epigenomic data". In: *Genome Biology* 21.1 (May 2020), p. 124. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02038-8. URL: https://doi.org/10.1186/s13059-020-02038-8.

[30] Ivar Grytten, Knut D. Rand, Alexander J. Nederbragt, Geir O. Storvik, Ingrid K. Glad, and Geir K. Sandve. "Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes". In: *PLOS Computational Biology* 15.2 (Feb. 2019). Publisher: Public Library of Science, e1006731. DOI: 10.1371/journal.pcbi.1006731. URL: https://doi.org/10.1371/journal.pcbi.1006731.

[31] M S Hayden, A P West, and S. Ghosh. "NF-κB and the immune response". In: *Oncogene* 25.51 (Oct. 2006), pp. 6758–6780. ISSN: 1476-5594. DOI: 10.1038/sj.onc.1209943. URL: https://doi.org/10.1038/sj.onc.1209943.

[32] Clara Hermant and Maria-Elena Torres-Padilla. "TFs for TEs: the transcription factor repertoire of mammalian transposable elements". In: *Genes & Development* 35.1-2 (Jan. 2021), pp. 22–39. DOI: 10.1101/gad.344473.120. URL: http://genesdev.cshlp.org/content/35/1-2/22.abstract.

[33] Glenn Hickey, David Heller, Jean Monlong, Jonas A. Sibbesen, Jouni Sirén, Jordan Eizenga, Eric T. Dawson, Erik Garrison, Adam M. Novak, and Benedict Paten. "Genotyping structural variants in pangenome graphs using the vg toolkit". In: *Genome Biology* 21.1 (Feb. 2020), p. 35. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1941-7. URL: https://doi.org/10.1186/s13059-020-1941-7.

[34] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. "Unsupervised pattern discovery in human chromatin structure through genomic segmentation". In: *Nature Methods* 9.5 (May 2012), pp. 473–476. ISSN: 1548-7105. DOI: 10.1038/nmeth.1937. URL: https://doi.org/10.1038/nmeth.1937.

[35] Robert Hubley, Robert D Finn, Jody Clements, Sean R Eddy, Thomas A Jones, Weidong Bao, Arian F A Smit, and Travis J Wheeler. "The Dfam database of repetitive DNA families". eng. In: *Nucleic acids research* 44.D1 (Jan. 2016). Edition: 2015/11/26 Publisher: Oxford University Press, pp. D81–D89. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1272. URL: https://pubmed.ncbi.nlm.nih.gov/26612867.

[36] Masamichi Isobe et al. "The CD300e molecule in mice is an immune-activating receptor". In: *Journal of Biological Chemistry* 293.10 (Mar. 2018). Publisher: Elsevier, pp. 3793–3805. ISSN: 0021-9258. DOI: 10.1074/jbc.RA117.000696. URL: https://doi.org/10.1074/jbc.RA117.000696 (visited on 07/22/2021).

[37] Songmi Kim, Chun-Sung Cho, Kyudong Han, and Jungnam Lee. "Structural Variation of Alu Element and Human Disease". eng. In: *Genomics & informatics* 14.3 (Sept. 2016). Edition: 2016/09/30 Publisher: Korea Genome Organization, pp. 70–77. ISSN: 1598-866X. DOI: 10.5808/GI.2016.14.3.70. URL: https://pubmed.ncbi.nlm.nih.gov/27729835.

[38] Jacob O Kitzman. "Haplotypes drop by drop". In: *Nature Biotechnology* 34.3 (Mar. 2016), pp. 296–298. ISSN: 1546-1696. DOI: 10.1038/nbt.3500. URL: https://doi.org/10.1038/nbt.3500.

[39] L Kolberg, U Raudvere, I Kuzmin, J Vilo, and H Peterson. "gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler [version 2; peer review: 2 approved]". In: *F1000Research* 9.709 (2020). DOI: 10.12688/f1000research.24956.2.

[40] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". eng. In: *Bioinformatics (Oxford, England)* 28.6 (Mar. 2012). Edition: 2012/01/17 Publisher: Oxford University Press, pp. 882–883. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts034. URL: https://pubmed.ncbi.nlm.nih.gov/22257669.

[41] Adrien Leger et al. "Genomic variations and epigenomic landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel". In: *bioRxiv* (Jan. 2021), p. 2021.05.17.444424. DOI: 10.1101/2021.05.17.444424. URL: http://biorxiv.org/content/early/2021/05/17/2021.05.17.444424.abstract.

[42] Heng Li. "Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly". In: *Bioinformatics* 28.14 (May 2012), pp. 1838–1844. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts280. URL: https://doi.org/10.1093/bioinformatics/bts280 (visited on 07/24/2020).

[43] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (May 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191. URL: https://doi.org/10.1093/bioinformatics/bty191 (visited on 07/16/2020).

[44] Patrick Marks et al. "Resolving the Full Spectrum of Human Genome Variation using Linked-Reads". en. In: (Jan. 2018). DOI: 10.1101/230946. URL: http://biorxiv.org/lookup/doi/10.1101/230946 (visited on 07/13/2018).

[45] Dmitry Meleshko, Patrick Marks, Stephen Williams, and Iman Hajirasouliha. "Detection and assembly of novel sequence insertions using Linked-Read technology". In: *bioRxiv* (Jan. 2019), p. 551028. DOI: 10.1101/551028. URL: http://biorxiv.org/content/early/2019/02/15/551028.abstract.

[46] Nicholas R Meyerson, Ligang Zhou, Yusong R Guo, Chen Zhao, Yizhi J Tao, Robert M Krug, and Sara L Sawyer. "Nuclear TRIM25 Specifically Targets Influenza Virus Ribonucleoproteins to Block the Onset of RNA Chain Elongation". eng. In: *Cell host & microbe* 22.5 (Nov. 2017). Edition: 2017/11/05, 627–638.e7. ISSN: 1934-6069. DOI: 10.1016/j.chom.2017.10.003. URL: https://pubmed.ncbi.nlm.nih.gov/29107643.

[47] Alina Ott, James C Schnable, Cheng-Ting Yeh, Linjiang Wu, Chao Liu, Heng-Cheng Hu, Clifton L Dalgard, Soumik Sarkar, and Patrick S Schnable. "Linked read technology for assembling large complex and polyploid genomes". eng. In: *BMC genomics* 19.1 (Sept. 2018). Publisher: BioMed Central, pp. 651–651. ISSN: 1471-2164. DOI: 10.1186/s12864-018-5040-z. URL: https://pubmed.ncbi.nlm.nih.gov/30180802.

[48] Paz Polak and Eytan Domany. "Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes". In: *BMC Genomics* 7.1 (June 2006), p. 133. ISSN: 1471-2164. DOI: 10.1186/1471-2164-7-133. URL: https://doi.org/10.1186/1471-2164-7-133.

[49] Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha A. Brugmann, Ryan A. Flynn, and Joanna Wysocka. "A unique chromatin signature uncovers early developmental enhancers in humans". In: *Nature* 470.7333 (Feb. 2011), pp. 279–283. ISSN: 1476-4687. DOI: 10.1038/nature09692. URL: https://doi.org/10.1038/nature09692.

[50] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." In: *Bioinformatics (Oxford, England)* 26.1 (Jan. 2010), pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616. URL: http://dx.doi.org/10.1093/bioinformatics/btp616.

[51] Andrey A Shabalin. "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". eng. In: *Bioinformatics (Oxford, England)* 28.10 (May 2012). Edition: 2012/04/06 Publisher: Oxford University Press, pp. 1353–1358. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts163. URL: https://pubmed.ncbi.nlm.nih.gov/22492648.

[52] Gordon K Smyth, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber. "Limma: linear models for microarray data". In: *Bioinformatics and computational biology solutions using r and bioconductor*. New York: Springer, 2005, pp. 397–420.

[53] Noah Spies, Ziming Weng, Alex Bishara, Jennifer McDaniel, David Catoe, Justin M Zook, Marc Salit, Robert B West, Serafim Batzoglou, and Arend Sidow. "Genome-wide reconstruction of complex structural variants using read clouds". In: *Nature methods* 14.9 (Sept. 2017), pp. 915–920. ISSN: 1548-7091. DOI: 10.1038/nmeth.4366. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5578891/.

[54] John D. Storey, Andrew J. Bass, Alan Dabney, and David Robinson. *qvalue: Q-value estimation for false discovery rate control*. manual. 2021. URL: http://github.com/jdstorey/qvalue.

[55] Ming Su, Dali Han, Jerome Boyd-Kirkup, Xiaoming Yu, and Jing-Dong J. Han. "Evolution of Alu Elements toward Enhancers". In: *Cell Reports* 7.2 (Apr. 2014), pp. 376–385. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2014.03.011. URL: http://www.sciencedirect.com/science/article/pii/S2211124714001892.

[56] Peter H. Sudmant et al. "An integrated map of structural variation in 2,504 human genomes". In: *Nature* 526.7571 (Oct. 2015), pp. 75–81. ISSN: 1476-4687. DOI: 10.1038/nature15394. URL: https://doi.org/10.1038/nature15394.

[57] Darren Taylor, Robert Lowe, Claude Philippe, Kevin C. L. Cheng, Gael Cristofari, and Miguel R. Branco. "Locus-specific chromatin profiling of evolutionarily young transposable elements". In: *bioRxiv* (Jan. 2021), p. 2021.08.25.457666. DOI: 10.1101/2021.08.25.457666. URL: http://biorxiv.org/content/early/2021/08/27/2021.08.25.457666.abstract.

[58] The 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526 (Sept. 2015), p. 68. URL: https://doi.org/10.1038/nature15393.

[59] "The Gene Ontology resource: enriching a GOld mine." eng. In: *Nucleic acids research* 49.D1 (Jan. 2021), pp. D325–D334. ISSN: 1362-4962 0305-1048. DOI: 10.1093/nar/gkaa1113.

[60] Marco Trizzino, YoSon Park, Marcia Holsbach-Beltrame, Katherine Aracena, Katelyn Mika, Minal Caliskan, George H. Perry, Vincent J. Lynch, and Christopher D. Brown. "Transposable elements are the primary source of novelty in primate gene regulation". In: *Genome Research* 27.10 (Oct. 2017), pp. 1623–1633. DOI: 10.1101/gr.218149.116. URL: http://genome.cshlp.org/content/27/10/1623.abstract.

[61] Neil I. Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M. Church, and David B. Jaffe. "Direct determination of diploid genome sequences". In: *Genome Research* 27.5 (2017). _eprint: http://genome.cshlp.org/content/27/5/757.full.pdf+html, pp. 757–767. DOI: 10.1101/gr.214874.116. URL: http://genome.cshlp.org/content/27/5/757.abstract.

[62] Julia H Wildschutte, Alayna Baron, Nicolette M Diroff, and Jeffrey M Kidd. "Discovery and characterization of Alu repeat sequences via precise local read assembly". eng. In: *Nucleic acids research* 43.21 (Dec. 2015). Edition: 2015/10/25 Publisher: Oxford University Press, pp. 10292–10307. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1089. URL: https://pubmed.ncbi.nlm.nih.gov/26503250.

[63] Gabriel E Zentner, Paul J Tesar, and Peter C Scacheri. "Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions". eng. In: *Genome research* 21.8 (Aug. 2011). Edition: 2011/06/01 Publisher: Cold Spring Harbor Laboratory Press, pp. 1273–1283. ISSN: 1549-5469. DOI: 10.1101/gr.122382.111. URL: https://pubmed.ncbi.nlm.nih.gov/21632746.

[64] Xiao-Ou Zhang, Thomas R. Gingeras, and Zhiping Weng. "Genome-wide analysis of polymerase III–transcribed Alu elements suggests cell-type–specific enhancer function". In: *Genome Research* 29.9 (Sept. 2019), pp. 1402–1414. DOI: 10.1101/gr.249789.119. URL: http://genome.cshlp.org/content/29/9/1402.abstract.

[65] Xiuwen Zheng, David Levine, Jess Shen, Stephanie M. Gogarten, Cathy Laurie, and Bruce S. Weir. "A high-performance computing toolset for relatedness and principal component analysis of SNP data". In: *Bioinformatics* 28.24 (Dec. 2012), pp. 3326–3328. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts606. URL: https://doi.org/10.1093/bioinformatics/bts606 (visited on 01/02/2021).

[66] Xiaoyu Zhuo, Alan Y. Du, Erica C. Pehrsson, Daofeng Li, and Ting Wang. "Epigenomic differences in the human and chimpanzee genomes are associated with structural variation". In: *Genome Research* (Dec. 2020). DOI: 10.1101/gr.263491.120. URL: http://genome.cshlp.org/content/early/2021/01/15/gr.263491.120.abstract.

Supplementary Tables

MEI-eQTL	Condition	Marks	Gene	Beta	FDR
chr1:160047665	NI	H3K4me1, ATAC	ENSG00000177807	1.02	4.10×10^{-2}
chr5:1812640	NI	H3K4me1	ENSG00000171421	0.17	2.48×10^{-3}
chr5:78442870	Flu	H3K4me1	ENSG00000152409	0.26	4.01×10^{-2}
chr7:128209146	NI	H3K4me1, ATAC	ENSG00000242588	0.15	1.13×10^{-3}
chr7:128217333	NI	ATAC	ENSG00000242588	0.16	3.87×10^{-4}
chr8:42039896	NI	H3K4me1, H3K27ac, ATAC	ENSG00000070718	0.15	4.12×10^{-3}
chr11:47806640	NI	H3K27ac	ENSG00000030066	0.28	3.99×10^{-7}
chr11:47806640	NI	H3K27ac	ENSG00000252874	0.41	1.06×10^{-2}
chr12:124066475	NI	H3K4me1, ATAC	ENSG00000111364	−0.19	1.13×10^{-3}
chr14:92586933	NI	H3K4me1	ENSG00000183648	−0.08	4.44×10^{-2}
chr14:92586933	Flu	H3K4me1	ENSG00000165934	−0.15	9.45×10^{-4}
chr15:41100318	NI	H3K4me1, H3K27ac, ATAC	ENSG00000166140	−0.11	9.58×10^{-3}
chr17:54947569	Flu	H3K4me1, H3K27ac	ENSG00000153933	0.91	9.56×10^{-16}
chr17:54947569	Flu	H3K4me1, H3K27ac	ENSG00000214226	0.44	9.69×10^{-9}
chr17:54947569	Flu	H3K4me1, H3K27ac	ENSG00000121060	0.24	2.13×10^{-8}
chr17:54947569	Flu	ATAC	ENSG00000153933	0.91	9.56×10^{-16}
chr17:54947569	Flu	ATAC	ENSG00000214226	0.44	9.69×10^{-9}
chr17:54947569	Flu	ATAC	ENSG00000121060	0.24	2.13×10^{-8}
chr20:13704145	NI	H3K4me1, ATAC	ENSG00000101247	0.14	3.69×10^{-4}

Table S1: MEI-eQTLs that support epigenomic marks and are associated with gene expression at a false discovery rate below 5×10^{-2} in the flu-infected or the non-infected condition.

Gene Ontology: Biological Process immune related terms of MEI-eQTL genes

- **NUP160** – intracellular transport of virus, viral life cycle, viral process, viral transcription
- **DGKE** – platelet activation
- **TRIM25** – defense response to virus, immune system process, innate immune response, interferon-gamma-mediated signaling pathway, negative regulation of type I interferon production, negative regulation of viral entry into host cell, negative regulation of viral release from host cell, positive regulation of I-kappaB kinase/NF-kappaB signaling, positive regulation of NF-kappaB transcription factor activity, regulation of viral entry into host cell, regulation of viral release from host cell, RIG-I binding, viral process

Supplementary Figures

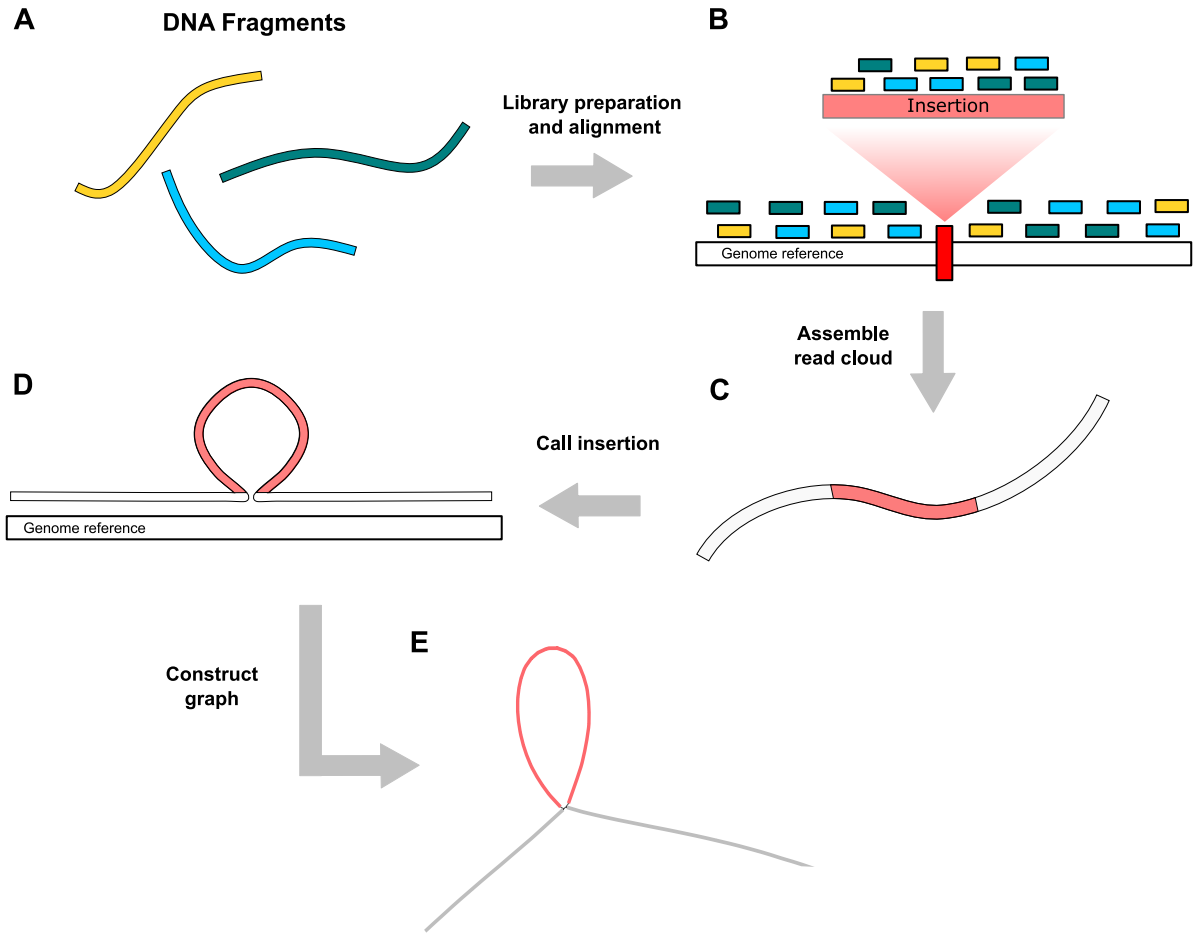

Figure S1: A) DNA fragments that are associated with barcodes (yellow, blue, green). B) Throughout the sequencing library protocol and alignment to the reference genome with lariat, the short reads remain associated with the barcode of the source fragment. We enumerate the barcodes observed around an insertion site. C) We assemble the reads tagged by the previously enumerated barcodes site with fermi-lite to obtain a contig that covers the insertion (red). D) We realign the contig back to the genomic locus to identify the boundaries of the inserted sequence. E) We augment the genome graph with a bubble that represents the assembled insertion.

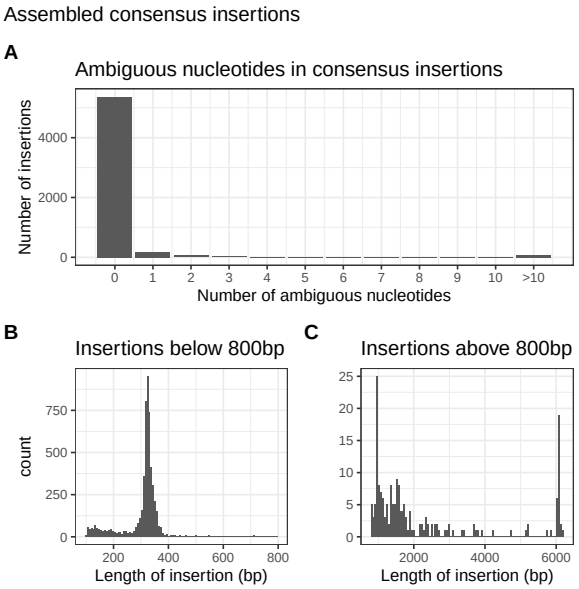

Figure S2: A) The distribution of ambiguous nucleotides in the sequence of each consensus insertion. B) C) The length distribution of insertions assembled through the population consensus approach.

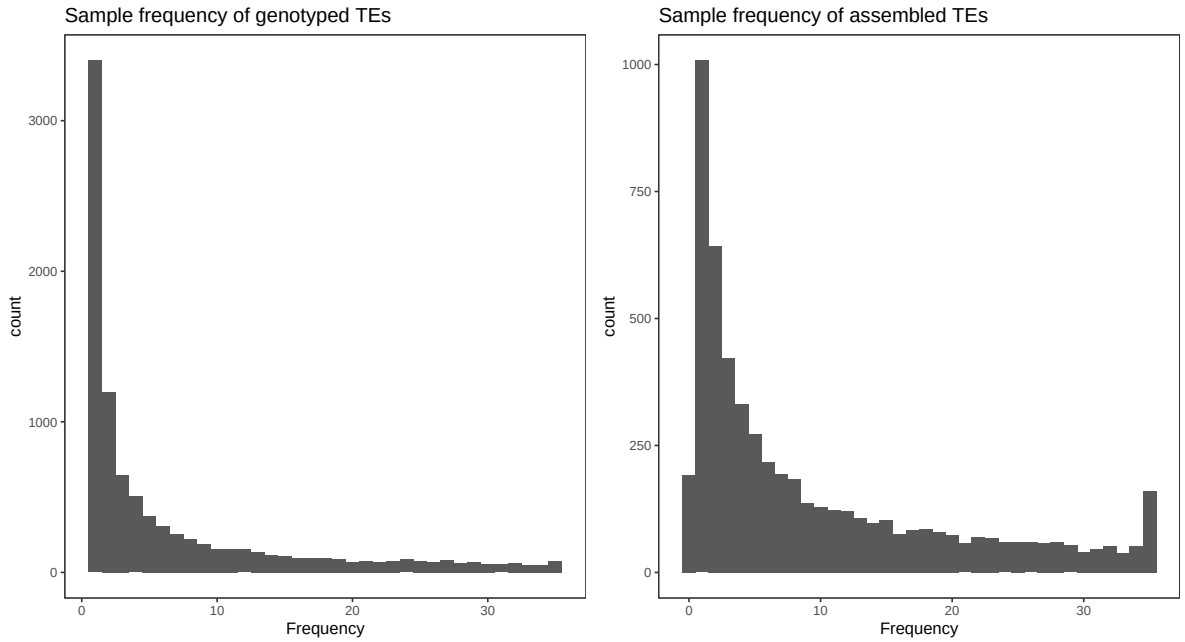

Figure S3: Population frequency of the A) genotyped and B) assembled MEIs in the cohort.

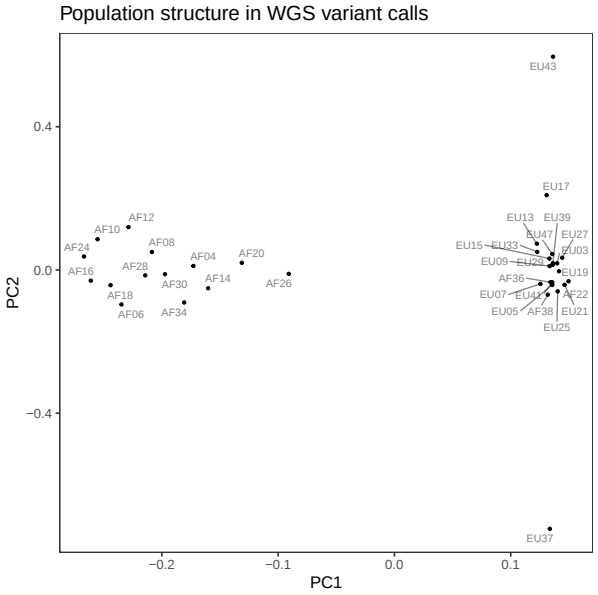

Figure S4: PCA projection showing the population structure observed in whole genome sequencing genotypes.

Figure S5: Frequency of altered peak calls as percentages of all peaks in the sample.

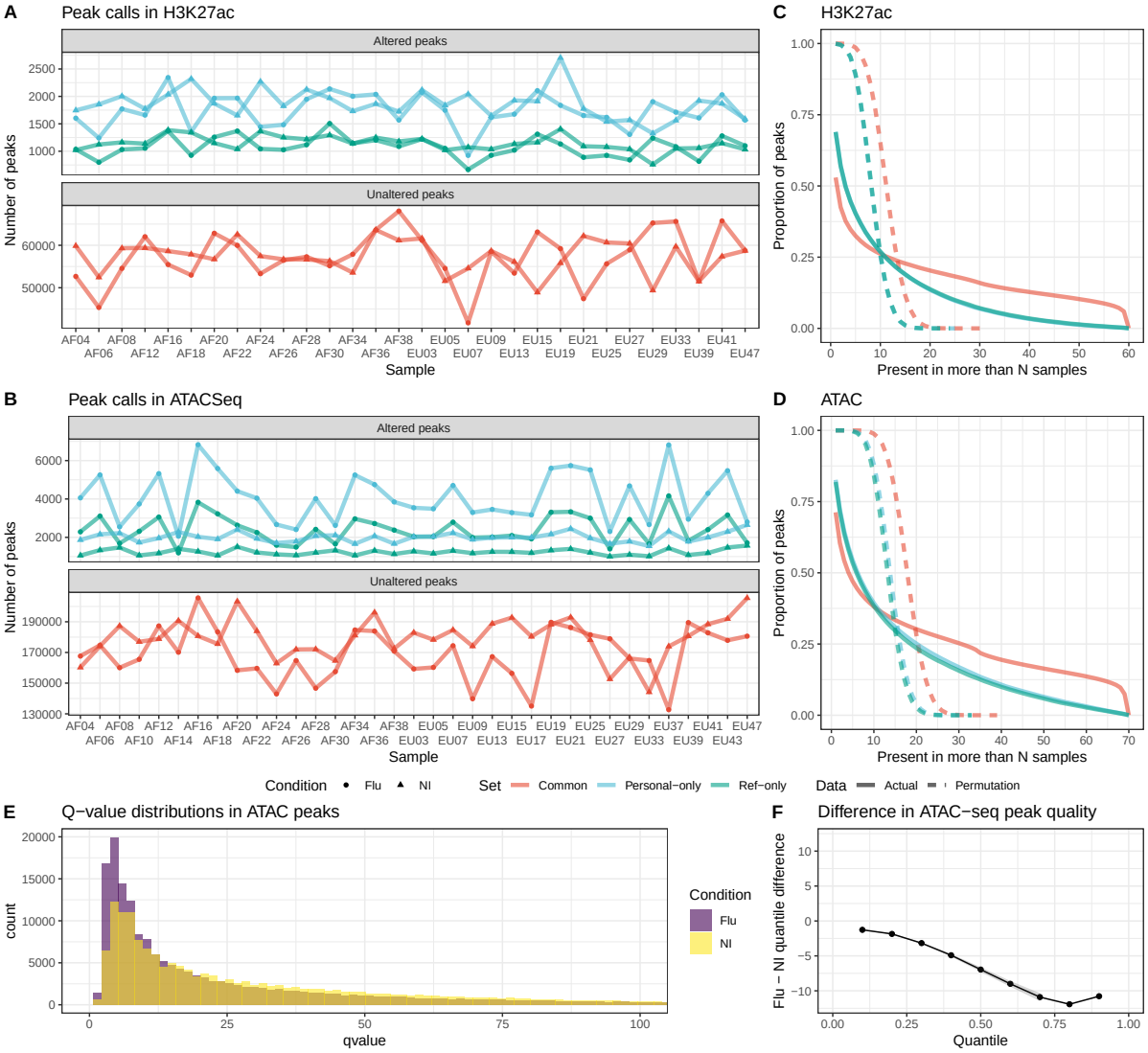

Figure S6: A) B) The number of altered (personal-only, ref-only) and unaltered (common) peaks between the cohort and the reference genome graphs for H3K27ac and ATAC-seq. C) D) Inverse cumulative distributions describing how many H3K27ac and ATAC-seq peaks are observed in more than a number of samples E) Q-value distribution of ATAC peaks in flu-infected and non-infected samples. F) The shift function of the two distributions, describing the difference in q-values at the same quantiles.

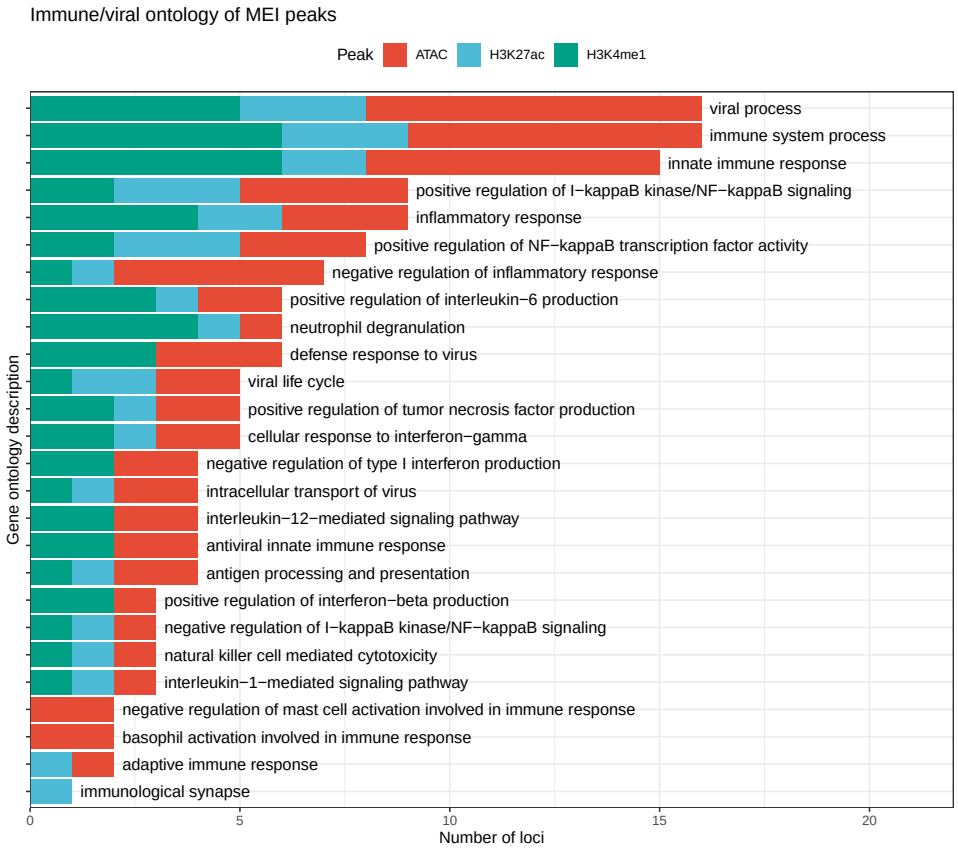

Figure S7: Top ranking GO terms related to immunity and viral infection of genes that are within 10 kbp of MEI peaks.

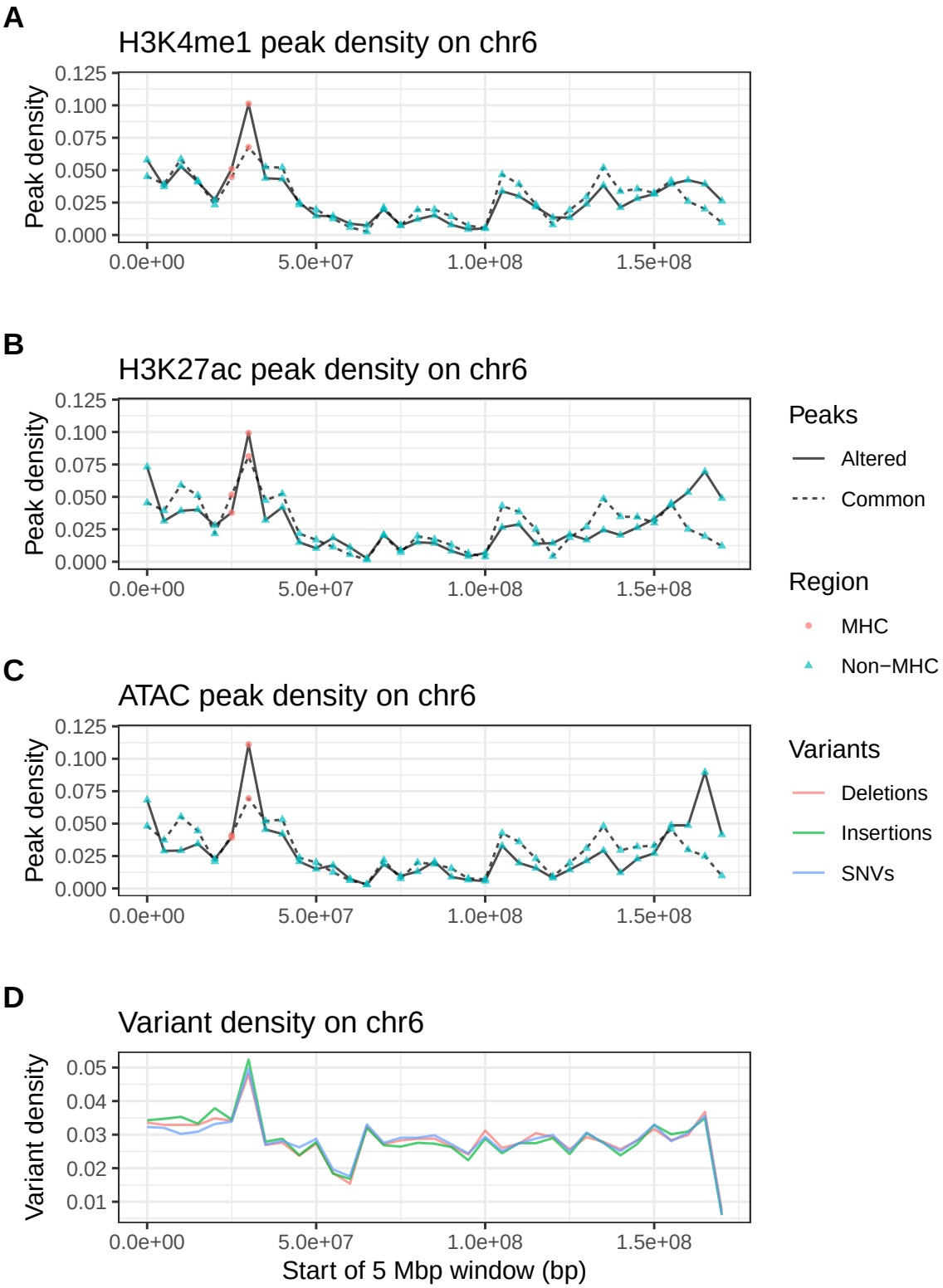

Figure S8: A) B) C) Density of common and altered peak calls along chr6, with the MHC locus highlighted. D) In parallel, density of insertions, deletions and SNVs.

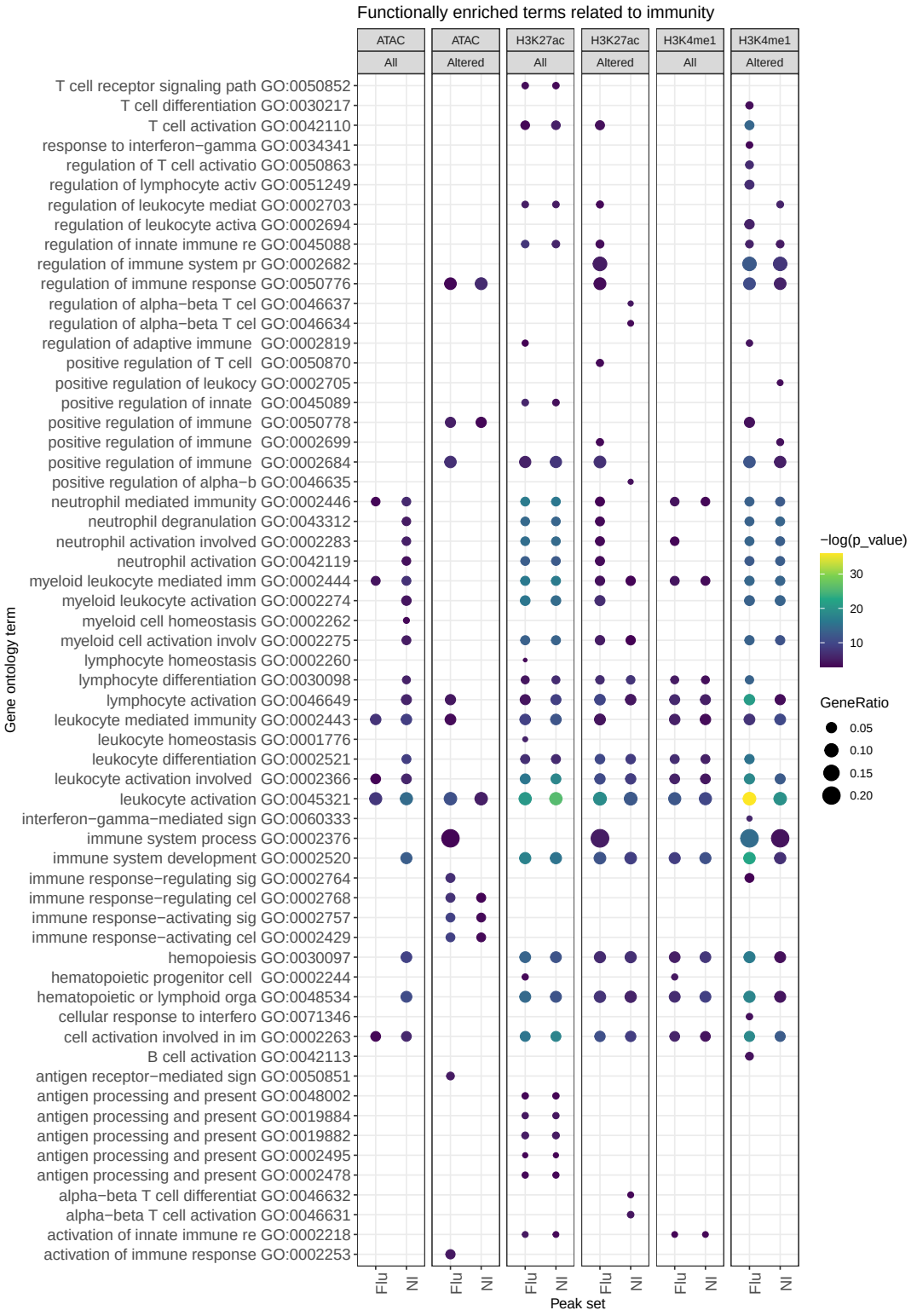

Figure S9: Functional enrichment of genes that are within 10 kbp of H3K4me1, H3K27ac and ATAC common and altered (personal-only) peaks. We show only the enriched gene ontology terms that relate to immunity and viral infection. The gene ratio is the share of genes that are associated with a GO term.

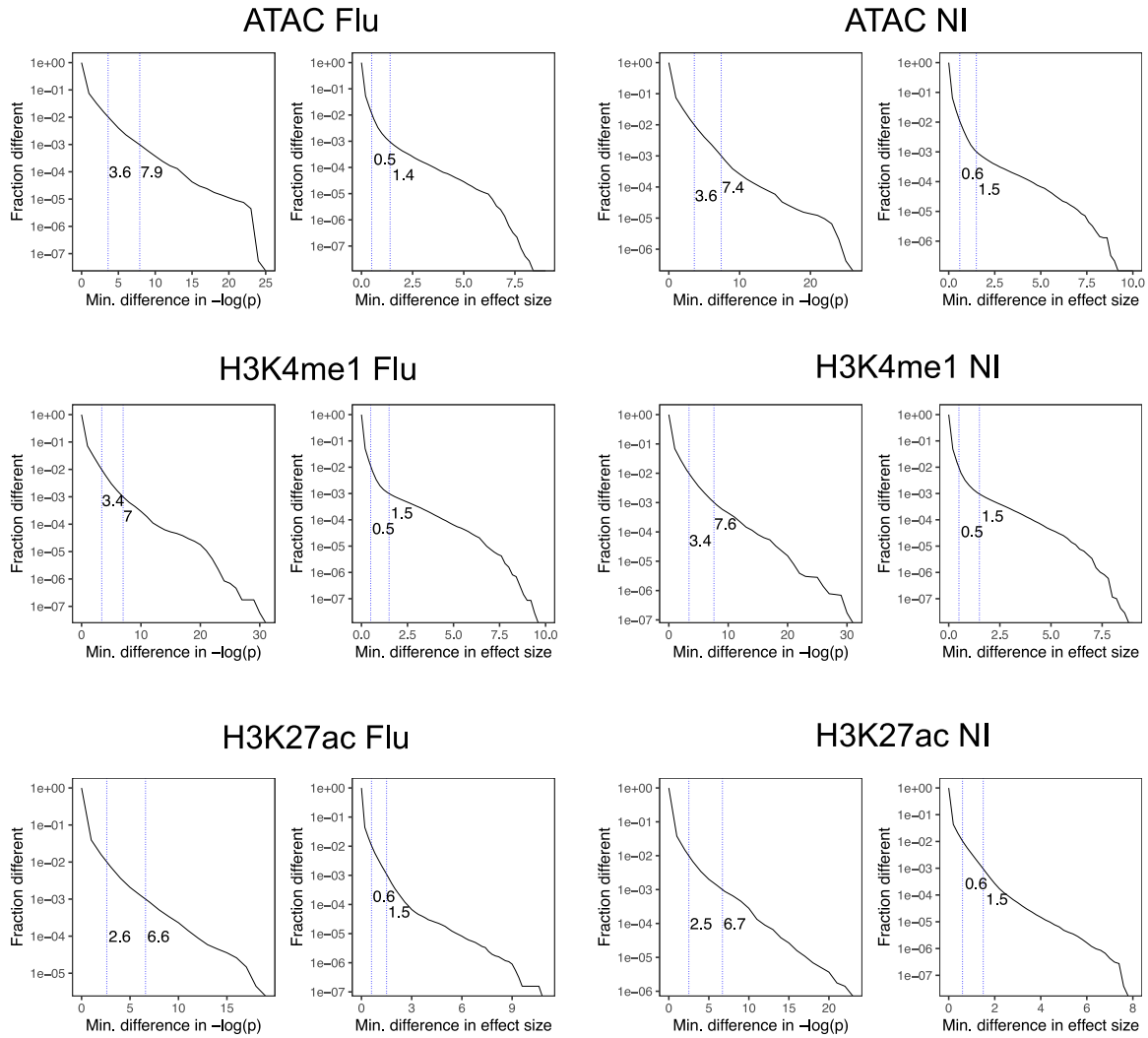

Figure S10: ATAC, H3K4me1, H3K27ac quantitative trait loci are estimated using counts derived from reference and cohort graph alignments in flu-infected and non-infected conditions. We show the fraction of QTLs (y-axis) for which the p-values or effect sizes change by a minimum amount (x-axis). The first line marks the 99th percentile and the second line marks the 99.9th percentile.

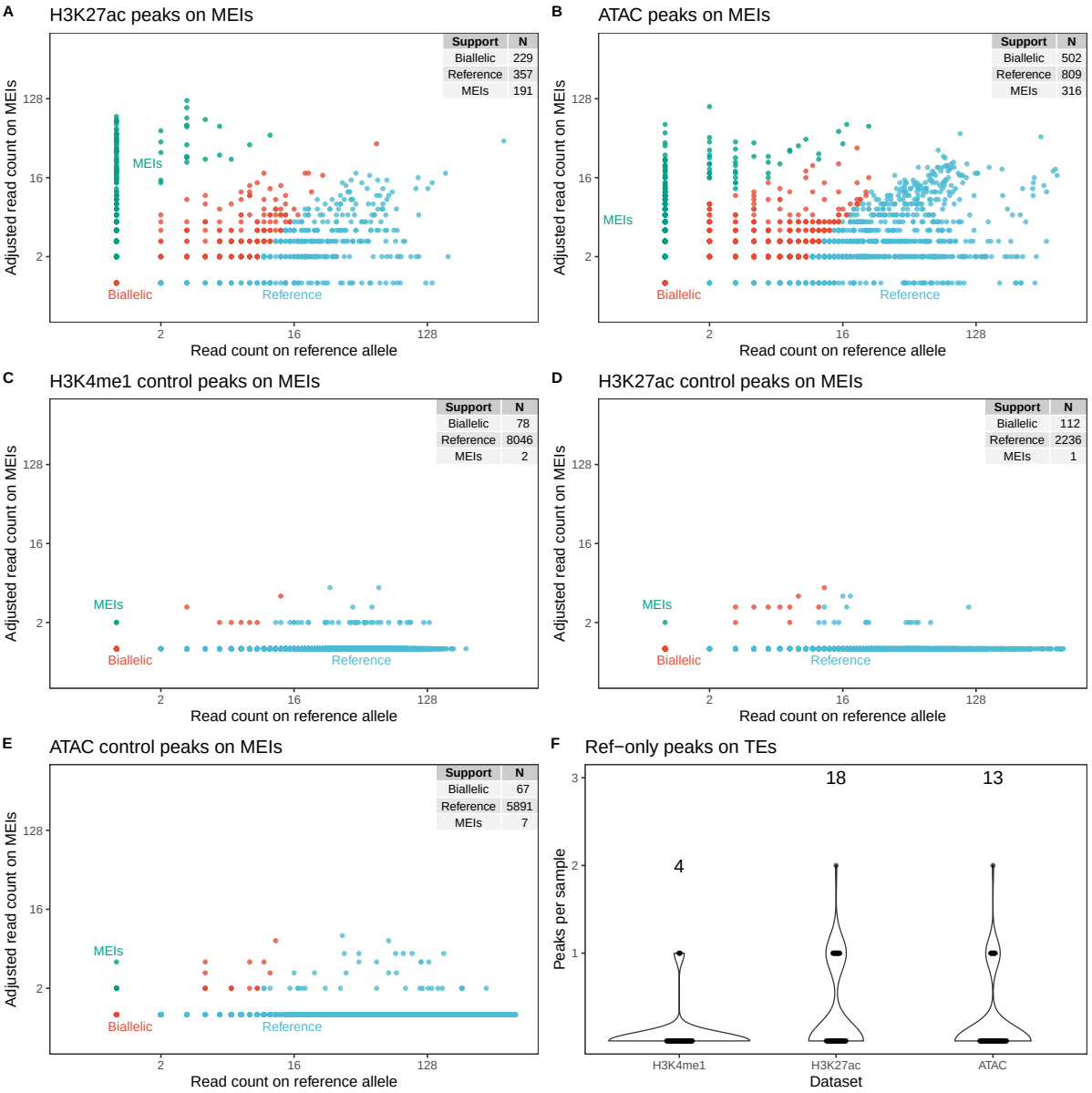

Figure S11: A) B) Partitioning of reads between the reference allele and the alternative allele in peaks that overlap heterozygous or homozygous MEIs in H3K27ac and ATAC-seq. The number of peaks with MEI, biallelic and reference read support is summarized in the table. C) D) E) The same for loci where the samples are known to be homozygous for the reference allele. F) The number of ref-only peaks that overlap MEIs.

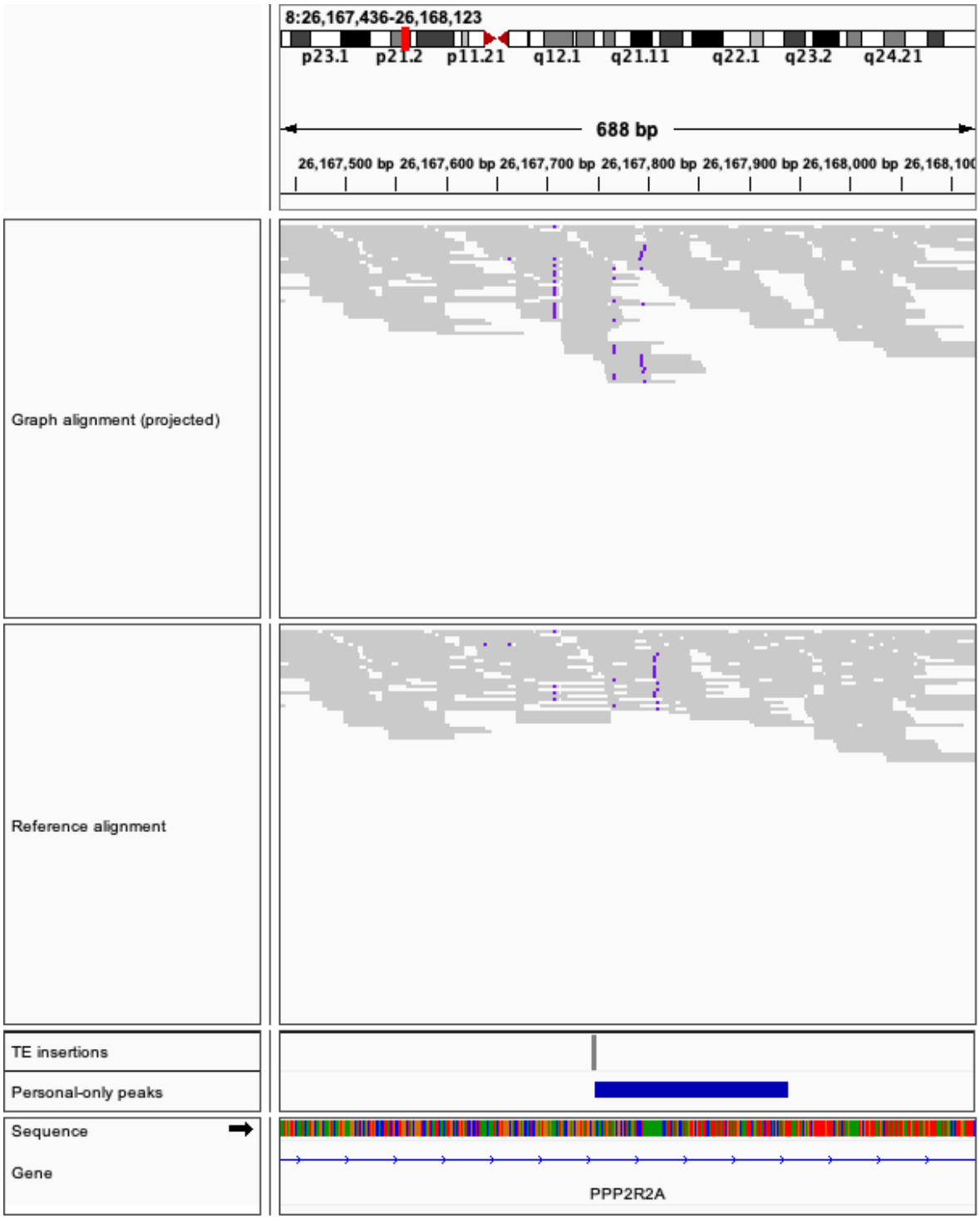

Figure S12: Linear surjection of the example locus (top) versus the reference alignment (bottom) in a Alu polymorphic locus. Note that the actual read pileup is deformed by the surjection to the linear reference. Reads with a MAPQ below 10 have been filtered out. This peak is personal-only in five non-infected samples (AF04, AF08, AF16, EU13, EU47).

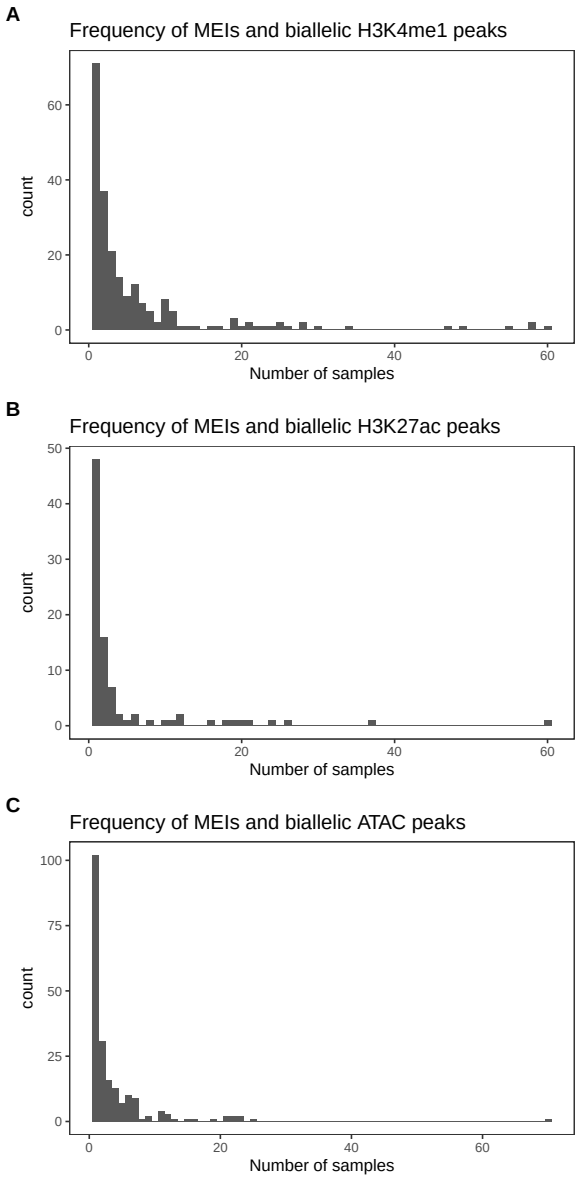

Figure S13: The number times we observe a given MEI and biallelic peak in the cohort at the same locus in A) H3K4me1, B) H3K27ac and C) ATAC-seq.

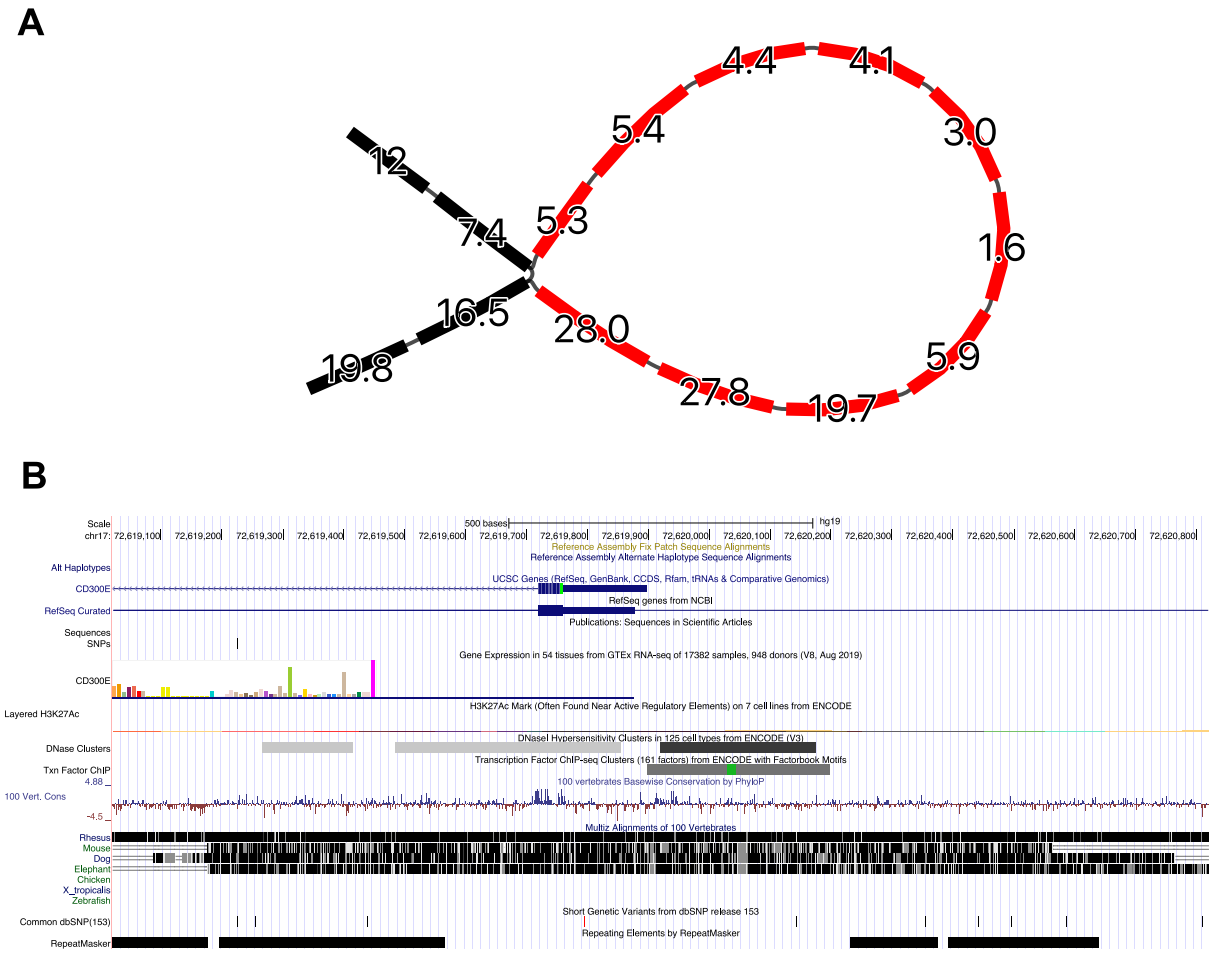

Figure S14: A) Alu mobile element insertion that supports a H3K27ac peak. B) This insertion is immediately upstream of the CD300E gene, and is within a DNase cluster and a ChIP-seq transcription factor cluster.