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

Cristian Groza[a], Xun Chen[b], Alain Pacis[c], Marie-Michelle Simon[d], Albena Pramatarova[d], Katherine A. Aracena[e], Tomi Pastinen[f], Luis B. Barreiro[g,h,i], Guillaume Bourque[b,c,d,j,*]

[a] *Quantitative Life Sciences, McGill University, Montréal, QC, Canada*
[b] *Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan*
[c] *Canadian Centre for Computational Genomics, McGill University, Montréal, QC, Canada*
[d] *McGill Genome Centre, Montréal, QC, Canada*
[e] *Human Genetics, University of Chicago, Chicago, IL, USA*
[f] *Genomic Medicine Center, Children's Mercy Hospital and Research Institute, KC, MO, USA*
[g] *Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, IL, USA*
[h] *Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, IL, USA*
[i] *Committee on Immunology, University of Chicago, Chicago, IL, USA*
[j] *Human Genetics, McGill University, Montréal, QC, Canada*

Abstract

Genetic variants, including mobile element insertions (MEIs), are known to impact the epigenome. We hypothesized that the use of a genome graph, which encapsulates genetic diversity, could reveal missing epigenomic signal. Given the contributions of mobile elements to the evolution of primate innate immunity, we tested this in monocyte-derived macrophages obtained from 35 individuals before and after *Influenza* virus infection. After characterizing genetic variants in this cohort using linked-reads, including 5140 Alu, 316 L1, 94 SVAs and 48 ERVs, we incorporated them into a genome graph. Mapping epigenetic data to this graph revealed 2.5%, 3.0% and 2.3% novel peaks for H3K4me1 and H3K27ac ChIP-seq and ATAC-seq respectively. Notably, using a genome graph also modified quantitative trait loci estimates and we observed 375 polymorphic MEIs in active epigenomic state. For example, we found an AluYh3 polymorphism whose chromatin state changed after infection and that was associated with the expression of *TRIM25*, a gene that restricts influenza RNA synthesis. Our results demonstrate that graph genomes can reveal regulatory regions that would have been overlooked by other approaches.

Keywords: epigenomics, mobile elements, pangenomics, influenza

Introduction

Structural variants (SVs) contribute the largest number of variable nucleotides in an individual [1], have larger effect sizes on gene expression [2] and are associated with functionally relevant epigenetic differences between humans and chimpanzees [3]. A particular class of SVs, mobile element insertions (MEIs), likely influences the epigenome since fixed mobile elements are known to harbor transcription factor binding sites [4, 5] and have contributed primate specific regulatory regions [6]. The epigenetic features that occur on SVs are not immediately accessible when mapping to a linear and incomplete reference genome [7, 8] but could potentially be accessed using a graph genome. Indeed,

*Corresponding author: guil.bourque@mcgill.ca

using the personal genome of a single individual we have shown previously that genome graphs can recover epigenomic signal in genetically variable regions of the genome [9]. Genome graph approaches have also been used to find differential CpG methylation within SVs in twelve medaka fish genomes [10].

Obtaining accurate maps of SVs with short read libraries can be challenging for several reasons [11]. First, repeats are abundant in eukaryotic genomes and resolving variation in these regions can be more difficult since read mapping is often ambiguous. Second, mapping short reads reveals only the break points of insertions and does not provide their actual sequence without assembling the reads into larger contigs. Third, short read assembly algorithms cannot distinguish between highly similar sequences and tend to collapse copy number variation. To mitigate these shortcomings, paired-end and linked read libraries [12] have been developed. Linked read libraries go further than paired-end libraries by labeling each read with a barcode that represent the DNA fragments from which it originates. This provides long range positional information in regions of the human genome that cannot be reached by short reads alone. Linked reads have been used to genotype [13, 14, 15], identify SVs [16, 17, 18], detect MEI polymorphisms [19] and assemble genomes [20, 21].

Given that mobile elements have been found to be co-opted in innate immunity [22], we decided to study the impact on the epigenome of genetic variants and MEIs in the response to *Influenza* virus (IAV) infection. Specifically, we used data obtained from monocyte-derived macrophages from 35 individuals of African- or European-descent before and after *in-vitro* IAV infection [23]. This included whole-genome sequence (WGS) data together with H3K4me1, H3K27ac ChIP-seq, ATAC-seq and RNA-seq data to characterize the transcriptome and the chromatin state. First, we developed a new approach to build a genome graph that includes MEIs by resolving the sequence of insertions using locally assembled linked reads (Fig S1). We validated the method by generating ChIP-seq and ATAC-seq data from the NA12878 benchmark genome [24] following the protocols used in Aracena et al. [23]. Next, we generated linked read data for the 35 individuals in the IAV infected cohort and applied our new approach to build a graph that includes SNPs, indels and MEIs. Using this genome graph, we showed we could identify regulatory sequences that would have been missed otherwise.

Results

Adding MEIs to the NA12878 genome graph reveals additional epigenomic signal

We chose the NA12878 genome to develop and benchmark our approach since WGS, linked reads and a haplotype resolved assembly were already available [24]. First, we ran MELT and ERVcaller on paired-end WGS data to identify and genotype 2175 Alu, 351 LINE1, 106 SVA and 6 ERV insertions (Fig 1A). Of these calls, 66% (1738) were previously observed by Ebert et al. [25]. MELT and ERVcaller could only predict the breakpoints of MEIs and to obtain the sequence of these insertions, we developed a local linked read assembly tool, which we called `BarcodeAsm` (Methods). By applying this tool, we successfully assembled the sequences of 1054 Alu, 117 LINE1, 17 SVA and 35 ERV instances (Fig 1B), sometimes recovering both copies of homozygous insertions corresponding to 963 loci (Fig 1A, right). We noted an excess of ERV annotations due to a set of misannotated short sequences that appear with low frequency in the local assembly windows. Next, to validate these assembled insertions, we aligned them against the haplotype resolved assembly of NA12878 and calculated the proportion of their length

that matched (Fig 1C). Despite expecting some errors in our local assembly and in the *de novo* assembly, we found that 925 loci (96.1%) had a match over 95% of their length. These matches were also equally distributed between the two haplotypes of the *de novo* assembly, confirming the robustness of the approach.

Previously, we introduced an axis that orders reference genomes according to how similar they are to the truth [9]. This axis ranged from the least accurate sequence (the reference genome) to the most representative (the *de novo* assembly). As intermediate stages along this axis, we now defined multiple genome graphs: a +SNVs graph containing only SNVs, a +indels graph containing SNVs and indels, and a +MEIs graph containing SNVs, indels and MEIs (Methods). We also created a *de novo* graph with all the variants called in the NA12878 haplotype-resolved assembly. Next, we simulated 600 million standard WGS reads from this *de novo* graph and aligned them to the genomes along this axis. As expected, we observed that the number of true positive alignments increased as the graphs became more complete (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 per variant as compared to 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 SVs.

Having established that our genome graph recovers more true mappings, we looked for MEIs that support active histone marks and chromatin accessibility signal. To mimic the data obtained from [23], we used cells derived from NA12878 and generated three replicates for H3K4me1, H3K27ac ChIP-seq and ATAC-seq (Methods). These data were then mapped to the MEI graph and we called peaks using Graph Peak Caller [27]. We observed 58 H3K4me1, 22 H3K27ac and 19 ATAC-seq peaks that overlapped MEIs in at least one of the replicates (Fig 1E). Similarly to other peaks, roughly half were observed in all three replicates. Most loci were covered on both the MEI and the reference allele and a smaller subset were covered on only one of the alleles (Fig 1F). While the number of events is small in a single genome, they show that using a graph we can profile the chromatin in regions that are missing from the reference.

Linked reads recover the sequences of polymorphic MEIs in a cohort

We looked to extend the method and apply it to a cohort with the 35 individuals that were exposed to IAV infection [23]. First, using MELT and ERVcaller on the WGS data, we identified and genotyped 7362 Alu, 1344 LINE1, 649 SVA and 19 ERV insertion loci (Fig 2A). Next, we generated linked read data from the same individuals and introduced a population consensus approach before attempting to identify the final MEI sequence at each locus with `BarcodeAsm` (Methods). Using this approach we were able to assemble and annotate the insertions for 5140 Alu, 316 LINE1, 94 SVA and 48 ERV (Fig 2A, right). The population consensus approach allowed us to recover a larger fraction (60% versus 37%) of MEIs 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 also the least likely to be assembled (Fig S2). As expected, the length distributions of the consensus insertions identified include the Alu peak at 300 bp and a long tail associated with longer truncated and full length MEIs such as the

LINE1 (Fig 2B-C). The resulting multiple sequence alignments showed few ambiguous nucleotides, suggesting that the consensus insertions are representative of most samples (Fig S3A). When we ordered mobile element families based on abundance (Fig 2D), we observed unsurprisingly that AluY families were the most common amongst Alus and that the L1HS sub-family was ranked highly [28, 29]. Similarly, the human specific SVA_F sub-family [30] was found to be the most abundant among SVA families in our dataset.

Next, we looked at the distance between insertions and the nearest exon, enhancer and repetitive sequences in the reference genome (Methods). We found that 87 insertions, mostly Alus, were within the bodies of exons (Fig 2E), which is notable since Alus are associated with alternative transcription events [31]. We also observed 135 insertions located within enhancers (Fig 2E) and that more than half of the insertions (2941) were nested within a repetitive sequence (data not shown). Overall, the distribution of these MEIs around exons, enhancers and other repeats is similar to those in the 1000 Genomes Project [32] (Fig S3B-D). Lastly, we checked if the insertions that we reconstructed with `BarcodeAsm` and their genotype would recapitulate the known African and European ancestry structure of the cohort. We found two clusters in the principal component analysis (Fig 2F) that were consistent with the genetic ancestry and what we see from SNV calls from WGS data (Fig S4). Overall this confirms that we were able to recover a high quality set of MEIs for our cohort together with their assembled sequences.

A genome graph increases the number of peak calls near immune genes and impacts QTL estimates

We wanted to explore the extent to which a cohort genome graph would impact read mapping and peak calling for epigenomic datasets. As expected [32], we observed more variants in samples from African ancestry (Fig 3A). Combining all these variants, we built a genome graph containing a total of 1.6×10^7 SNVs, 1.1×10^6 insertions, 1.3×10^6 deletions and 5.6×10^3 MEIs. We then mapped ChIP-seq and ATAC-seq datasets before and after IAV infection [23] and observed a decrease in unmapped reads and an increase in perfectly mapped reads that can reach up to 0.15% of reads in a data set (Fig 3B). Next, we called peaks using either the reference genome or the cohort genome graph. Peaks called with both approaches were called `common peaks`, while those called only in the reference graph are `ref-only` and those called only in the cohort graph are `graph-only`. Among H3K4me1 samples, we observed an average of 4700 (2.5%) graph-only peaks and 2200 (1.2%) ref-only peaks per sample (Fig 3C, S5A). The net increase in the number of peaks is caused by additional mapped reads that push peaks above the significance threshold. The fact that graph-only peaks are approximately twice as numerous as ref-only is consistent with what was observed previously [9]. For H3K27ac, we counted 1800 (3.0%) graph-only peaks and 1100 (1.9%) ref-only peaks (Fig S6A, S5B) on average. Among ATAC-seq datasets, graph-only events averaged 4000 (2.3%) peaks in flu-infected samples and ref-only events average 2400 (1.4%) peaks per sample (Fig S6B). In non-infected samples, the same numbers and proportions are roughly halved (S5C). 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).

To confirm whether altered peaks were associated with sequence variants, we estimated the influence of genotype on common and graph-only peaks by logistic regression on SNPs and indels within the peaks while controlling for peak width (Table S1). We found that the log-odds for a peak to be graph-only increased by 0.11 to 0.19 when SNPs

were present and by 0.45 to 0.56 when indels were present. Overall this confirms that indels have a stronger influence on peak calling as compared to SNPs. A similar analysis could not be performed with MEIs because of the small numbers. This multi-sample epigenomic dataset was also an opportunity to better understand how reliable the altered peaks were compared to common peaks. We contrasted the population frequency of graph-only, ref-only and common peaks with a peak replication cumulative distribution curve (Methods). H3K4me1, H3K27ac and ATAC peak sets generated very similar curves, with common peaks having the longest tails followed by graph-only and ref-only peaks, with the curves expected by chance decaying the fastest (Fig 3D, S6C-D). Under the random simulations, none of the peaks were observed in more than 20 datasets, but a proportion of graph-only peaks were replicated in more than 40 datasets (each individual has a non-infected and an infected dataset).

To understand the relevance of the newly identified peaks in the response to IAV, we retrieved the ontological descriptions of genes within 10 Kbp of a peak (Methods). As expected, we found that the genes near common peaks were functionally enriched for immune biological processes (Fig S7). Notably, genes near graph-only peaks were enriched for similar ontological terms (Fig S7 and Fig 3E). For example, 281, 203 and 249 genes related to "positive regulation of NF-kappaB" had graph-only peaks for H3K4me1, H3K27ac and ATAC, respectively. Finally, the use of a genome graph could affect methods to identify quantitative trait loci (QTLs), which aim to characterize the impact of genetic variants [33]. To measure this, we mapped caQTLs (chromatin accessibility) and hQTLs (H3K4me1 and H3K27ac histone modifications) using reference-based and graph-based read count estimates (Methods). We measured changes in effect size and p-value and found that while most QTLs remained the same, some do change. For example, when mapping caQTLs, the estimated effect size changes by 1.4 or more for one QTL in 1000 and the observed p-value (as $-log(p)$) changes by 7.9 or more (Fig 3F). The mapping of H3K4me1-QTLs and H3K27ac-QTLs is similarly affected (Fig S8). We checked the genes that are nearby these top changing QTLs and found enrichment for mostly immune pathways, which is consistent with the cell type (Fig S9). For example, the gene *HLA-DQA1* is near such changing QTLs for H3K4me1, H3K27ac and ATAC-seq. Therefore, removing reference bias from the analysis of epigenomic data using genome graphs can reveal novel peaks and improve QTL discovery.

Genome graphs measure epigenomic signal on MEIs

We wanted to focus next on the MEIs that were assembled and introduced in the cohort genome graph. First, we took advantage of the genotypes of the individuals in the cohort 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). Reassuringly, we find that the graph genotypes are highly consistent with the calls made by MELT and ERVcaller (Fig 4A). When genotyping, the graph recapitulates between 1274 to 1563 genotypes per sample and only misses 60 to 95 insertions. The graph also gains between 110 and 196 insertions per sample. These gained genotypes are not necessarily false positives and could be explained by an increase in sensitivity since the exact location and sequence of these polymorphisms were known *a priori* with the graph genotyping algorithm but needed to be found *de novo* by MELT and ERVcaller. Having confirmed that read mapping was reliable in MEIs, we extracted H3K4me1, H3K27ac ChIP-seq and ATAC-seq peaks overlapping MEIs (Fig 4B, S10A-B). The peaks were labeled either as reference peaks, biallelic peaks or MEI peaks (binomial

test, Methods). Notably, we found 714 MEI peaks for H3K4me1, 191 for H3K27ac and 316 for ATAC-seq. As an additional negative control, we repeated the same with peaks in MEI loci for which the samples were homozygous reference. We found that reads in these peaks were overwhelmingly assigned to the reference allele (Fig S10C-E).

MEIs peaks for H3K4me1 and H3K27ac were mostly from AluY families or other Alu elements (Fig 4C). L1 (L1HS, L1PA2) and ERV (LTR5_Hs) insertions also support a small number of peaks. Furthermore, we detected 30 SVA insertions in open chromatin states (Fig 4C). Notably, Chen et al. [34] found that SVA families become enriched in open chromatin after influenza infection in macrophages and are variable between individuals (Fig S11). Consistent with this, the overwhelming majority of ATAC-seq peaks that lie on SVA insertions were detected in the infected condition (92 of 108 peaks, 85.2%). Finally, we counted how many of these peaks were 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 graph-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 were graph-only. On the other hand, MEIs rarely disrupt peaks in the reference graph (Fig S10F). We show one Alu insertion in the graph that supports a H3K4me1 graph-only peak (Fig 4E), and its linear surjection (Fig S12).

Cohort data reveals MEIs that act as potential enhancers

So far, we have focused on detecting single MEI alleles that support chromatin marks in individual genomes. Next, we wanted to obtain an overview of MEIs at the level of the entire cohort. In total, we identified 375 MEIs that carry at least one epigenomic peak. Specifically, 218 MEIs support H3K4me1 peaks, 90 support H3K27ac peaks and 210 support ATAC peaks (Fig 5A), many of which were found in more than one sample (Fig S13). Among the H3K4me1 MEI peaks, 58 were unique to flu-infected samples, 41 were unique to non-infected samples and 119 were found in both conditions (Fig 5B). Peaks in H3K27ac and ATAC-seq were also well balanced between the two conditions. Ordering MEIs by allele frequency, shows variable levels of occupancy by peaks (Fig 5A, right side).

Next, we looked for MEIs that carry combinations of epigenomic marks that are characteristic of enhancer sequences (Fig 5C). For example, 56 MEIs supported both H3K4me1 and ATAC, a combination which suggests poised enhancers in open chromatin. Similarly, 16 instances showed H3K27ac and ATAC and 26 insertions were marked by H3K4me1, H3K27ac and ATAC, which is strong evidence for active enhancers in open chromatin. There is a noticeable number of SVA elements showing ATAC peaks, second only to Alu elements. We find that peaks on MEIs have a similar distribution relative to exons, introns, genes and enhancers as compared to other peaks and quite distinct from a random distribution (Fig 5D). Additionally, 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 S14). For example, an Alu insertion that supports H3K27ac peaks (Fig S15A) in 21 samples is immediately upstream of *CD300E*, an immune-activating receptor gene [35]. This frequent MEI peak is also located within DNase and transcription factor clusters (Fig S15B), which is further evidence for active chromatin.

Finally, since Alu polymorphisms could alter gene transcript levels [36], we asked if any MEIs were eQTLs for expressed genes. We mapped MEI-eQTLs and found 18 MEIs in the flu-infected condition and 34 MEIs in the non-infected condition that were associated with gene expression (FDR lower than 5×10^{-2}, Methods). In total, there are also 354 MEIs

that are ca/hQTLs for at least one mark or condition (Fig S16). Of the MEI-eQTLs, 3 have a supporting caQTL or hQTL in the flu-infected condition and 9 in the non-infected condition (Table S2). In particular, in the flu-infected condition, we detected an AluYh3 MEI that is an eQTL for *DGKE* and *TRIM25*, a gene that restricts influenza RNA synthesis [37], and is also a QTL for 2 H3K4me1 peaks, 2 H3K27ac peaks and 3 ATAC peaks (Table S3). We show the average read depth at this locus in flu-infected samples that carry the MEI (Fig 6A), the flu-infected samples that do not carry the MEI (Fig 6B), in non-infected samples that carry the MEI (6C) and in non-infected samples that do not carry the MEI (Fig 6D). The more traditional linear projection of the average read depth at this locus confirms that the signal is higher in flu-infected samples that carry the MEI when using an accurate genome graph (Fig 6E-F). Overall, this MEI exists in a flu-specific active chromatin state, is associated with *TRIM25* and *DGKE* gene expression and would have been missed with the linear reference genome.

Discussion

We have constructed a genome graph to encapsulate the genetic diversity of a cohort of 35 individuals by re-sequencing their genome with linked-reads. These data allowed us to reliably discover and phase SNVs, small indels and a large number of MEIs; the latter through the development of a local assembly method. We note that, because of the limitations of the linked-read technology, the MEIs included were skewed towards the shorter insertions and missed a significant portion of larger events. Indeed, out of the 1344 L1 polymorphisms detected in the cohort, 1048 (78%) could not be assembled and integrated into the genome graph. We anticipate that long read technologies [25] will be needed to construct more complete genome graphs that represent even more complex SVs.

We showed that our cohort genome graph improves read mapping for epigenomic data. This lead to a net increase in the number of peak calls, which are replicated across samples and occur in functionally interesting loci such as in the vicinity of immune genes. In addition, we showed that using a genome graph had an impact on the discovery of histone and chromatin accessibility QTLs in polymorphic regions. However, these new results were restricted to regions where sequence variants were added the graph, which based on our current approach remain relatively small (SNPs, indels and short MEIs). Therefore, the gains were mostly associated smaller epigenomic features such as narrow histone marks or chromatin accessibility. We anticipate that adding larger and more complex structural variations to the graph, such as segmental duplications, will reveal even more epigenomic features that are present but difficult to measure using traditional approaches.

Furthermore, we used the cohort genome graph to re-genotype 5598 MEIs using short read WGS data and reliably assign epigenomic signal to MEI alleles. While rare, we sometimes observed heterozygous MEIs where the majority of the epigenomic signal was either on the insertion or the reference allele, revealing homologous loci that are in different chromatin states. This shows that genome graphs can be exploited to search for epigenomic signal that is specific to alleles in a cohort of individuals. Meanwhile, current methods that measure allelic specific signal are limited in their ability to represent complex structural variants or relate genomes within a population to one another. Therefore, genome graphs could prove a powerful framework to study the chromatin state of polymorphic structural variants in a large number of genomes.

Finally, with the recent completion of the first full human genome [38], there is a general appreciation of the need to use a new baseline for genomic and epigenomic analysis,

including graph genomes that can capture genetic diversity. The Human Pangenome Reference Consortium is making great strides towards this goal [39] and new approaches are needed to demonstrate the benefits of these richer references. Once a human pangenome graph is developed from a set of high quality and diverse genome assemblies, a more comprehensive set of structural variants will become easily accessible. The hope is that similar analyses as was done here could be employed on a multitude of cell types and phenotypes to identify functionally relevant SVs without the challenges associated with the construction of cohort-specific genome graphs.

Methods

ATAC-seq and ChIPmentation library preparation

ATAC-Seq library preparation was performed according to the Omni-ATAC protocol [40]. 50,000 GM12878 cultured cells were resuspended in 1 ml of cold ATAC-seq resuspension buffer (RSB; 10 mM Tris-HCl pH 7.4, 10 mM NaCl, and 3 mM MgCl2 in water). Cells were centrifuged at 500 g for 5 min in a pre-chilled (4 °C) fixed-angle centrifuge. After centrifugation, supernatant was aspirated and cell pellets were then resuspended in 50 μl of ATAC-seq RSB containing 0.1% IGEPAL, 0.1% Tween-20, and 0.01% digitonin by pipetting up and down three times. This cell lysis reaction was incubated on ice for 3 min. After lysis, 1 ml of ATAC-seq RSB containing 0.1% Tween-20 (without IGEPAL and digitonin) was added, and the tubes were inverted to mix. Nuclei were then centrifuged for 10 min at 500 rcf in a pre-chilled (4 °C) fixed-angle centrifuge. Supernatant was removed and nuclei were resuspended in 50 μL transposition mix (2x TD Buffer, 100 nM final transposase, 16.5 μL PBS, 0.5 μL 1% digitonin, 0.5 μL 10% Tween-20, 5 μL H2O). Transposition reactions were incubated at 37 °C for 30 min in a thermomixer with shaking at 1000 rpm. Reactions were cleaned up with Zymo DNA Clean and Concentrator 5 columns. Primers (i5 and i7) were added by amplification (12 cycles) using NEBNext 2x MasterMix. Sequencing of the ATAC-Seq libraries was performed on the Illumina NovaSeq 6000 system using 100-bp paired-end sequencing.

ChIPmentation library preparation was performed on 5 million cross-linked cells (1% formaldehyde). After cell lysis, sonication of nuclei was performed on a BioRuptor UCD-300 targeting 150-500 bp size. Immunoprecipitation and library preparation for the histone marks H3K27ac and H3K4me1 was performed following the Auto-ChIPmentation protocol for Histones (Diagenode inc, Denville, USA) according to the manufacturer's indications. Sequencing of the ChIPmentation libraries was performed on the Illumina NovaSeq 6000 system using 100-bp paired-end sequencing.

Whole genome sequencing with linked reads and genotyping

High molecular weight DNA was extracted using the MagAttract HMW kit from Qiagen and quantified using the Qubit dsDNA HS assay kit (ThermoFisher) for 35 samples. In order to generate the linked read libraries 1ng of HMW DNA was loaded on a 10x Chromium device using 10x Genome Sequencing Solution v2 reagents (10X Genomics). Illumina compatible libraries were prepared as per the 10x Genomics protocol and loaded as one library per lane on the Illumina HiSeqX instrument for 150bp paired-end sequencing. SNVs and indels were called using the `longranger` pipeline.

Locally assembling linked reads with BarcodeAsm

For the purpose of locally assembling linked reads, we wrote BarcodeAsm [41]. The inputs to BarcodeAsm are a BED file describing the regions to be locally assembled and a BAM file that was aligned with `lariat` [17]. This BAM file must be provided twice, once sorted by position and once sorted by barcode (with `bxtools`).

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 (800 bp for Alus, 8000 bp for other TEs) 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, have very low mapping quality, or be multi-mapped.

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 [44] and MELT [45] were used to genotype novel insertions of Alu, LINE1, SVA, and ERV transposable elements using short reads. 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 ($\approx$ 300 bp) and long insertions ($>$ 300 bp) with different parameters. For Alus, a minimum read overlap of 30 bp, a maximum mapping quality of 10 (as assigned by `lariat`), 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 (see `BarcodeAsm/scripts/`). Here, contigs that contain insertions are selected but not immediately used to recover an insertion. Instead, `scripts/msa.py` generates a multiple alignment using MUSCLE and calculates a consensus contig for a particular locus across samples. 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 [24] using `minimap2 -H` [43] and selecting the haplotype with the best mapping quality.

Annotating assembled insertions

To annotate the insertions (without any flanking sequences), we used RepeatMasker with the Dfam [46] database and the longest annotation was selected for each insertion. For each assembled TE polymorphism, the distance to the nearest enhancer in the GeneHancer annotation [47] and the distance to the nearest exon in the GENCODE annotation [48] were calculated. The same was done with the MEIs from the 1000 Genome Project [32] 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` [49].

Creating and benchmarking genome graphs

The genomes graphs were generated with `vg construct` [26] on VCF file containing SNVs, indels and MEIs. For the benchmark genome graph, the NA12878 Platinum callset [50] was merged with the MEI VCF file. For the cohort genome graph, SNPs and indels were called using the LongRanger pipeline [21] 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 a matching but separate WGS data set of the same samples (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` [51] and the graph genotypes were compared to ERV and MELT genotypes.

To evaluate the impact of the genome graphs on alignment, a WGS 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. First, this read set was aligned to the hg19 reference graph. Then, the alignment of the simulated reads was repeated on increasingly complete genome graphs, first by including only SNVs, then indels and lastly MEIs. Finally, the simulated reads were aligned to the NA12878 de novo 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 reference graph to obtain reference peaks. The reference peaks were intersected with the graph peaks, and categorized into common peaks, graph-only and ref-only peaks. Common peaks are unaltered peaks that are found with both the reference and cohort genome graph. Graph-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` (see `peak_variants.R`) [52]. The number of common and altered peaks were balanced by subsampling common peaks. We report the cross-validation mean coefficients and median AUC. To summarize the population properties of peaks, curves were generated for common, graph-only and ref-only peaks (see `Rscripts/peak_replication.R`). 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` [53] 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` [54]). In each case, we calculated age and batch corrected expression matrices. Here, batch is a categorical variable. We calculated normalization factors to scale the raw library sizes using `calcNormFactors` in `edgeR (v 3.28.1)` [55] and used the `voom` function in `limma (v 3.42.2)` [56] 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 [57]. 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 Table S4.

Mapping was performed combining individuals in order to increase power, thus, we included the first eigenvector obtained from a PCA on the SNP data as a covariate in our linear model in order to account for population structure. 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` [58] 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 (see `pathways.R`). 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 graph-only H3K4me1, H3K27ac and ATAC peaks. Peaks that occur only in one sample within a condition were excluded. In a separate but similar analysis, we compiled gene sets that are within 10 Kbp of a peak that is under the control of a QTL, separately for H3K4me1-QTLs, H3K27ac-QTLs and ca-QTLs, and separately in the infected and non-infected condition.

Then, gene sets were checked for functional enrichment [59, 60] in Biological Process Gene Ontology (GO:BP) terms with `gprofiler2` [61]. 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 [27] 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 (see `Rscripts/allele_support.R`) 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. We also counted the MEIs that support combinations of marks to support their functional relevance [62, 63, 64, 65].

References

[1] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature 2015;526(7571):75–81. URL: `https://doi.org/10.1038/nature15394`. doi:`10.1038/nature15394`.

[2] Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nature Genetics 2017;49(5):692–9. URL: `https://doi.org/10.1038/ng.3834`. doi:`10.1038/ng.3834`.

[3] Zhuo X, Du AY, Pehrsson EC, Li D, Wang T. Epigenomic differences in the human and chimpanzee genomes are associated with structural variation. Genome Research 2020;URL: `http://genome.cshlp.org/content/early/2021/01/15/gr.263491.120.abstract`. doi:`10.1101/gr.263491.120`.

[4] Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Research 2008;18(11):1752–62. URL: `http://genome.cshlp.org/content/18/11/1752.abstract`. doi:`10.1101/gr.080663.108`.

[5] Wang Ting , Zeng Jue , Lowe Craig B. , Sellers Robert G. , Salama Sofie R. , Yang Min , et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proceedings of the National Academy of Sciences 2007;104(47):18613–8. URL: https://doi.org/10.1073/pnas.0703637104. doi:10.1073/pnas.0703637104; publisher: Proceedings of the National Academy of Sciences.

[6] Jacques PE, Jeyakani J, Bourque G. The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. PLOS Genetics 2013;9(5):e1003504. URL: https://doi.org/10.1371/journal.pgen.1003504. doi:10.1371/journal.pgen.1003504; publisher: Public Library of Science.

[7] Daron J, Slotkin RK. EpiTEome: Simultaneous detection of transposable element insertion sites and their DNA methylation levels. Genome Biology 2017;18(1):91. URL: https://doi.org/10.1186/s13059-017-1232-0. doi:10.1186/s13059-017-1232-0.

[8] Gershman A, Sauria ME, Hook PW, Hoyt S, Razaghi R, Koren S, et al. Epigenetic Patterns in a Complete Human Genome. bioRxiv 2021;:2021.05.26.443420URL: http://biorxiv.org/content/early/2021/05/27/2021.05.26.443420.abstract. doi:10.1101/2021.05.26.443420.

[9] Groza C, Kwan T, Soranzo N, Pastinen T, Bourque G. Personalized and graph genomes reveal missing signal in epigenomic data. Genome Biology 2020;21(1):124. URL: https://doi.org/10.1186/s13059-020-02038-8. doi:10.1186/s13059-020-02038-8.

[10] Leger A, Brettell I, Monahan J, Barton C, Wolf N, Kusminski N, et al. Genomic variations and epigenomic landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel. bioRxiv 2021;:2021.05.17.444424URL: http://biorxiv.org/content/early/2021/05/17/2021.05.17.444424.abstract. doi:10.1101/2021.05.17.444424.

[11] Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 2012;13(1):36–46. URL: https://doi.org/10.1038/nrg3117. doi:10.1038/nrg3117.

[12] Kitzman JO. Haplotypes drop by drop. Nature Biotechnology 2016;34(3):296–8. URL: https://doi.org/10.1038/nbt.3500. doi:10.1038/nbt.3500.

[13] Chu C, Borges-Monroy R, Viswanadham VV, Lee S, Li H, Lee EA, et al. Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nature Communications 2021;12(1):3836. URL: https://doi.org/10.1038/s41467-021-24041-8. doi:10.1038/s41467-021-24041-8.

[14] Spies N, Weng Z, Bishara A, McDaniel J, Catoe D, Zook JM, et al. Genome-wide reconstruction of complex structural variants using read clouds. Nature methods 2017;14(9):915–20. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5578891/. doi:10.1038/nmeth.4366.

[15] Meleshko D, Marks P, Williams S, Hajirasouliha I. Detection and assembly of novel sequence insertions using Linked-Read technology. bioRxiv 2019;:551028URL:

http://biorxiv.org/content/early/2019/02/15/551028.abstract. doi:10.1101/551028.

[16] Garcia S, Williams S, Xu AW, Herschleb J, Marks P, Stafford D, et al. Linked-Read sequencing resolves complex structural variants. bioRxiv 2017;URL: http://biorxiv.org/lookup/doi/10.1101/231662. doi:10.1101/231662.

[17] Bishara A, Liu Y, Weng Z, Kashef-Haghighi D, Neuburger DE, West R, et al. Read clouds uncover variation in complex regions of the human genome. Genome Research 2015;:gr.191189.115URL: http://genome.cshlp.org/content/early/2015/08/18/gr.191189.115. doi:10.1101/gr.191189.115.

[18] Marks P, Garcia S, Martinez Barrio A, Belhocine K, Bernate J, Bharadwaj R, et al. Resolving the Full Spectrum of Human Genome Variation using Linked-Reads. bioRxiv 2018;URL: http://biorxiv.org/lookup/doi/10.1101/230946. doi:10.1101/230946.

[19] Wildschutte JH, Baron A, Diroff NM, Kidd JM. Discovery and characterization of Alu repeat sequences via precise local read assembly. Nucleic acids research 2015;43(21):10292–307. URL: https://pubmed.ncbi.nlm.nih.gov/26503250. doi:10.1093/nar/gkv1089; edition: 2015/10/25 Publisher: Oxford University Press.

[20] Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Research 2017;27(5):757–67. URL: http://genome.cshlp.org/content/27/5/757.abstract. doi:10.1101/gr.214874.116; _eprint: http://genome.cshlp.org/content/27/5/757.full.pdf+html.

[21] Ott A, Schnable JC, Yeh CT, Wu L, Liu C, Hu HC, et al. Linked read technology for assembling large complex and polyploid genomes. BMC genomics 2018;19(1):651–. URL: https://pubmed.ncbi.nlm.nih.gov/30180802. doi:10.1186/s12864-018-5040-z; publisher: BioMed Central.

[22] Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science (New York, NY) 2016;351(6277):1083–7. URL: https://pubmed.ncbi.nlm.nih.gov/26941318. doi:10.1126/science.aad5497.

[23] Aracena KA, Lin YL, Luo K, Pacis A, Gona S, Mu Z, et al. Epigenetic variation impacts ancestry-associated differences in the transcriptional response to influenza infection. submitted 2022;.

[24] Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, et al. Chromosome-scale, haplotype-resolved assembly of human genomes. Nature Biotechnology 2020;URL: https://doi.org/10.1038/s41587-020-0711-0. doi:10.1038/s41587-020-0711-0.

[25] Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 2021;:eabf7117URL: http://science.sciencemag.org/content/early/2021/02/24/science.abf7117.abstract. doi:10.1126/science.abf7117.

[26] Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature Biotechnology 2018;36:875. URL: https://doi.org/10.1038/nbt.4227.

[27] Grytten I, Rand KD, Nederbragt AJ, Storvik GO, Glad IK, Sandve GK. Graph Peak Caller: Calling ChIP-seq peaks on graph-based reference genomes. PLOS Computational Biology 2019;15(2):e1006731. URL: https://doi.org/10.1371/journal.pcbi.1006731. doi:10.1371/journal.pcbi.1006731; publisher: Public Library of Science.

[28] Deininger P. Alu elements: know the SINEs. Genome Biology 2011;12(12):236. URL: https://doi.org/10.1186/gb-2011-12-12-236. doi:10.1186/gb-2011-12-12-236.

[29] Hermant C, Torres-Padilla ME. TFs for TEs: the transcription factor repertoire of mammalian transposable elements. Genes & Development 2021;35(1-2):22–39. URL: http://genesdev.cshlp.org/content/35/1-2/22.abstract. doi:10.1101/gad.344473.120.

[30] Bantysh OB, Buzdin AA. Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. Biochemistry (Moscow) 2009;74(12):1393–9. URL: https://doi.org/10.1134/S0006297909120153. doi:10.1134/S0006297909120153.

[31] Kim S, Cho CS, Han K, Lee J. Structural Variation of Alu Element and Human Disease. Genomics & informatics 2016;14(3):70–7. URL: https://pubmed.ncbi.nlm.nih.gov/27729835. doi:10.5808/GI.2016.14.3.70; edition: 2016/09/30 Publisher: Korea Genome Organization.

[32] The 1000 Genomes Project Consortium , Auton A, Abecasis GR, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, et al. A global reference for human genetic variation. Nature 2015;526:68. URL: https://doi.org/10.1038/nature15393.

[33] Abiola O, Angel JM, Avner P, Bachmanov AA, Belknap JK, Bennett B, et al. The nature and identification of quantitative trait loci: a community's view. Nature reviews Genetics 2003;4(11):911–6. URL: https://pubmed.ncbi.nlm.nih.gov/14634638. doi:10.1038/nrg1206.

[34] Chen X, Pacis AS, Aracena K, Gona S, Tony K, Groza C, et al. Transposable elements are associated with the variable response to influenza infection. submitted 2022;.

[35] Isobe M, Izawa K, Sugiuchi M, Sakanishi T, Kaitani A, Takamori A, et al. The CD300e molecule in mice is an immune-activating receptor. Journal of Biological Chemistry 2018;293(10):3793–805. URL: https://doi.org/10.1074/jbc.RA117.000696. doi:10.1074/jbc.RA117.000696; publisher: Elsevier.

[36] Payer LM, Steranka JP, Kryatova MS, Grillo G, Lupien M, Rocha PP, et al. Alu insertion variants alter gene transcript levels. Genome Research 2021;URL: http://genome.cshlp.org/content/early/2021/11/19/gr.261305.120.abstract. doi:10.1101/gr.261305.120.

[37] Meyerson NR, Zhou L, Guo YR, Zhao C, Tao YJ, Krug RM, et al. Nuclear TRIM25 Specifically Targets Influenza Virus Ribonucleoproteins to Block the Onset of RNA Chain Elongation. Cell host & microbe 2017;22(5):627–638.e7. URL: https://pubmed.ncbi.nlm.nih.gov/29107643. doi:10.1016/j.chom.2017.10.003; edition: 2017/11/05.

[38] Nurk Sergey , Koren Sergey , Rhie Arang , Rautiainen Mikko , Bzikadze Andrey V. , Mikheenko Alla , et al. The complete sequence of a human genome. Science 2022;376(6588):44–53. URL: https://doi.org/10.1126/science.abj6987. doi:10.1126/science.abj6987; publisher: American Association for the Advancement of Science.

[39] Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature 2022;604(7906):437–46. URL: https://doi.org/10.1038/s41586-022-04601-8. doi:10.1038/s41586-022-04601-8.

[40] Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nature methods 2017;14(10):959–62. URL: https://pubmed.ncbi.nlm.nih.gov/28846090. doi:10.1038/nmeth.4396; edition: 2017/08/28.

[41] Cristian G. cgroza/BarcodeAsm: publication version. 2021. URL: https://doi.org/10.5281/zenodo.5510086. doi:10.5281/zenodo.5510086.

[42] Li H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics 2012;28(14):1838–44. URL: https://doi.org/10.1093/bioinformatics/bts280. doi:10.1093/bioinformatics/bts280.

[43] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34(18):3094–100. URL: https://doi.org/10.1093/bioinformatics/bty191. doi:10.1093/bioinformatics/bty191.

[44] Chen X, Li D. ERVcaller: identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. Bioinformatics 2019;35(20):3913–22. URL: https://doi.org/10.1093/bioinformatics/btz205. doi:10.1093/bioinformatics/btz205.

[45] Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Research 2017;27(11):1916–29. URL: http://genome.cshlp.org/content/27/11/1916.abstract.

[46] Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic acids research 2016;44(D1):D81–9. URL: https://pubmed.ncbi.nlm.nih.gov/26612867. doi:10.1093/nar/gkv1272; edition: 2015/11/26 Publisher: Oxford University Press.

[47] Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes

in GeneCards. Database : the journal of biological databases and curation 2017;2017:bax028. URL: https://pubmed.ncbi.nlm.nih.gov/28605766. doi:10.1093/database/bax028; publisher: Oxford University Press.

[48] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research 2019;47(D1):D766–73. URL: https://doi.org/10.1093/nar/gky955. doi:10.1093/nar/gky955.

[49] Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 2012;28(24):3326–8. URL: https://doi.org/10.1093/bioinformatics/bts606. doi:10.1093/bioinformatics/bts606.

[50] Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, 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. Genome Research 2017;27(1):157–64. URL: http://genome.cshlp.org/content/27/1/157.abstract.

[51] Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biology 2020;21(1):35. URL: https://doi.org/10.1186/s13059-020-1941-7. doi:10.1186/s13059-020-1941-7.

[52] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 2010;33(1):1–22. URL: http://www.jstatsoft.org/v33/i01/.

[53] Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics (Oxford, England) 2012;28(10):1353–8. URL: https://pubmed.ncbi.nlm.nih.gov/22492648. doi:10.1093/bioinformatics/bts163; edition: 2012/04/06 Publisher: Oxford University Press.

[54] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) 2013;29(1):15–21. URL: https://pubmed.ncbi.nlm.nih.gov/23104886. doi:10.1093/bioinformatics/bts635; edition: 2012/10/25 Publisher: Oxford University Press.

[55] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) 2010;26(1):139–40. URL: http://dx.doi.org/10.1093/bioinformatics/btp616. doi:10.1093/bioinformatics/btp616.

[56] Smyth GK, R. Gentleman , Carey V, Dudoit S, Irizarry R, Huber W. Limma: linear models for microarray data. In: Bioinformatics and computational biology solutions using r and bioconductor. New York: Springer; 2005, p. 397–420.

[57] Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics (Oxford, England) 2012;28(6):882–3. URL: https://pubmed.ncbi.nlm.nih.

gov/22257669. doi:10.1093/bioinformatics/bts034; edition: 2012/01/17 Publisher: Oxford University Press.

[58] Storey JD, Bass AJ, Dabney A, Robinson D. qvalue: Q-value estimation for false discovery rate control. manual; 2021. URL: http://github.com/jdstorey/qvalue.

[59] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 2000;25(1):25–9. URL: https://pubmed.ncbi.nlm.nih.gov/10802651. doi:10.1038/75556.

[60] The Gene Ontology resource: enriching a GOld mine. Nucleic acids research 2021;49(D1):D325–34. doi:10.1093/nar/gkaa1113.

[61] Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. gprofiler2 – an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler [version 2; peer review: 2 approved]. F1000Research 2020;9(709). doi:10.12688/f1000research.24956.2.

[62] Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences 2010;107(50):21931. URL: http://www.pnas.org/content/107/50/21931.abstract. doi:10.1073/pnas.1016071107.

[63] Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome research 2011;21(8):1273–83. URL: https://pubmed.ncbi.nlm.nih.gov/21632746. doi:10.1101/gr.122382.111; edition: 2011/06/01 Publisher: Cold Spring Harbor Laboratory Press.

[64] Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 2011;470(7333):279–83. URL: https://doi.org/10.1038/nature09692. doi:10.1038/nature09692.

[65] Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. Genome Research 2017;27(12):2096–107. URL: http://genome.cshlp.org/content/27/12/2096.abstract. doi:10.1101/gr.226233.117.

[66] Groza C. Genome graphs detect human polymorphisms in active epigenomic states during influenza infection: code and processed data. 2022. URL: https://doi.org/10.5281/zenodo.6525192. doi:10.5281/zenodo.6525192.

[67] Aracena K. katiearacena/Groza_et_al_mapping: for_preprint. 2021. URL: https://doi.org/10.5281/zenodo.5519627. doi:10.5281/zenodo.5519627.

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 [41].
- Additional code and processed data to reproduce the analysis, figures and manuscript are available on the Zenodo repository [66].
- Code for QTL mapping is available on the Zenodo repository [67].

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

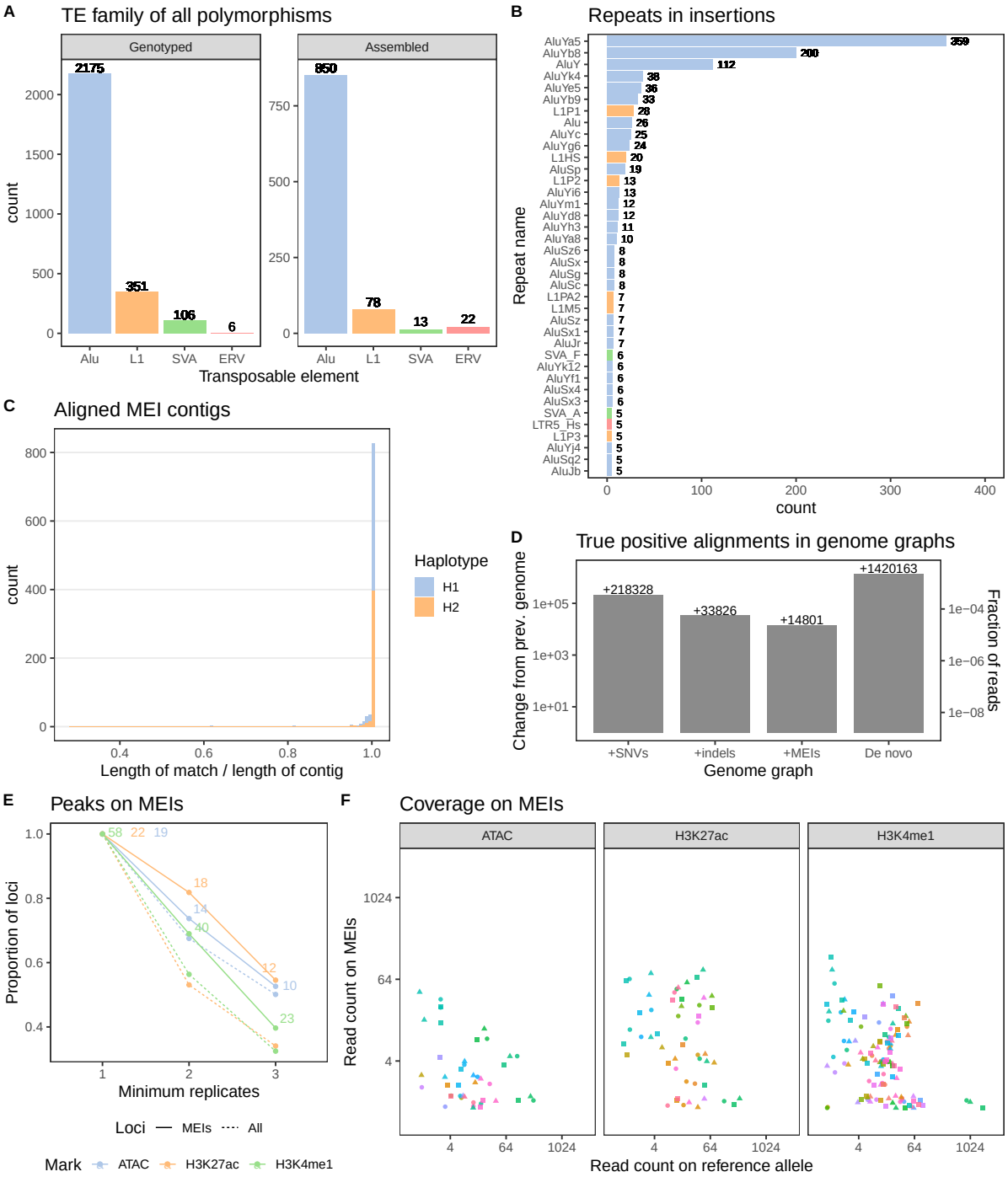

Figure 1: Adding MEIs to the NA12878 genome graph reveals additional epigenomic signal. A) The number of MEI breakpoints detected using ERVcaller and MELT and the full-length MEIs recovered by BarcodeAsm. B) The reassembled families ordered by frequency in the NA12878 genome. Homozygous insertions that are assembled twice are double counted. 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) Proportion and number of peaks on MEI (full line) that were called at least once, twice and three times in the replicates. Proportion also shown for all peaks (dashed lined). F) MEI and reference allele coverage at the peak calls that overlap MEIs, stratified by locus (color) and replicate (shape).

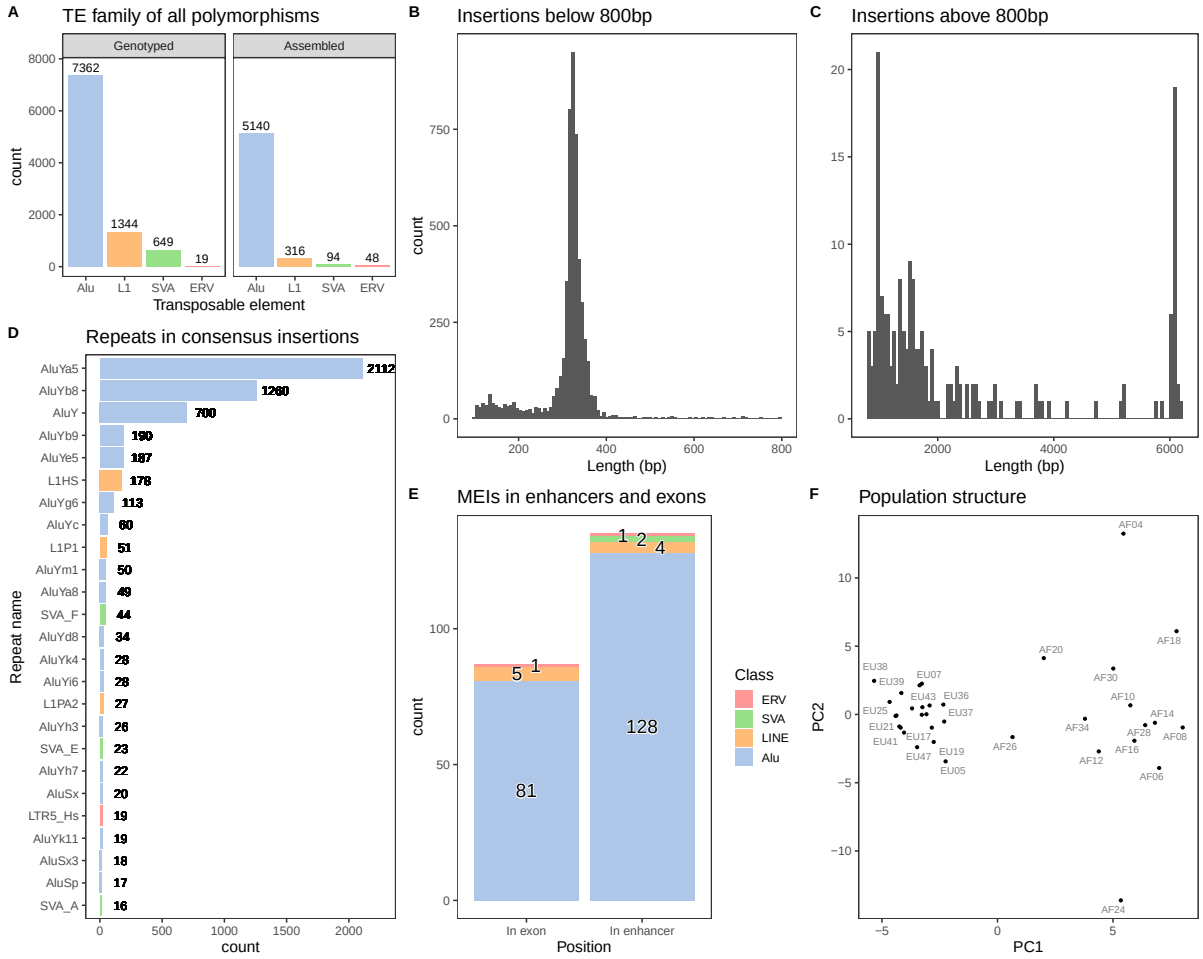

Figure 2: BarcodeAsm recovers the sequences of polymorphic MEIs in a cohort. A) The number and family of MEIs genotyped from short read sequencing data using ERVcaller and MELT in the entire cohort. The lengths of assembled MEIs B) below 800 bp and C) above 800 bp. D) The reassembled families ordered by frequency in the cohort. E) Number of MEIs inserted in enhancers or exons. F) The observed population structure in MEI genotypes as projected by principal component analysis.

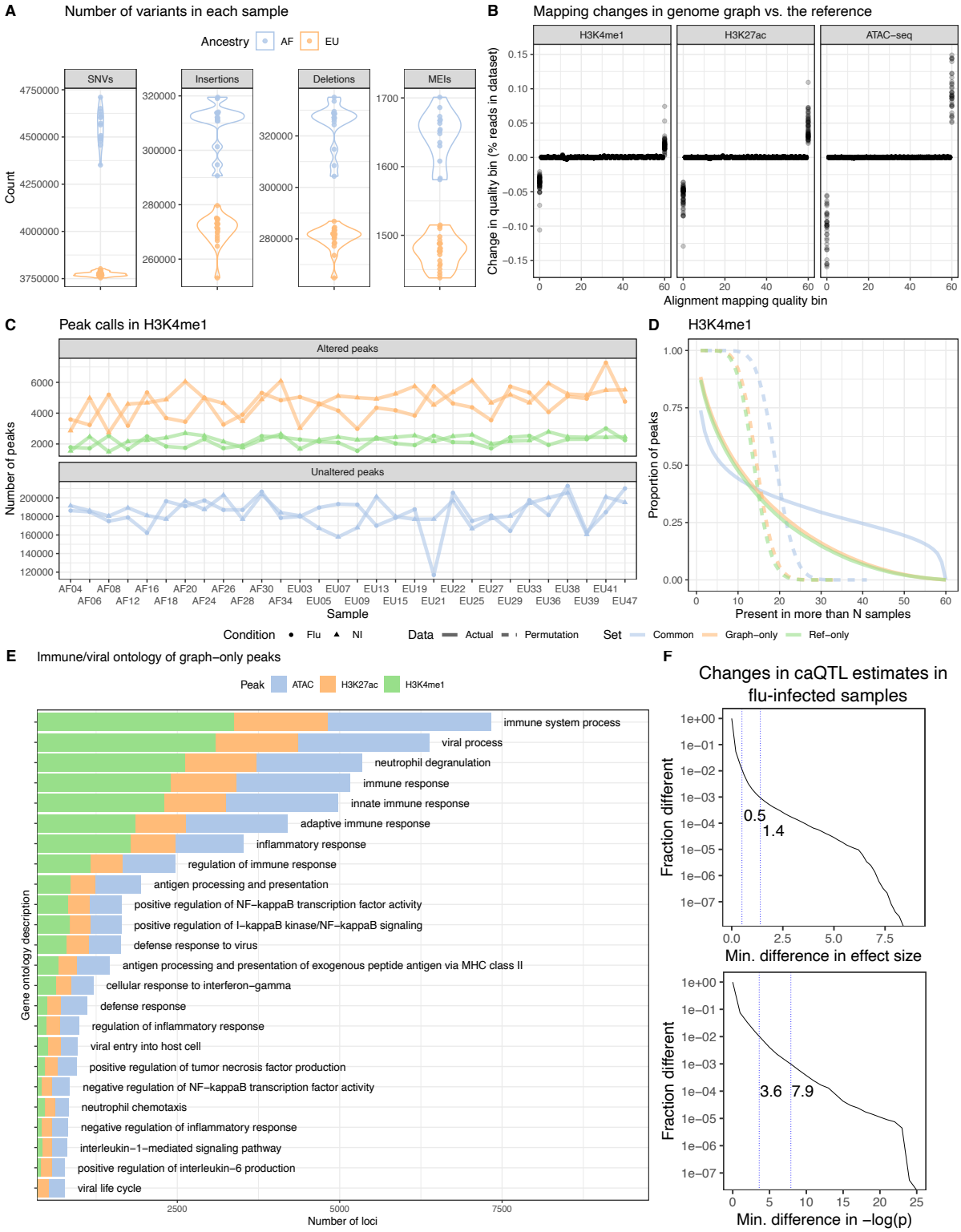

Figure 3: Genome graphs with millions of variants impact downstream results. A) Number of SNVs, insertions, deletions and MEIs in the cohort graph for each sample. B) Changes in the distribution of read mapping quality when using a cohort genome graph relative to the reference genome. Each point is an individual sample. C) The number of H3K4me1 altered (graph-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 graph-only peaks. One gene may contribute multiple descriptions. F) Distributions showing caQTL effect size and p-value estimates that changed by a minimum amount between the genome graph and the reference. First and second vertical lines mark the 99th and 99.9th percentiles.

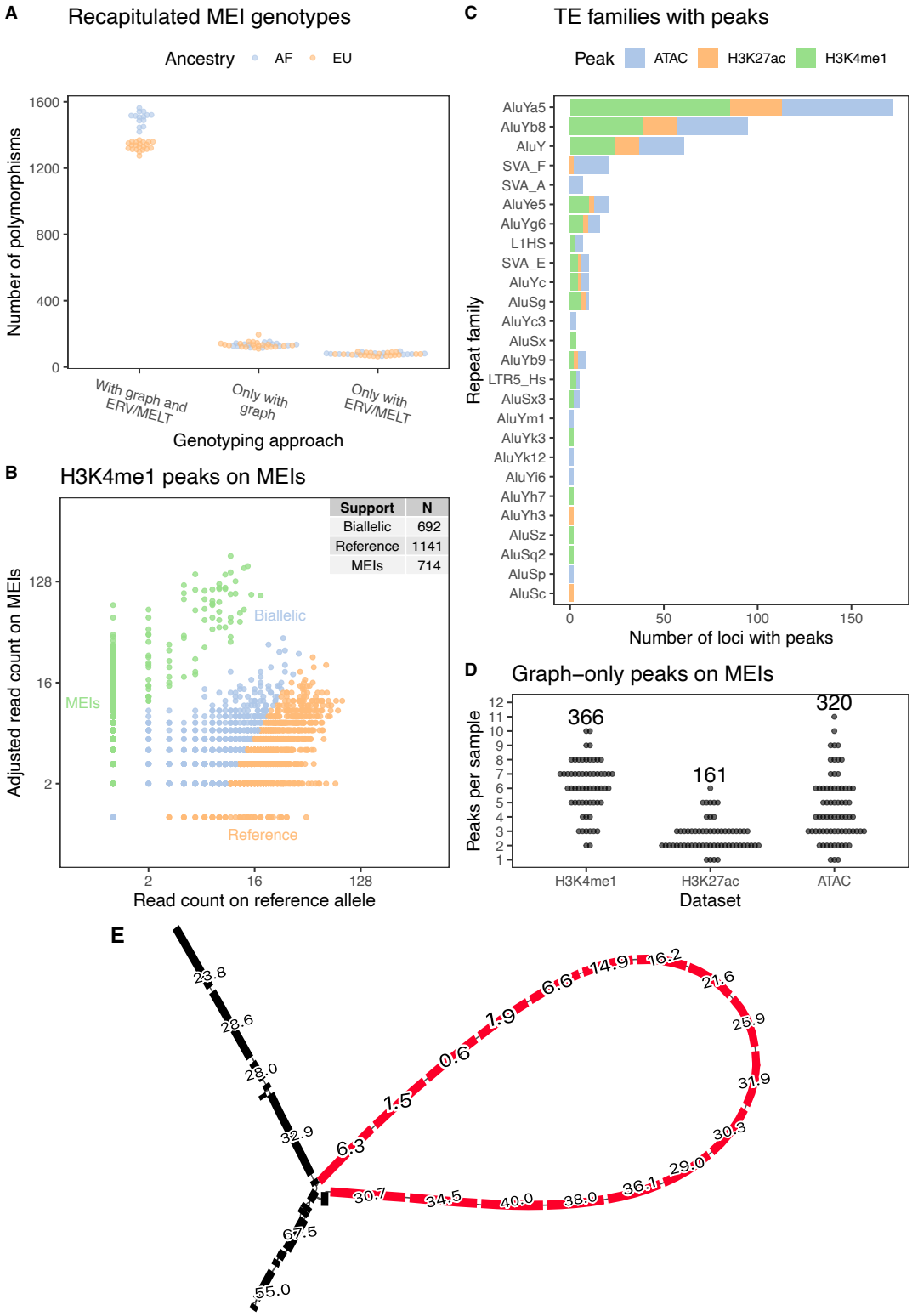

Figure 4: Genome graphs measure epigenomic signal on MEIs. A) Assembled MEIs that were regenotyped using the cohort genome graph. Graph genotypes are compared to the previous genotypes that were called with ERVcaller and MELT. B) Partitioning of reads between the reference and alternative allele in peaks that overlap heterozygous or homozygous MEIs. C) TE families that support peaks in at least one sample. Singletons not shown. D) Number of graph-only peaks that overlap MEIs in each sample, with total number across genomes annotated at the top. E) A graph region that represents a MEI locus. The black nodes represent the reference sequence and the red nodes represent the Alu insertion carried by some samples. The numbers show the average number of H3K4me1 reads that were mapped to each nucleotide over the span of the node. Nodes are at most 32 bp long.

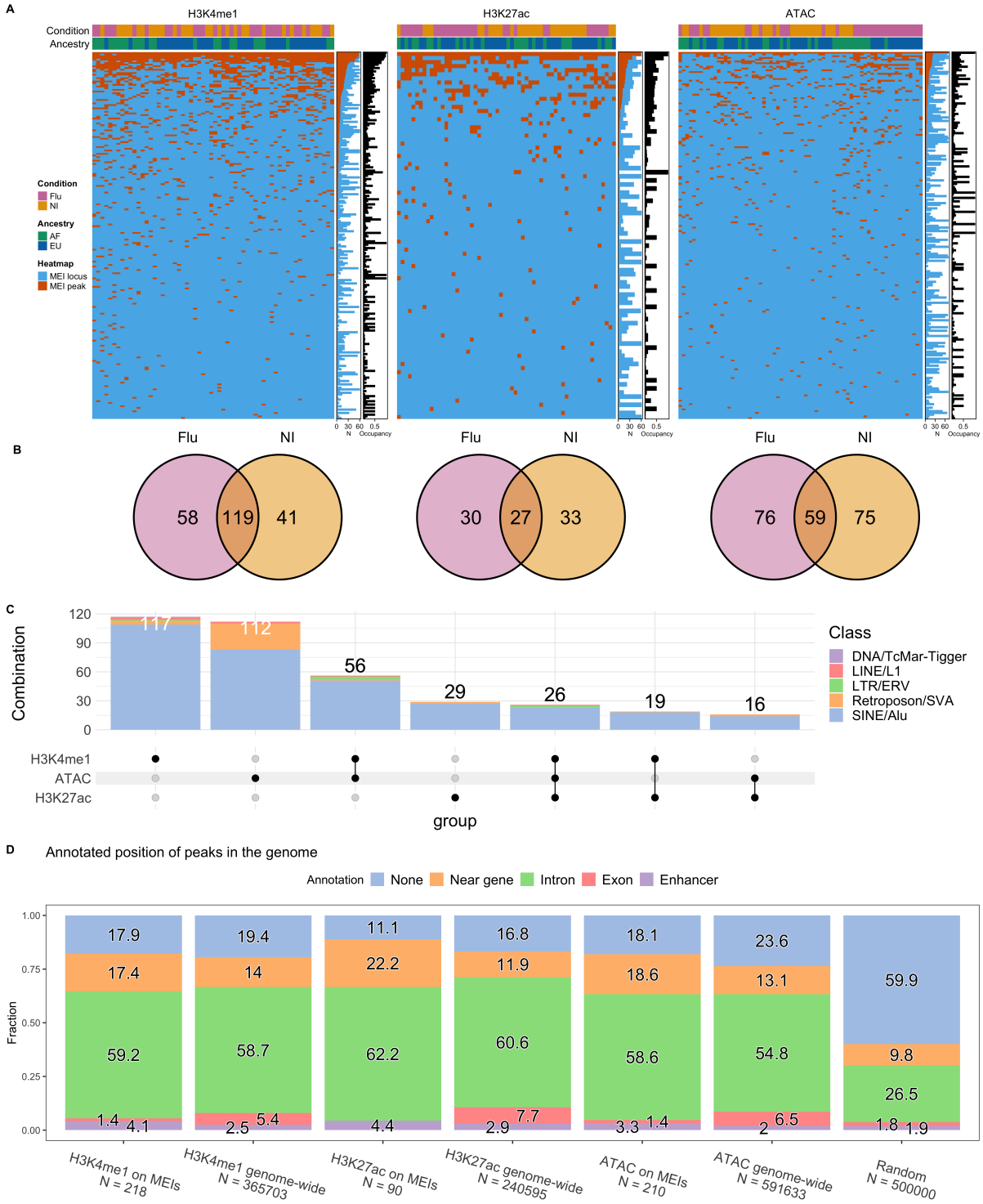

Figure 5: Cohort data reveals MEIs that act as potential enhancers. A) Summary of MEIs (rows) that support H3K4me1, H3K27ac and ATAC peaks in the cohort samples (columns). Occupancy is the ratio between samples that support peaks on the MEI (N - red) and those that carry the MEI (N - blue). B) Venn diagrams 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.

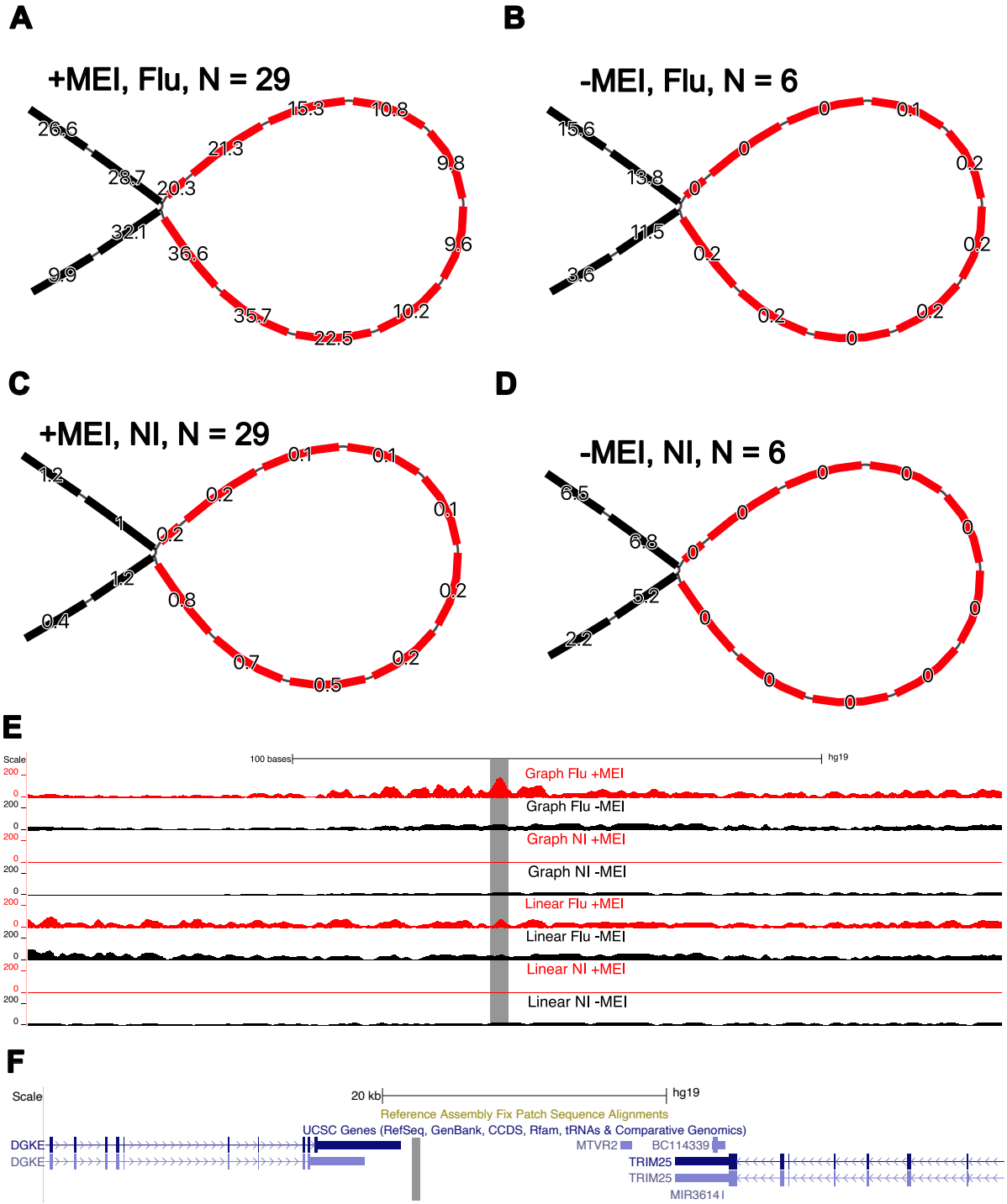

Figure 6: Example of flu-specific and genotype-specific peak on an MEI. A) Average H3K27ac read depth in the locus of an AluYh3 MEI-eQTL in flu-infected samples that carry the insertion, B) in flu-infected samples that do not carry the MEI, C) in non-infected samples that carry the MEI and D) in non-infected 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. F) Nearby genes that are associated with this MEI-eQTL (DGKE, TRIM25). The grey strips denote the position of the MEI.

Supplementary Tables

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

Table S1: Relative influence of peak width (100bp), SNPs and indels on the log-odds of a peak being found only with the cohort graph genome.

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, ATAC	ENSG00000153933	0.91	9.56×10^{-16}
chr17:54947569	Flu	H3K4me1, H3K27ac, ATAC	ENSG00000214226	0.44	9.69×10^{-9}
chr17:54947569	Flu	H3K4me1, H3K27ac, ATAC	ENSG00000121060	0.24	2.13×10^{-8}
chr20:13704145	NI	H3K4me1, ATAC	ENSG00000101247	0.14	3.69×10^{-4}

Table S2: 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.

QTL	Effect size	FDR
chr17:54947569 — H3K4me1, H3K27ac, ATAC — AluYh3		
TRIM25 - Flu	0.239	2.13×10^{-8}
DGKE - Flu	0.912	9.56×10^{-16}
H3K4me1:chr17:54944305-54949599 - Flu	0.435	1.68×10^{-5}
H3K4me1:chr17:54952157-54958631 - Flu	0.114	2.79×10^{-2}
H3K27ac:chr17:54945079-54949298 - Flu	0.507	8.85×10^{-7}
H3K27ac:chr17:54956812-54958346 - Flu	0.646	6.95×10^{-5}
ATAC:chr17:54946838-54948880 - Flu	0.923	3.85×10^{-12}
ATAC:chr17:54945058-54946521 - Flu	0.466	1.67×10^{-8}
ATAC:chr17:54952726-54958278 - Flu	0.252	1.15×10^{-5}
chr11:47806640 — H3K27ac — AluYh7		
NUP160 - NI	0.278	3.99×10^{-7}

Table S3: Summary of MEI-QTLs listing the regressed features (peaks within 10 Kbp or 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.

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

Table S4: Number of principal components performed on age and gene expression included in each QTL discovery analysis in order to correct for age and batch effects.

Supplementary Figures

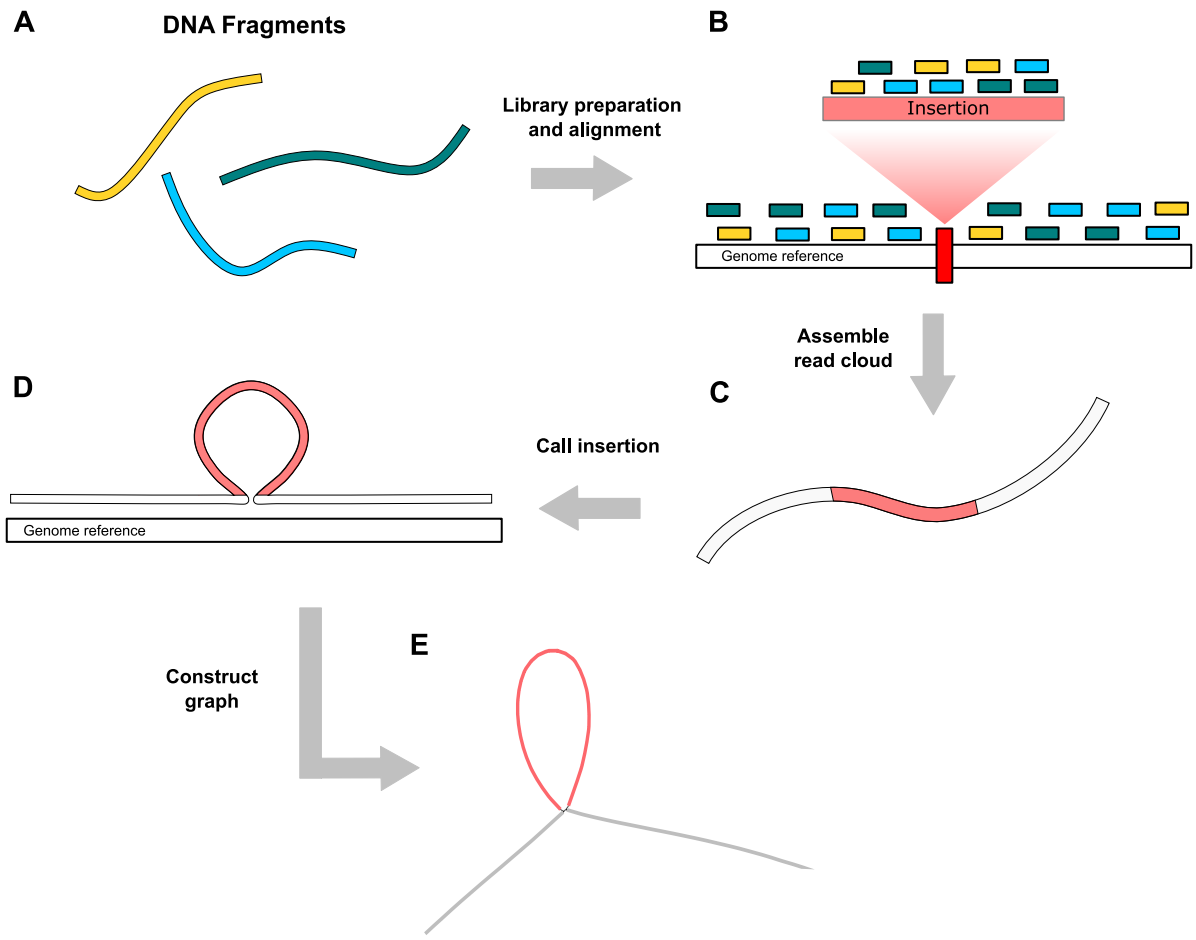

Figure S1: A) DNA fragments that are associated with barcodes (yellow, blue, green). B) Throughout the sequencing library protocol and alignment to the reference genome with lariat, the short reads remain associated with the barcode of the source fragment. We enumerate the barcodes observed around an 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.

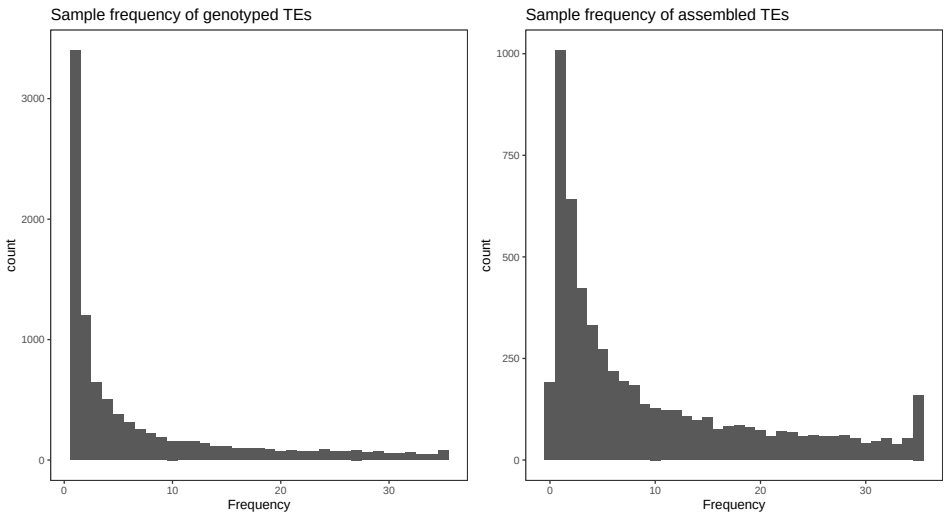

Figure S2: Population frequency of the A) genotyped and B) assembled MEIs in the cohort.

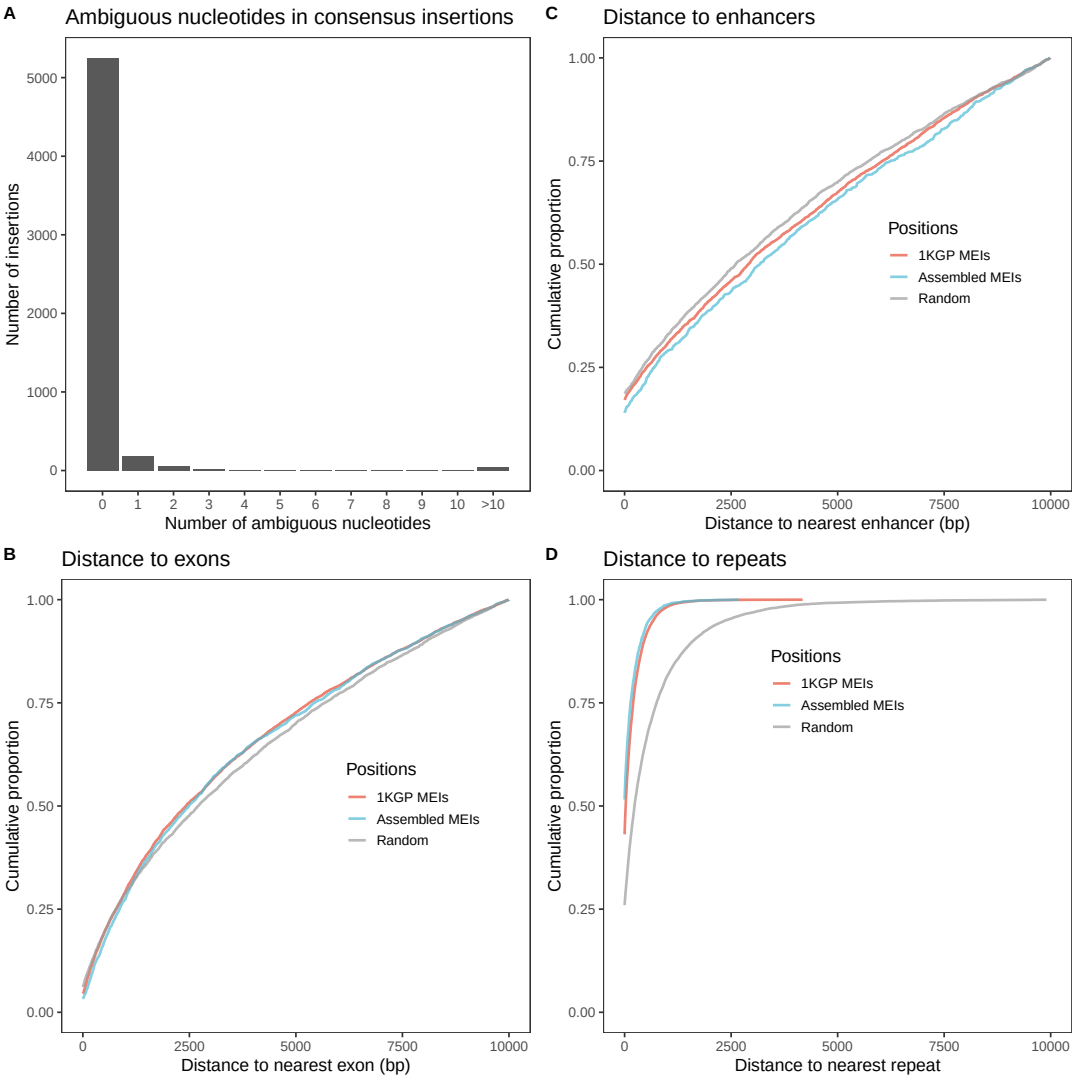

Figure S3: A) Number of ambiguous nucleotides per sequence in consensus insertions. The distributions of insertions in the genome around B) exons, C) enhancers and D) other repeats.

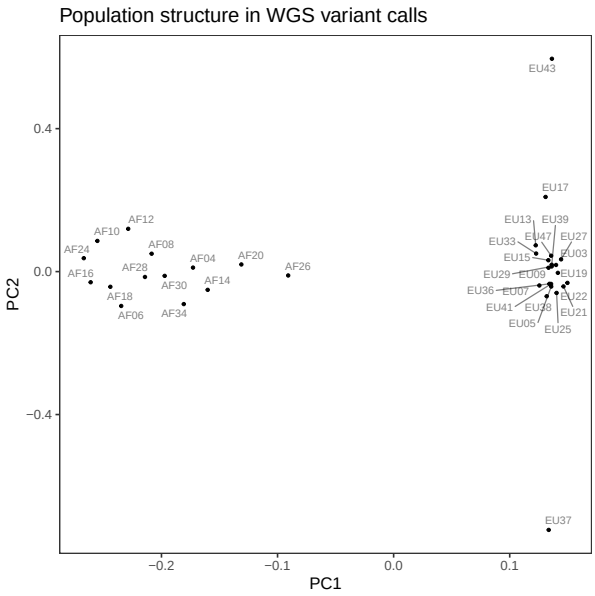

Figure S4: PCA projection showing the population structure observed in whole genome sequencing genotypes.

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

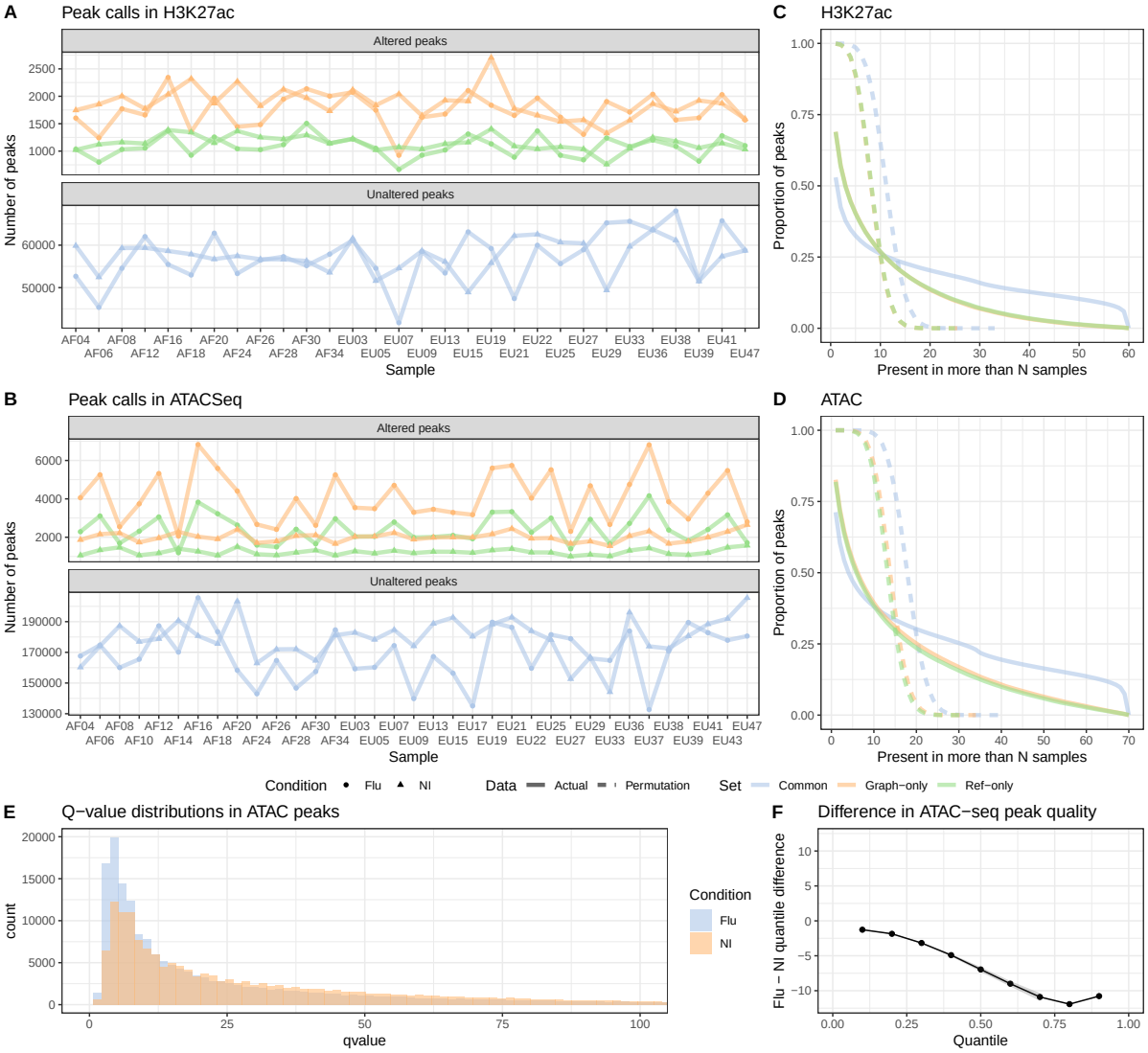

Figure S6: A) B) The number of altered (graph-only, ref-only) and unaltered (common) peaks between the cohort and the reference genome graphs for H3K27ac and ATAC-seq. C) D) Inverse cumulative distributions describing how many H3K27ac and ATAC-seq peaks are observed in more than a number of samples E) 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.

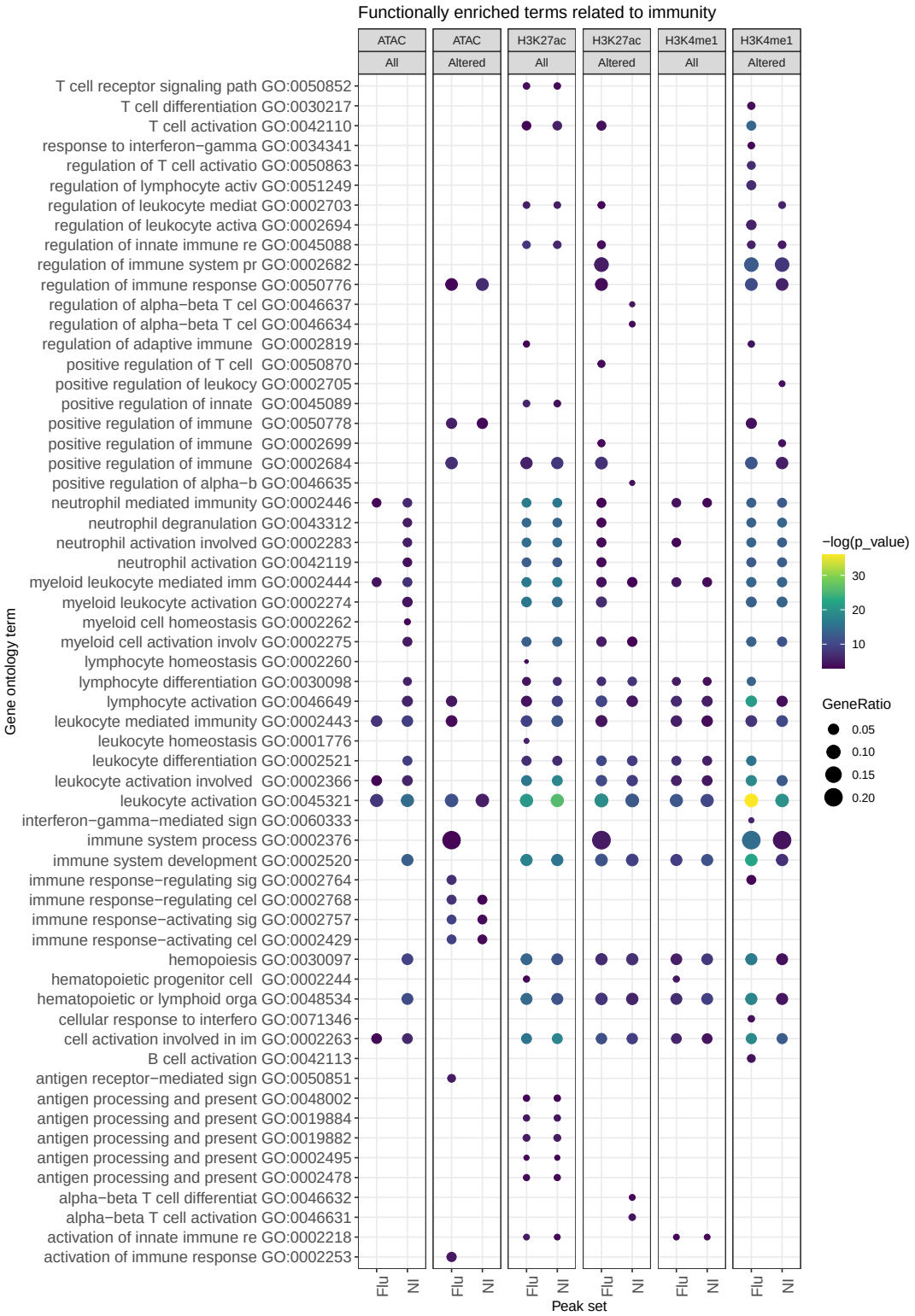

Figure S7: Functional enrichment of genes that are within 10 Kbp of H3K4me1, H3K27ac and ATAC common and altered (graph-only) peaks. We show only the enriched gene ontology terms that relate to immunity and viral infection. The gene ratio is the share of genes that are associated with a GO term.

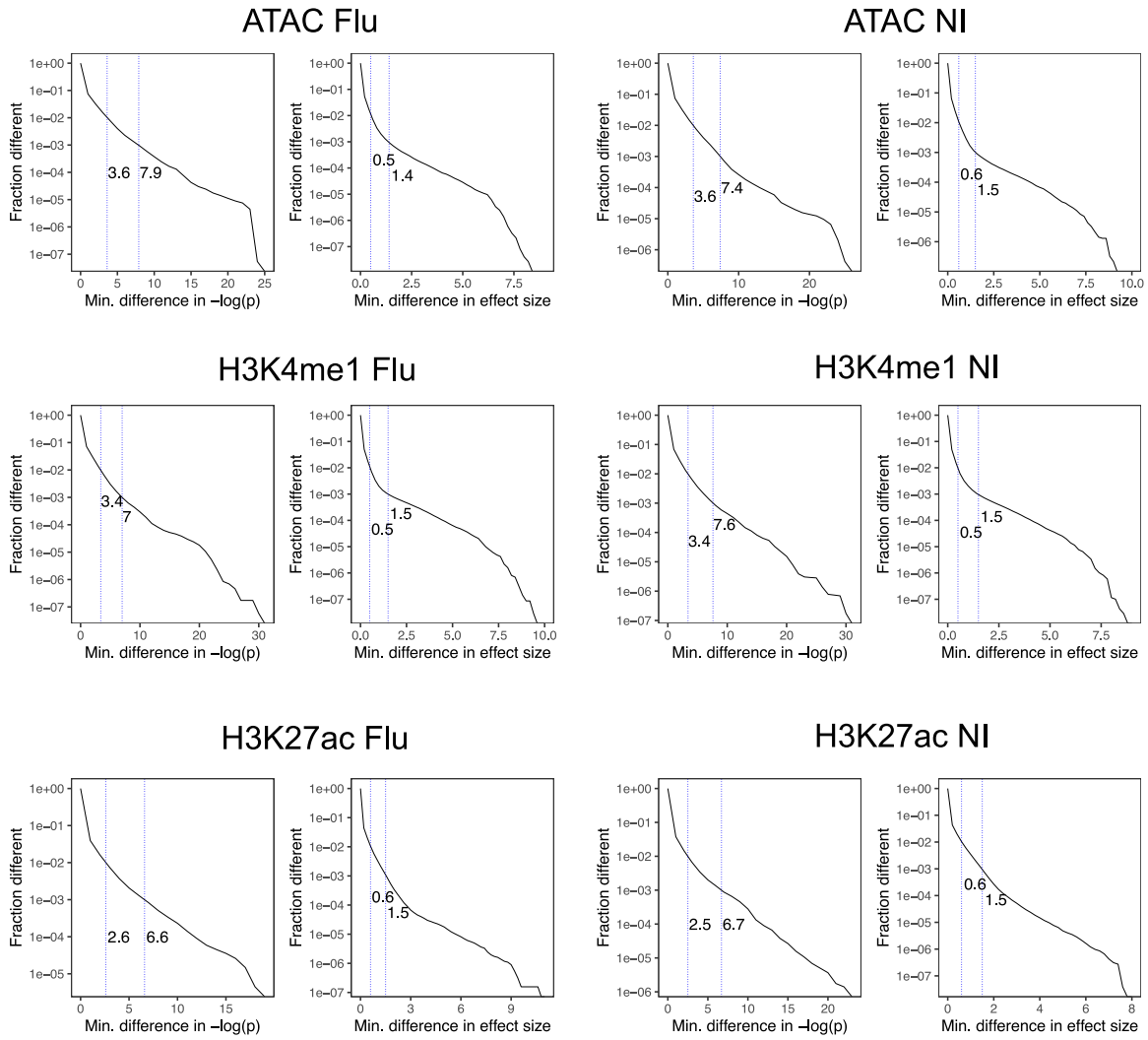

Figure S8: ATAC, H3K4me1, H3K27ac quantitative trait loci are estimated using counts derived from reference and cohort graph alignments in flu-infected and non-infected conditions. We show the fraction of QTLs (y-axis) for which the p-values or effect sizes change by a minimum amount (x-axis). The first line marks the 99th percentile and the second line marks the 99.9th percentile.

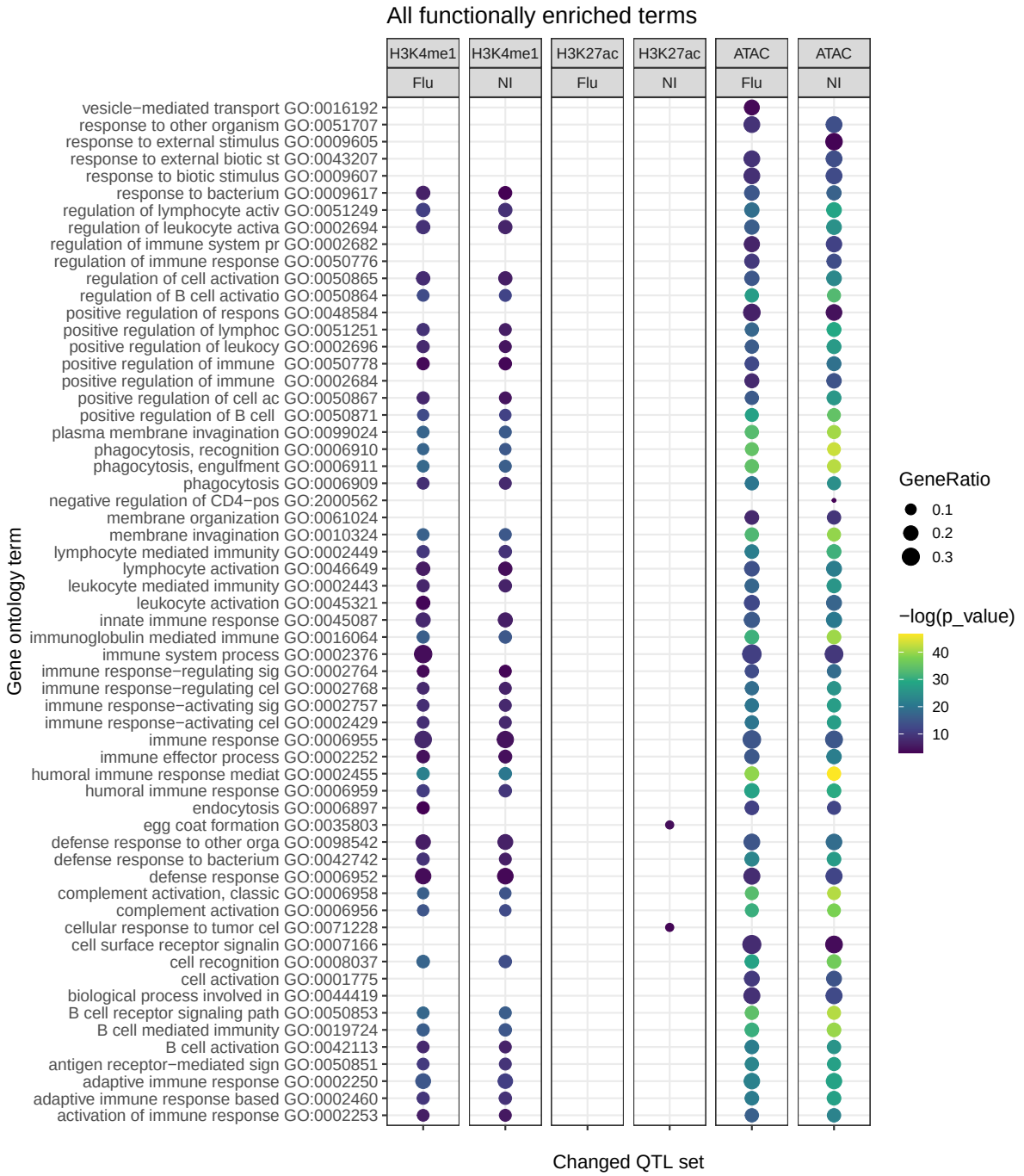

Figure S9: Gene ontology enrichment of genes located within 10 Kbp of peaks associated with hQTLs or caQTLs that change when using a genome graph. Only ca/hQTLs with p-value and effect size changes in the 99.9th percentile were selected. H3K27ac Flu and NI ca/hQTL showed little or no enrichment due a smaller number of ca/hQTLs.

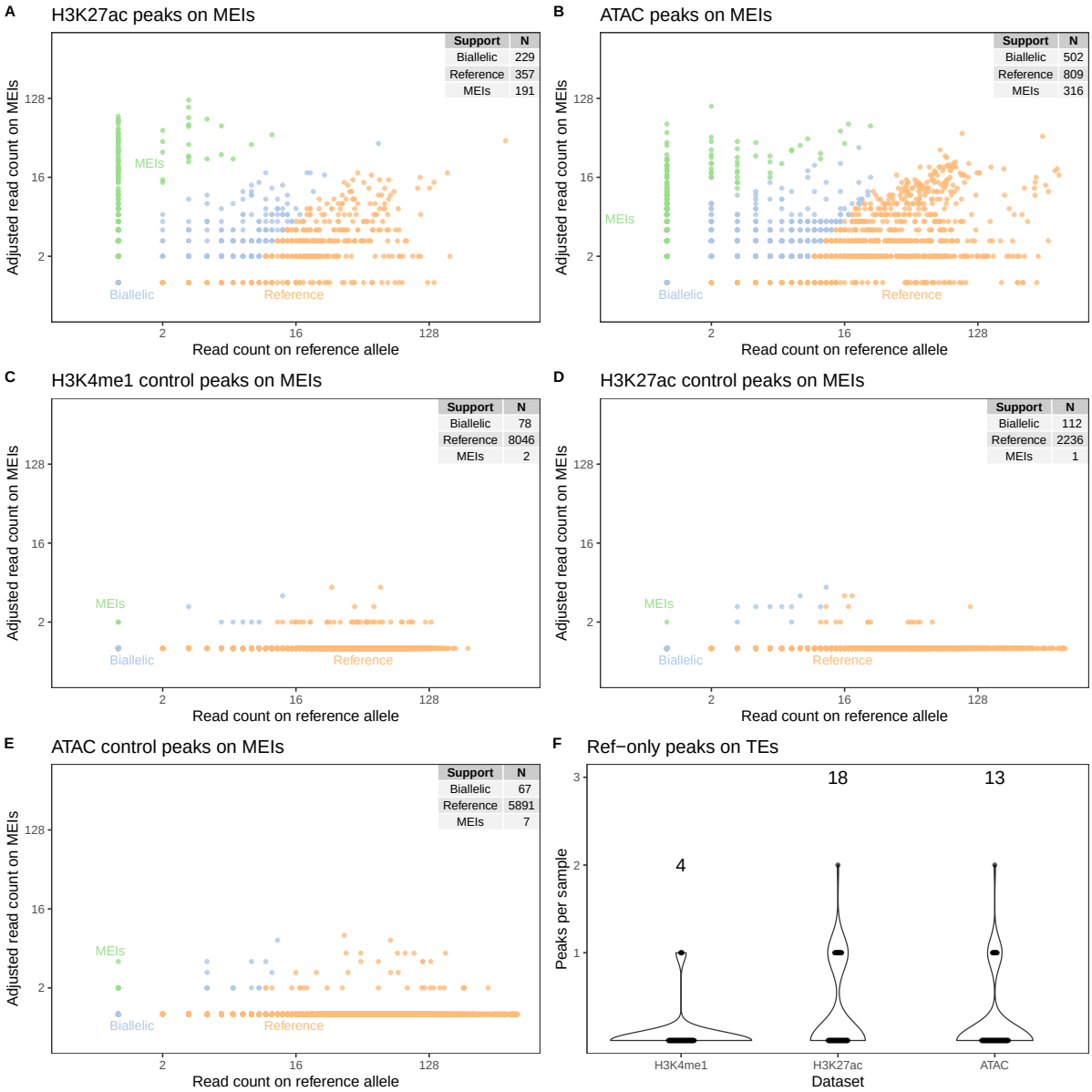

Figure S10: A) B) Partitioning of reads between the reference allele and the alternative allele in peaks that overlap heterozygous or homozygous MEIs in H3K27ac and ATAC-seq. The number of peaks with MEI, biallelic and reference read support is summarized in the table. C) D) E) The same for loci where the samples are known to be homozygous for the reference allele. F) The number of ref-only peaks that overlap MEIs.

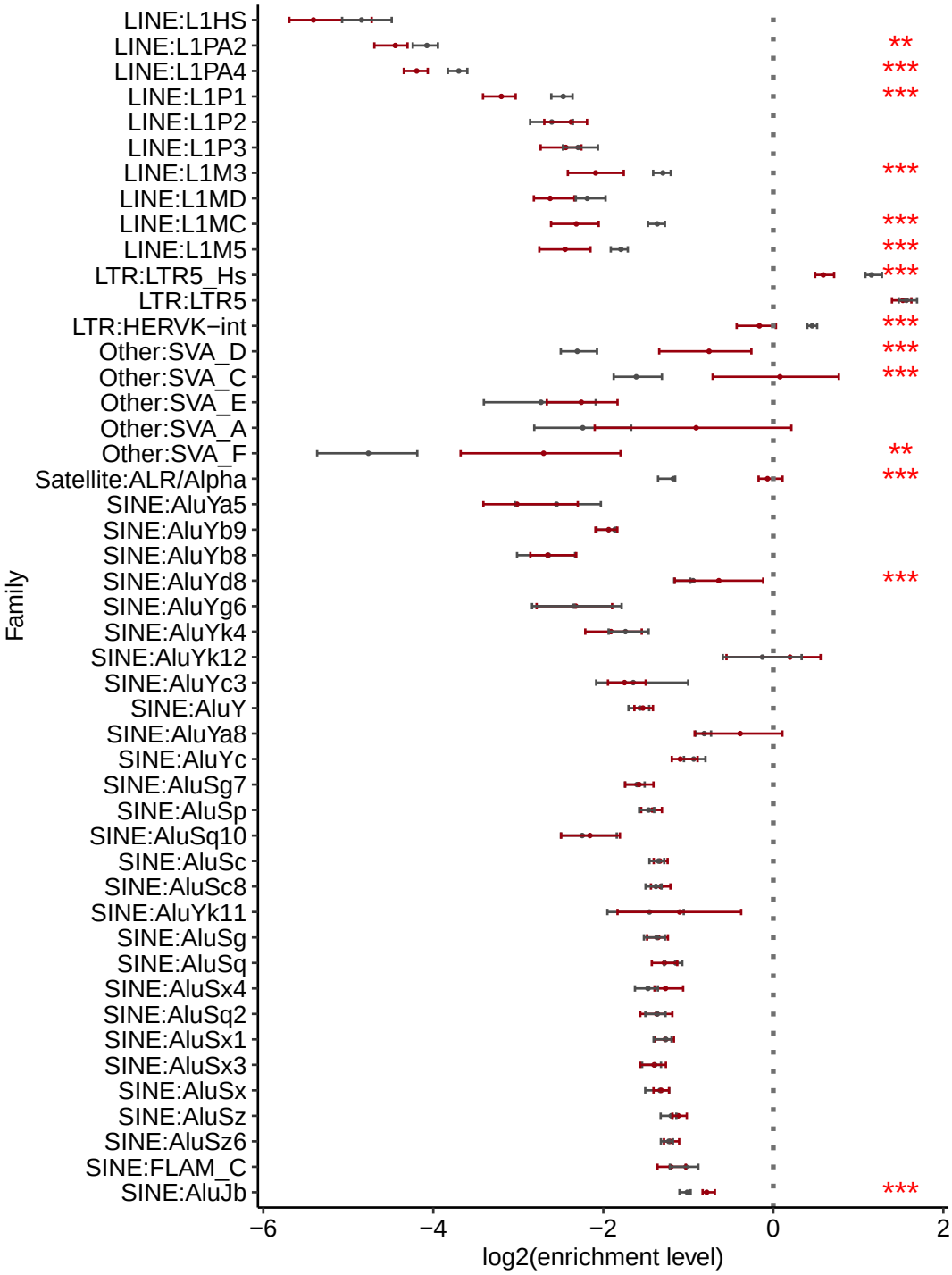

Figure S11: Distribution of enrichment levels of reassembled families within accessible chromatin regions (ATAC) in infected (red) and non-infected (black) samples. SVA families tend become more enriched in accessible chromatin after infection and are highly variable between individuals. Enrichment levels were calculated using methods described by Xun Chen et al [34].

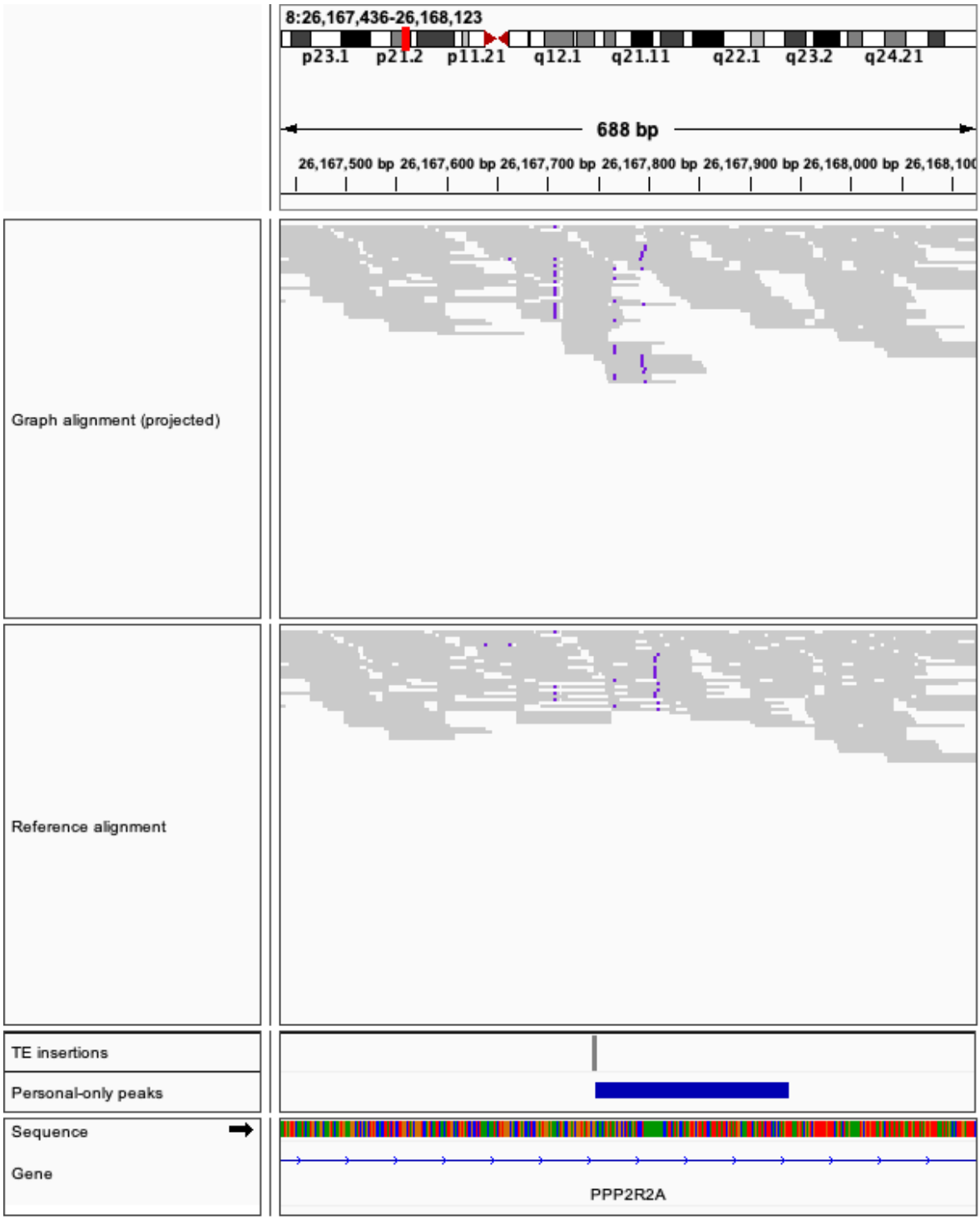

Figure S12: Linear surjection of the example locus (top) versus the reference alignment (bottom) in a Alu polymorphic locus. Note that the actual read pileup is deformed by the surjection to the linear reference. Reads with a MAPQ below 10 have been filtered out. This peak is graph-only in five non-infected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

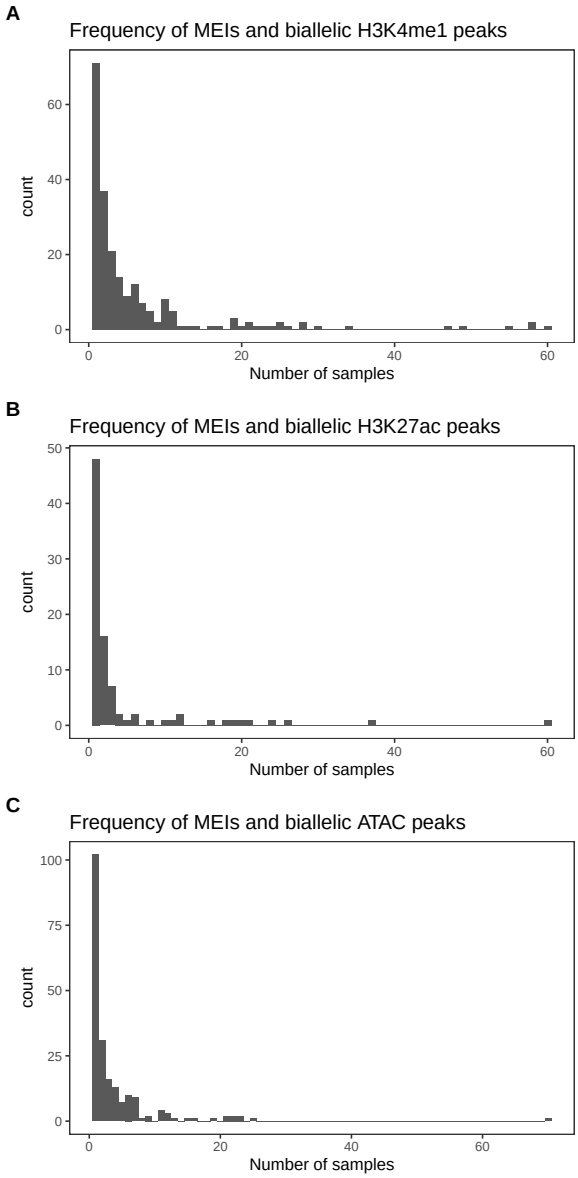

Figure S13: The number times we observe a given MEI and biallelic peak in the cohort at the same locus in A) H3K4me1, B) H3K27ac and C) ATAC-seq.

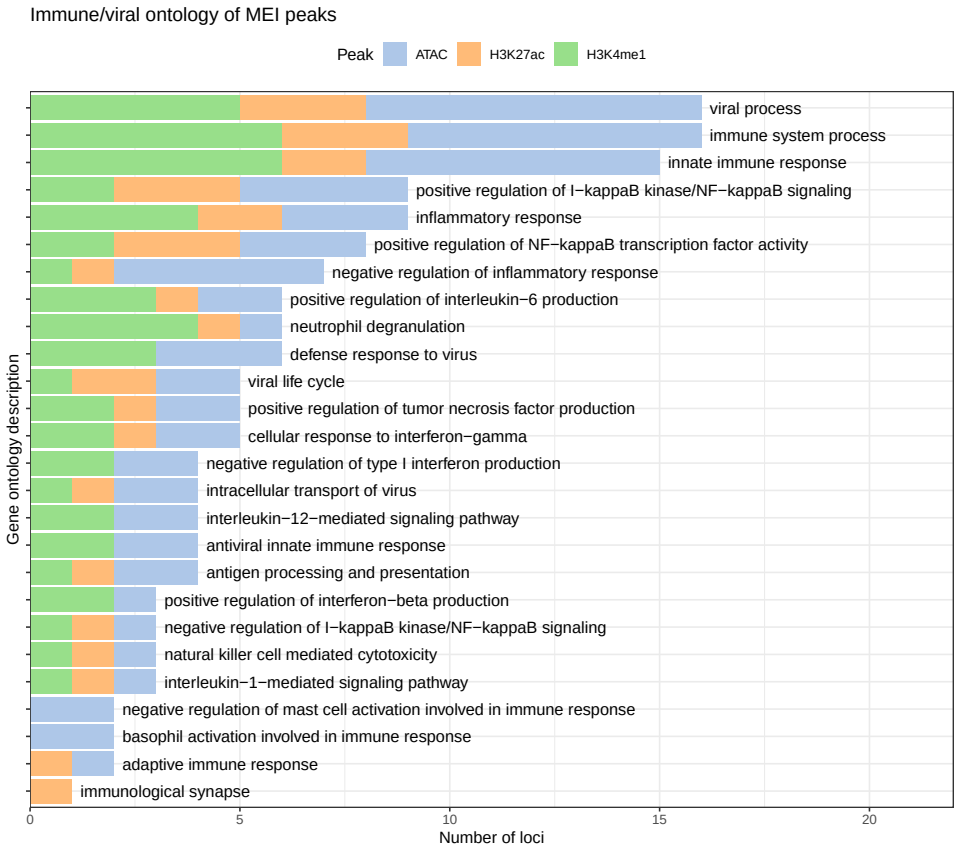

Figure S14: Top ranking GO terms related to immunity and viral infection of genes that are within 10 Kbp of MEI peaks.

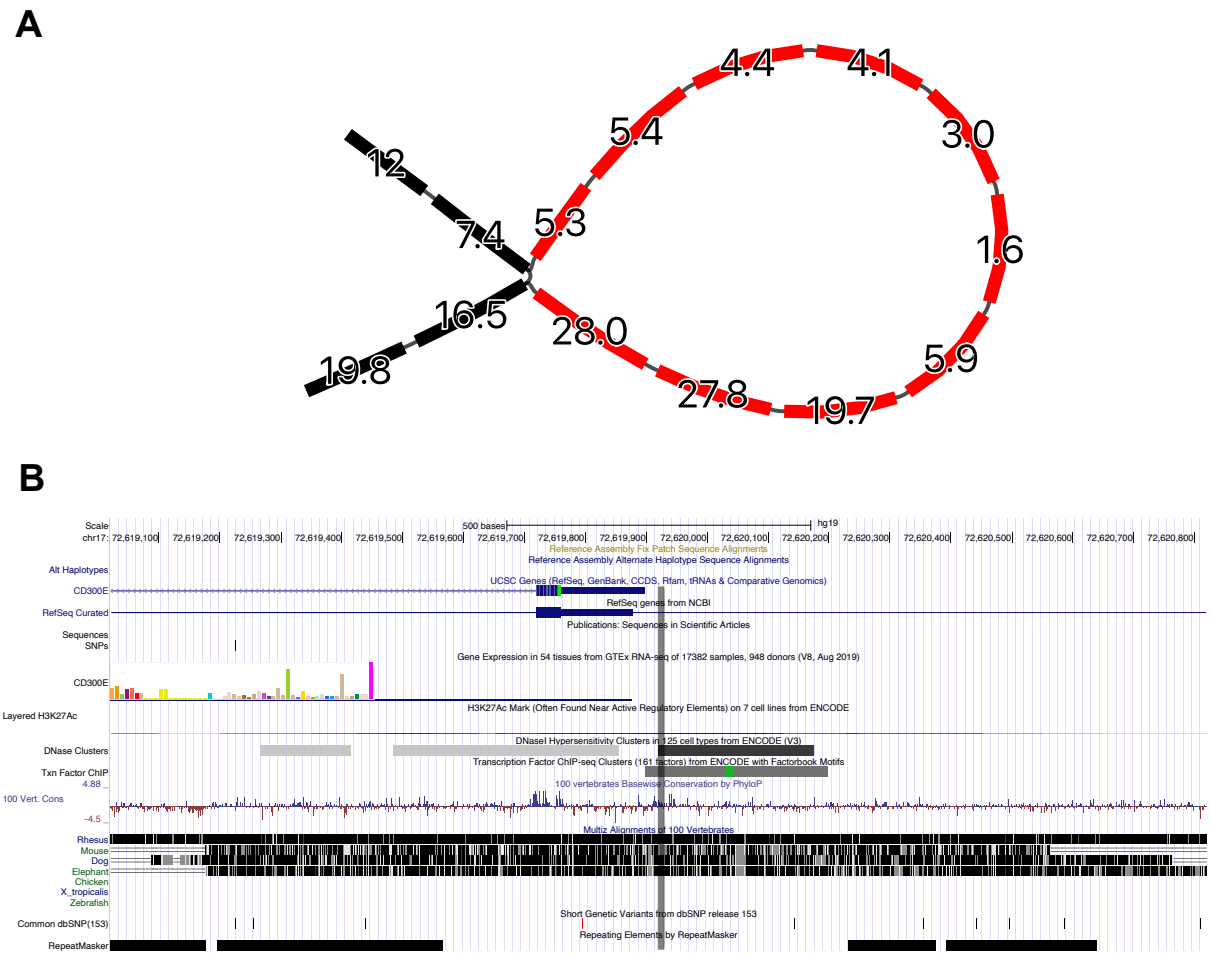

Figure S15: A) Alu mobile element insertion that supports a H3K27ac peak. B) This insertion is immediately upstream of the CD300E gene, and is within a DNase cluster and a ChIP-seq transcription factor cluster. Vertical line marks the site of insertion.

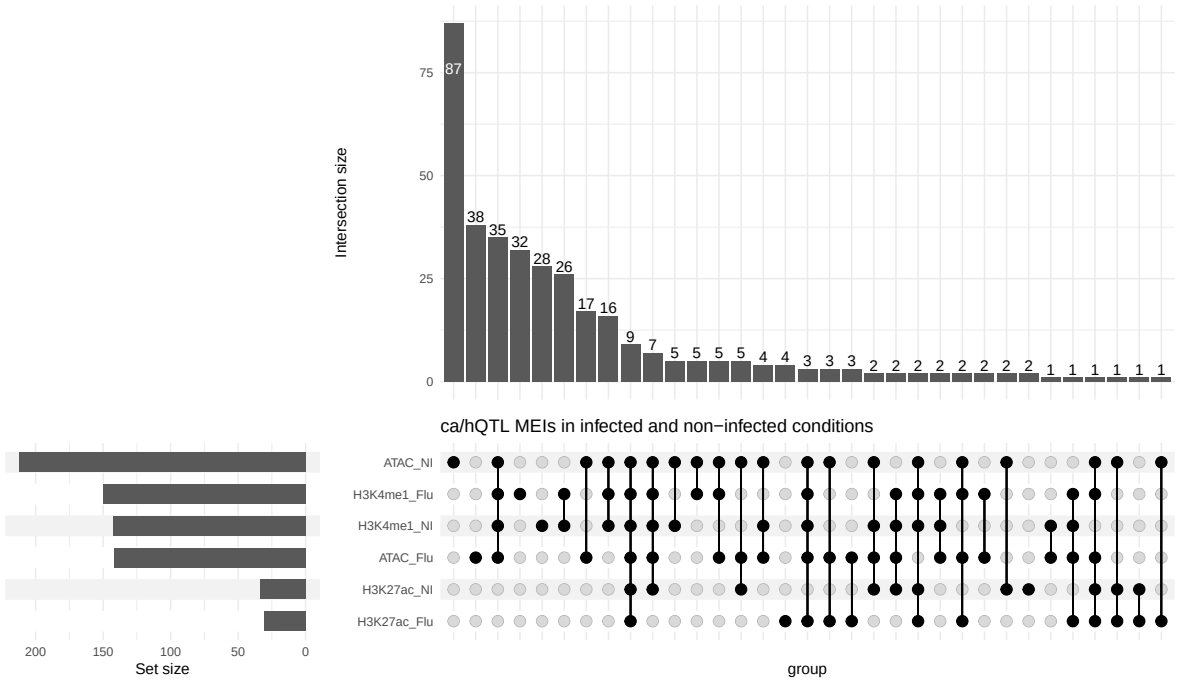

Figure S16: Number of MEIs that are ca/hQTLs in infected and non-infected macrophages.