













Mobile element variation contributes to population-specific genome diversification, gene regulation and disease risk

Received: 13 June 2022

Accepted: 4 April 2023

Published online: 11 May 2023

 Check for updates

Shohei Kojima ¹✉, Satoshi Koyama ^{2,3,4}, Mirei Ka^{1,4,1}, Yuka Saito^{1,5}, Erica H. Parrish¹, Mikiko Endo⁶, Sadaaki Takata ⁶, Misaki Mizukoshi⁶, Keiko Hikino ⁷, Atsushi Takeda^{8,9}, Asami F. Gelinas¹, Steven M. Heaton ¹, Rie Koide¹, Anselmo J. Kamada^{1,4,2}, Michiya Noguchi¹⁰, Michiaki Hamada^{8,9}, Biobank Japan Project Consortium*, Yoichiro Kamatani ^{11,12}, Yasuhiro Murakawa ^{13,14,15}, Kazuyoshi Ishigaki ¹⁶, Yukio Nakamura ¹⁰, Kaoru Ito ², Chikashi Terao ^{12,17,18}, Yukihide Momozawa ⁶ & Nicholas F. Parrish¹✉

Mobile genetic elements (MEs) are heritable mutagens that recursively generate structural variants (SVs). ME variants (MEVs) are difficult to genotype and integrate in statistical genetics, obscuring their impact on genome diversification and traits. We developed a tool that accurately genotypes MEVs using short-read whole-genome sequencing (WGS) and applied it to global human populations. We find unexpected population-specific MEV differences, including an *Alu* insertion distribution distinguishing Japanese from other populations. Integrating MEVs with expression quantitative trait loci (eQTL) maps shows that MEV classes regulate tissue-specific gene expression by shared mechanisms, including creating or attenuating enhancers and recruiting post-transcriptional regulators, supporting class-wide interpretability. MEVs more often associate with gene expression changes than SNVs, thus plausibly impacting traits. Performing genome-wide association study (GWAS) with MEVs pinpoints potential causes of disease risk, including a LINE-1 insertion associated with keloid and fasciitis. This work implicates MEVs as drivers of human divergence and disease risk.

MEs characteristically insert copies of themselves into new genome locations. The evolutionary innovations of MEs are constrained within the linear descent of their host genomes; thus, differences in the sequences, mobilization activity or insertion preferences of the MEs in a particular lineage can increase the rate at which descendant genomes accumulate mutations characteristic of that lineage. In other words, MEs can accelerate genomic divergence. MEs account for a large part of species-specific genomic differentiation¹, but the degree

to which MEs cause species-level phenotypic differences is difficult to dissect due to accumulation of other genetic variation. MEs may also be a force driving speciation, but direct evidence of within-species divergence driven by MEs is limited².

MEs influence the complex traits that differentiate humans and human populations, but our view of this landscape remains partial. Insertions of each of the MEs actively replicating in human genomes—namely, long interspersed nuclear element 1 (L1), SINE-VNTR-*Alu* (SVA)

A full list of affiliations appears at the end of the paper. ✉ e-mail: shohei.kojima@riken.jp; nicholas.parrish@riken.jp

and *Alu* elements—have been implicated in Mendelian diseases³. For example, an SVA insertion so far only reported in the Japanese population causes Fukuyama congenital muscular dystrophy⁴. Individuals carrying a *SLCO1B3* allele with exonic insertion of a proposed Japanese-specific highly active L1 (ref. 5) develop a benign form of hyperbilirubinemia⁶. Recent studies have identified ME polymorphisms associated with differential gene expression^{7–10} and differential polygenic disease risk^{11,12}, but the global influence on human traits remains unclear. MEs make up a large fraction of DNase hypersensitive sites¹³, which are enriched in complex trait heritability¹⁴, and are also the main source of novel regulatory elements in primate genomes¹⁵. Moreover, SVs, about a quarter of which are MEVs^{16,17}, are frequently in tight linkage disequilibrium (LD) with eQTL and trait-associated variants^{17,18}. Actively replicating MEs necessarily carry promoters and transcription-factor binding sites that drive their expression, and some MEs appear to have been coopted as lineage-specific gene regulatory elements^{19,20}. These observations provide a rationale to comprehensively assess the impact of ME polymorphisms on gene expression and complex traits, for example, by performing ME-oriented genotype-trait association studies.

One barrier to ME-phenotype correlation is the low accuracy of current methods used to genotype MEVs, lower than those available for single-nucleotide variants (SNVs) and often too low to derive meaningful hypotheses from statistical genetics approaches. Long-read and strand-specific sequencing are ideal to resolve MEVs and other SVs^{17,21}; however, the number of genomes studied using these methods is low and will remain orders of magnitude lower than those genotyped by short reads until new enabling technologies emerge^{22,23}.

Results

Development and benchmarking of MEGAnE

Accurate variant genotyping is required for statistical genetics. To enable both discovery and accurate MEV genotyping from genomes studied using short reads, we developed a new bioinformatic tool, mobile element genotype analysis environment (MEGANe; Supplementary Note). Compared to SVs resolved by long reads, MEGAnE discovers ME insertions (MEIs) and ME absences with false-positive rates of 3% and 6%, respectively (Fig. 1a and Supplementary Fig. 5). MEGAnE discovers more than 80% of the target-primed reverse transcription (TPRT)-mediated insertions that can be found using long reads, and more than 80% of MEVs are genotyped as accurately as using long reads or a graph-based genotyper. Less than 2% of genotype calls are inconsistent with Mendelian inheritance (Supplementary Fig. 9). To test the genotyping quality of MEGAnE by an orthogonal approach, we deep sequenced over 100 MEV target sites using DNA from 2,221 Japanese individuals. More than 95% of genotype calls were concordant with those determined by targeted deep sequencing (Fig. 1b,c and Supplementary Figs. 10–15). Accurate genotyping allows us to assign MEVs to haplotypes better than alternatives (Supplementary Fig. 16); more than 90% of ME genotypes imputed using MEGAnE's output were highly concordant with those inferred using graph-based pangenome references (Fig. 1b and Supplementary Figs. 16 and 17). Although read length imposes some intrinsic limitations on MEV discovery, the low false-positive rate and accurate genotyping of this tool enabled us to interrogate MEVs in short-read data at a resolution that was previously impossible.

Fig. 1 | Discovery and accurate genotyping of MEVs in global and Japanese populations. a, Concordance between MEV genotype called by MEGAnE and an SV callset generated by Phased Assembly Variant caller in 34 individuals. Dot color represents R^2 between the two genotyping results. **b**, Concordance between allele frequency called by MEGAnE, or imputed based on MEGAnE calls, and targeted deep sequencing. Genotypes of MEIs in 888 Japanese individuals were directly called by MEGAnE or imputed using haplotypes in the 1000GP and compared to those assessed by targeted deep sequencing. A total of 54 *Alu*, 27 L1, 9 SVA and 1 human endogenous retrovirus (HERV)-K were analyzed.

Characteristics of MEVs in diverse populations and Japanese

We applied MEGAnE to the 2,504 and 1,235 individuals sequenced at high coverage (30× and 25×) in the 1000 Genomes Project (1000GP) and BioBank Japan (BBJ), respectively. We detected 48,360 and 10,996 MEVs in these respective cohorts, with around 2,500 to 3,000 polymorphisms per individual (Supplementary Fig. 19). The top eight principal components (PCs) of MEVs were highly correlated with those of SNVs; like SNVs, MEVs reflect the geographical distribution of human populations (Fig. 1d and Supplementary Fig. 20). MEVs are more abundant in Africans, as are population-specific MEVs (Fig. 1e,f). Population-specific L1 and SVA are more abundant in East Asians, particularly in Japanese, than other non-African populations, whereas the abundance of *Alu* is similar (Fig. 1f). Over half of the MEVs observed as Japanese-specific singletons within 1000GP, which sequenced 104 Japanese individuals, were observed in other participants in BBJ (Supplementary Fig. 21). As expected, MEVs predominantly involve young elements known to be active germline mutagens (*Alu*, L1 and SVA) (Fig. 1g, Supplementary Fig. 31 and Supplementary Table 1).

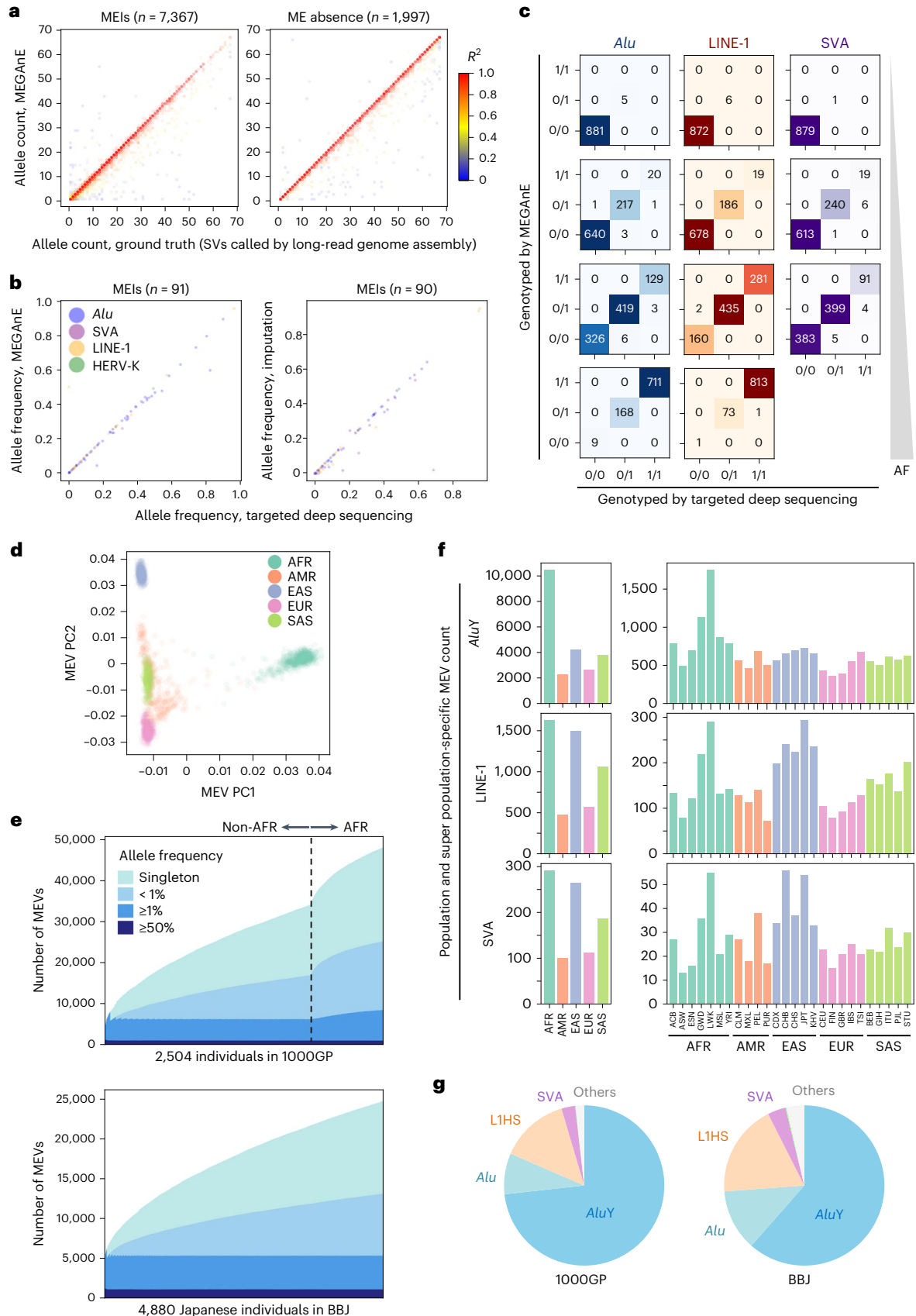
Fixed MEs enrich in distinct genome regions²⁴. To assess the genomic niches occupied by MEVs, we correlated MEV occurrences with genome features measured in H1-hESCs (Fig. 2a,b). L1 polymorphisms are positively correlated with markers of heterochromatin, such as DNA methylation and H3K9me3. SVA polymorphisms show the opposite trend, occurring more often in regions with active chromatin markers, such as H3K9ac and early replication timing. To reduce the degree to which selection may influence this observation, we also analyzed the association with rare, presumably recently acquired insertions. Singletons found in the 1000GP and BBJ exhibit a similar trend; polymorphisms of L1 and SVA show positive and negative correlations to heterochromatin markers, respectively (Supplementary Fig. 32a). In addition to singletons, which may have higher false-positive rate than non-singletons, we also used 15,718 family-specific heritable insertions (those private to a family yet inherited by at least one offspring) found in Simons Foundation Autism Research Initiative (SFARI) datasets (Supplementary Fig. 32b). These show the same trend, suggesting that this distribution results from biased insertion, rather than a consequence of selection or technical bias. The opposite insertional bias of these two MEs, which employ the same molecular machinery for insertion (ORF2p of L1), suggests that other factors, such as recruitment of different RNA-binding protein partners, influence insertional preference. Considering that L1 expression is a prerequisite for SVA transposition, different expression patterns of these RNAs in the context of germline development are unlikely to fully account for this difference. As previously reported, L1 and SVA MEVs exhibit the same motif at insertion breakpoints (T/AAAA; Supplementary Fig. 33), suggesting that the difference of insertion bias is not due to the differences in local sequence recognition by endonuclease.

Alu insertions from 1000GP and SFARI show weak enrichment in late-replicating domains, whereas this trend is mitigated in BBJ, suggesting that the insertion bias of *Alu* may differ between human populations (Fig. 2b,c). To examine this more closely, we focused on population-specific *Alu* insertions in 1000GP. Compared to other populations' specific *Alu* insertions, *Alu* found only in JPT show an opposite trend, occurring slightly more often in early-replicating domains (Fig. 2d). This is not a consequence of differences in the chromatin

c, Examples of MEV genotypes called by MEGAnE and targeted deep sequencing. **d**, Distribution of first two PCs of MEVs discovered in the 1000GP. Color indicates superpopulation. **e**, Discovery of MEVs from diverse populations in the 1000GP (top) and Japanese in BBJ (bottom). The color of bar plots is stratified based on allele frequency of MEVs. **f**, The number of superpopulation-specific (left three panels) and population-specific (right three panels) MEVs found in the 1000GP. **g**, Proportion of ME families found in 1000GP (left) and BBJ (right). In this figure, *Alu* represents *Alu* subfamilies other than *AluY*. AFR, African; AMR, American; EAS, East Asian; EUR, European; SAS, South Asian.

organization of Japanese individuals' genomes, at least as inferred from CpG methylation (Supplementary Fig. 34c). At the continental superpopulation level, *Alu* insertions specific to AFR, AMR, or EUR populations are more biased towards late-replicating domains than

those found only in the EAS population (Fig. 2e and Supplementary Fig. 34). Differences in *Alu* insertion distribution could result from various causes, including drift, selection and differences in *Alu* insertional mutations. However, when restricting this analysis to rare variants



expected to reflect mutational processes, rare *Alu* elements ($AF < 0.1\%$) in BBJ participants were distributed in earlier-replicating regions compared to those in PC-inferred Europeans in SFARI (Fig. 2f). Although we are unable to fully exclude the contribution of population-specific differences in selection acting on *Alu* insertions, we interpret these differences to suggest that *Alu* insertion preference has shifted in East Asians.

Regulatory effects depend on ME ontology and genomic context

To understand the consequences of MEVs on gene expression, we imputed MEVs in 838 individuals in GTEx and performed eQTL mapping in 49 tissues using both MEVs and SNVs. We defined 'ME-eQTLs' as MEVs that are either the lead variants or are in high LD with (hereafter, 'tagged') lead SNVs ($r^2 > 0.95$). After cross-tissue meta-analysis, we detected 1,073 ME-eQTLs consisting of 778 different MEVs. MEVs were the lead variants of 483 ME-eQTLs in at least one tissue (Fig. 3a). More than 60% of detected ME-eQTLs are tissue-specific (Supplementary Fig. 37a), and the tissue in which the most tissue-specific and total ME-eQTLs were detected was testis, consistent with frequent de-repression of MEs in this tissue (Fig. 3a)²⁵. MEVs were 1.2 times more frequently found in LD ($r^2 > 0.8$) with sentinel variants in testis eQTLs than SNVs (Fig. 3b; $P < 0.0001$), suggesting MEVs are a major factor creating variation of gene expression in testis and potentially other tissues (Supplementary Fig. 38).

In addition to tissue, gene regulatory effects of MEVs plausibly depend on the type of ME and the functional and epigenetic context of the genome, and ME-eQTLs allow us to dissect such determinants. MEVs in regions with active histone marks, such as H3K4me3, and accessible chromatin (represented as early-replicating domains and A compartments) are frequently ME-eQTLs. MEVs in exons, promoters (defined as 1 kb upstream of transcription start site), and introns are more often ME-eQTL, whereas those in intergenic regions are less likely to be detected as ME-eQTLs (Fig. 3c). Concordantly, ME-eQTLs in exons or promoter regions have larger effects than those in introns or intergenic regions (Fig. 3d). Consistent with the enrichment of genes in early-replicating domains, MEVs in early-replicating domains are more likely to associate with gene expression than those in late-replicating domains. Even when accounting for the increased number of MEV-gene pairs in early-replicating domains, the same trend was observed (Supplementary Fig. 39a). Together this indicates that MEVs in transcriptionally active regions, regulatory elements, and accessible chromatin often influence gene regulation.

Full-length *Alu* elements contain a Pol-III promoter, whereas LI, SVA, and human endogenous retrovirus (HERV)-K harbor Pol-II promoters. When comparing the distribution of the effect sizes of *Alu* ME-eQTLs to ME-eQTLs with Pol-II promoter-containing MEs, the latter have larger positive effects, but there was no clear difference when comparing the negative effects, suggesting that MEVs with a Pol-II promoter often function as enhancers of nearby genes (Fig. 3e,f). At the ME family level, SVA is more frequently an ME-eQTL in multiple tissues than *Alu* (Supplementary Fig. 37b, two-sided Fisher exact test,

$P = 0.046$), consistent with SVA having a more ubiquitous influence on nearby genes^{26,27}. Thus, MEs exert different gene regulatory functions depending on ME family and genomic context.

Compared to permutation, ME-eQTLs are more than twice as often found in high LD with SNVs in the GWAS Catalog than expected for non-eQTL MEVs (Fig. 3g), suggesting that MEV-associated modulation of gene expression could result in differences in complex traits; thus, the integration of ME-eQTLs with GWAS could help refine hypotheses about the molecular mechanisms driving complex traits. Moreover, the observation that MEVs regulate gene function based on ME family and context supports the possibility of interpreting (for example predicting the anticipated regulatory consequences of) some MEVs, a major challenge for other non-coding variants.

MEVs often attenuate enhancers

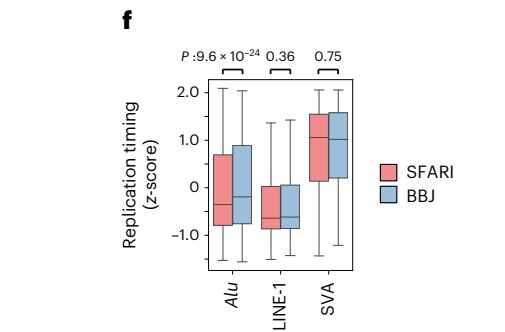
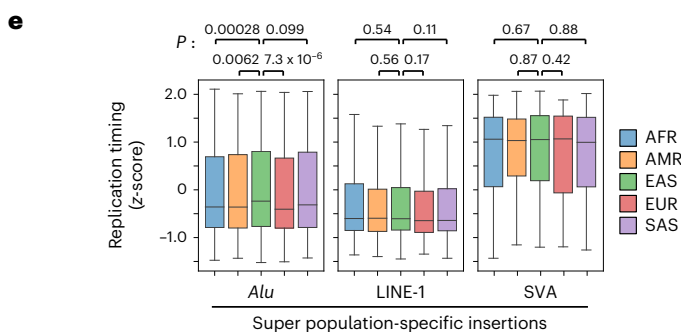
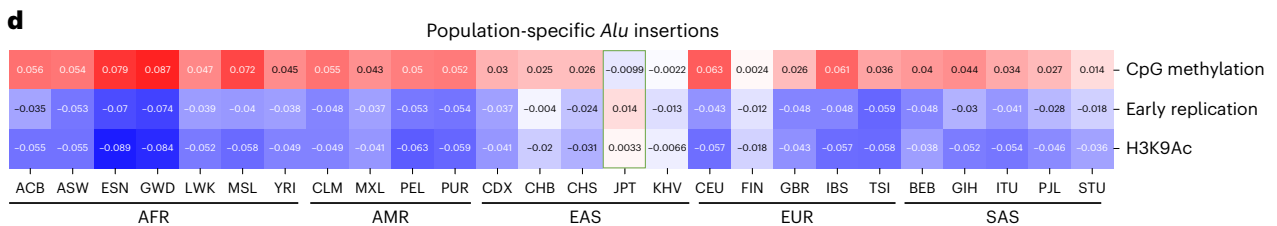
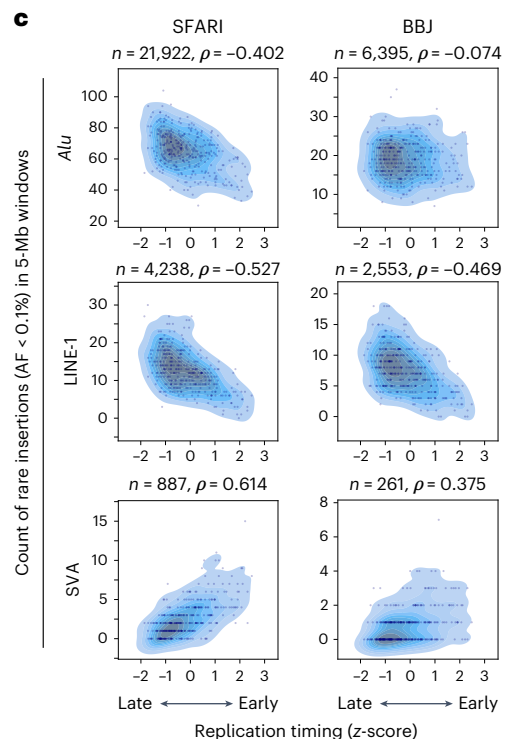
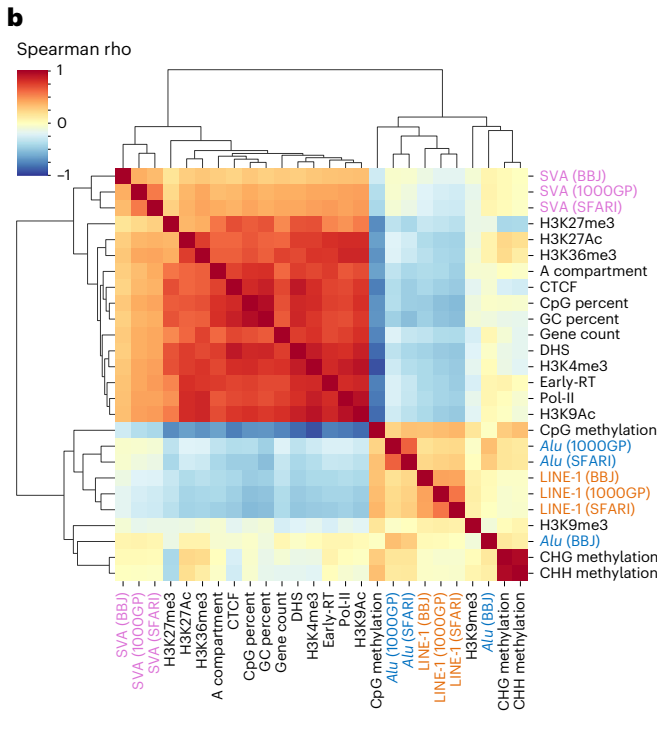
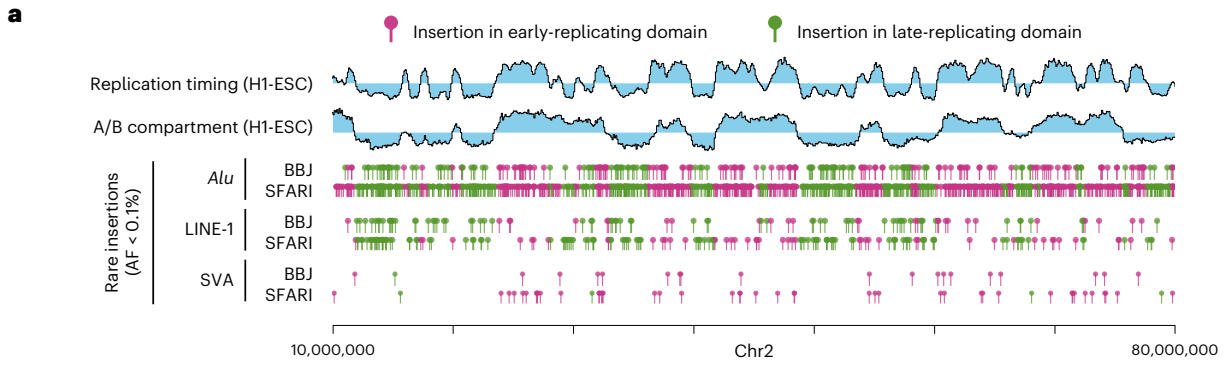
Although MEVs with Pol-II promoters often associate with increased expression of nearby genes, some MEVs have negative effects. We hypothesized that ME insertion into an existing gene regulatory element can attenuate that element's regulatory function, analogous to ME insertion into protein-coding exons generating hypomorphic and loss-of-function alleles. 45 out of 688 MEI-eQTLs fall into distal enhancer-like signatures (dELS) in the ENCODE cCRE dataset. Of these 45 ME-eQTLs, 30 were associated with negative regulation of nearby genes, compared to only 13 with upregulation (Fig. 4a; $P = 0.007$, Fisher exact test), suggesting that ME insertions into enhancers often decrease their enhancing activity. To test this, we studied an *Alu* insertion in dELS between genes *DGKE* and *TRIM25* (Fig. 4b). This 297-bp insertion overlaps with a DNase hypersensitive site detected in LCLs, is the lead variant in a *DGKE* eQTL, and is in high LD with the lead variant in a *TRIM25* eQTL, both in LCLs (Fig. 4c,d; $r^2 = 0.98$). For both eQTLs, the *Alu* insertion haplotype is associated with decreased gene expression, suggesting that the insertion attenuates enhancer activity. Consistent with this model, the dELS shows enhancer activity in LCLs, whereas the insertion of *Alu* reduced the reporter activity by half (Fig. 4e). This pattern, of *Alu* insertions into dELS associating with decreased expression of genes presumably regulated by these enhancers, is observed at multiple loci (for example, Supplementary Fig. 40).

Coherent regulation of gene expression by 3'UTR MEVs

In GTEx, 71 MEVs in 3'UTRs of protein-coding genes were used for eQTL mapping. Of these, 20 *Alu* were observed as ME-eQTLs of the genes; 16 were ME-eQTLs in two or more tissues. *Alu* in 3'UTR tended to associate with decreased gene expression (Fig. 5a–d). An *Alu* insertion in the 3'UTR of *HSD17B12* was previously reported to downregulate that gene's expression in iPSCs and LCLs⁷. This association was replicated in 40 tissues, including LCLs (Fig. 5b–d). To test whether other *Alu* insertions cause differential gene expression, we cloned 3'UTRs of *ADIPOQ* and *MAP3K21* genes (*Alu-ADIPOQ* and *Alu-MAP3K21*, respectively) in a reporter plasmid and generated isogenic controls lacking the *Alu* sequence. The *Alu-ADIPOQ* decreased reporter expression in LCLs, supporting the MEV as causal for the observed association (Fig. 5b–d).

Fig. 2 | Biased distribution of MEVs. **a**, Example of positional distribution of rare MEIs found in BBJ ($n = 4,880$ individuals) and individuals of PC-inferred European ancestry in SFARI ($n = 7,642$). Insertion sites of rare MEIs ($AF < 0.1\%$) in a 70-Mb region of chromosome 2 are shown. **b**, Heatmap showing correlations between the numbers of MEIs discovered from the 1000GP, SFARI and BBJ, and genome features of nonoverlapping 1-Mb windows measured in HI-hESCs. Dendrograms show results of hierarchical clustering. **c**, Distribution of replication timing and number of rare MEIs in nonoverlapping 5-Mb windows. Left three panels show the distributions of MEIs found in individuals of PC-inferred European ancestry ($n = 7,642$) in SFARI, whereas the right three panels show those of Japanese in BBJ. Kernel density of data points is shown with the actual data points. Spearman's correlation coefficients (ρ) are shown. **d**, Heatmap showing correlations between the number of population-specific MEIs discovered from the 1000GP and

genome features of nonoverlapping 1-Mb windows. Japanese in Tokyo (JPT) are highlighted by a green box. **e, f**, Distribution of replication timing of the regions in which superpopulation-specific MEVs are observed in 1000GP (**e**) or rare MEVs ($AF < 0.1\%$) found in the individuals of PC-inferred European ancestry in SFARI and BBJ (**f**). P of two-sided t -test is shown. Middle line of box plot represents median, and lower and upper whiskers represent the lowest datum above $Q1 - 1.5 \times IQR$ and highest below $Q3 + 1.5 \times IQR$, respectively, where $Q1$, $Q3$ and IQR are the first and third quartiles and interquartile range, respectively. **e**, *Alu*: $n = 11,500$ (AFR), 2,558 (AMR), 4,777 (EAS), 2,938 (EUR), 4,209 (SAS); LI: $n = 1,636$ (AFR), 483 (AMR), 1,508 (EAS), 579 (EUR), 1,094 (SAS); SVA: $n = 370$ (AFR), 122 (AMR), 317 (EAS), 140 (EUR), 235 (SAS). **f**, BBJ: $n = 10,160$ (*Alu*), 3,883 (LI), 509 (SVA); SFARI: $n = 23,606$ (*Alu*), 4,581 (LI), 1,184 (SVA).



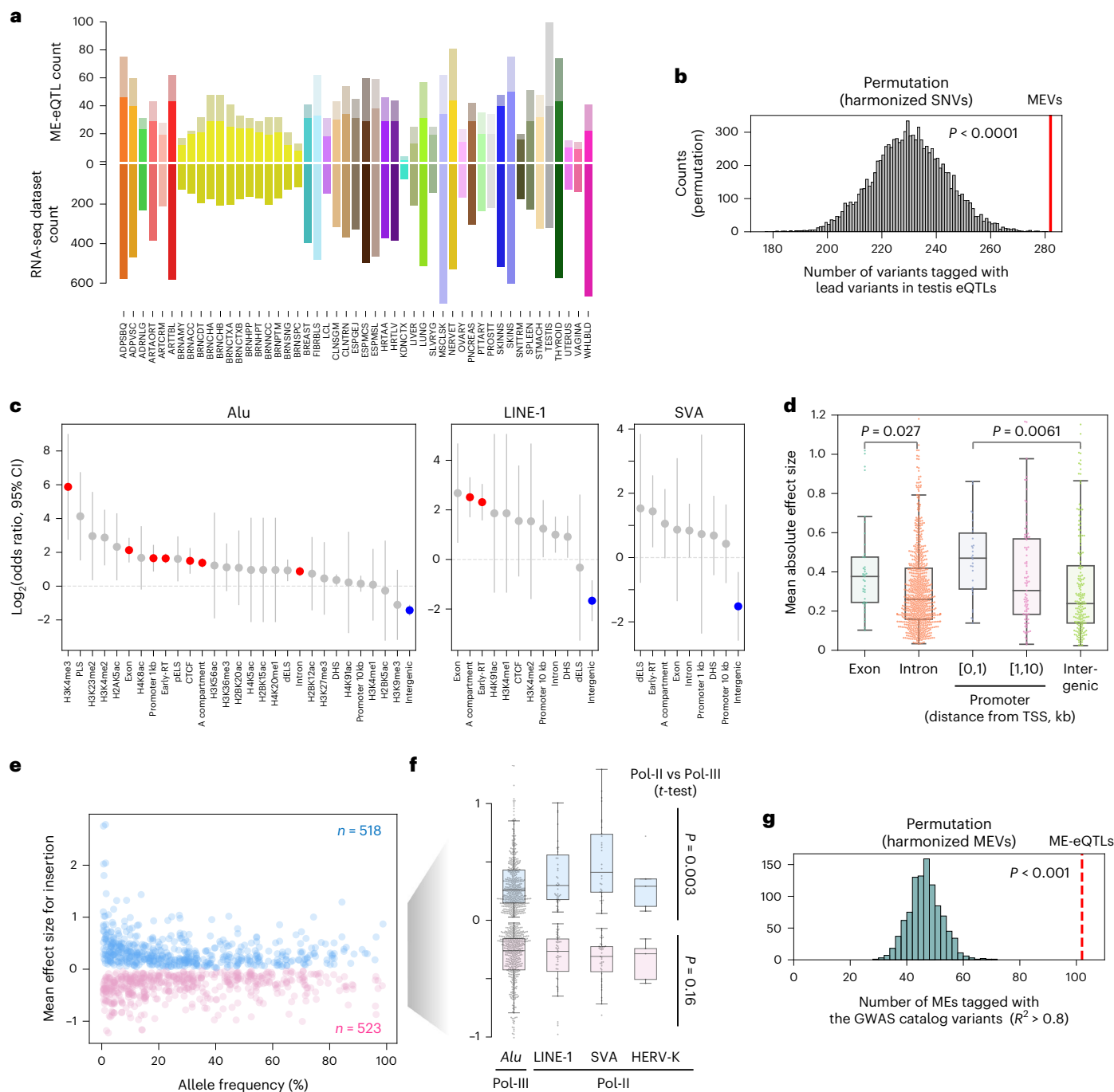


Fig. 3 | eQTL analysis with MEVs. **a**, Number of ME-eQTLs detected in GTEx. Bars with bright and subdued color in the top panel represent the number of multitissue and tissue-specific ME-eQTLs, respectively. The bottom panel shows the number of RNA-sequencing datasets used for eQTL analysis. Bar color represents tissue, specified along the horizontal axis. **b**, The number of MEVs in testis eQTLs. Histogram shows the result of harmonized SNVs by 10,000 permutations. Red line shows the actual number of MEVs tagged by lead variants in testis eQTLs. Empirical *P* of one-sided permutation test is shown. **c**, Odds ratios that an ME observed within a designated genome region is detected as an ME-eQTL. Red and blue points are significant enrichments or depletions (two-sided Fisher exact test *P* < 0.05 after Bonferroni correction). Odds ratios and these 95% confidence intervals (CIs) are shown. RT, replication timing. *n* = 7,859 (*Alu*), 1,108 (*LI*), 653 (*SVA*). **d**, Distribution of effect sizes of ME-eQTLs intersecting designated genome features. *n* = 39 (exon), 726 (intron),

30 (promoter 1 kb), 87 (promoter 1–10 kb), 190 (intergenic). TSS, transcription start site. **e**, Distribution of allele frequencies and effect sizes of ME-eQTLs. Effect sizes for presence of an ME are shown. Positive effect size: *n* = 430 (Pol-III), 78 (Pol-II); negative effect size: *n* = 418 (Pol-III), 92 (Pol-II). **f**, Distribution of effect sizes of ME-eQTLs by ME families. **d**, **f**, Two-sided *t*-test *P*s are shown. Middle line of box plot represents median, and lower and upper whiskers represent the lowest data point above $Q1 - 1.5 \times IQR$ and highest below $Q3 + 1.5 \times IQR$, respectively, where $Q1$, $Q3$ and IQR are the first and third quartiles and interquartile range, respectively. **d**–**f**, If a given ME-eQTL is detected in multiple tissues, the mean of the effect sizes across tissues was used for visualization. **e**, **f**, Thirty-two ME-eQTLs that have both positive and negative effects, differing by tissue, were excluded. **g**, The number of MEVs in ME-eQTLs in LD with variants in the GWAS Catalog. Histogram shows the result of 1,000 permutations. Red line shows the actual number of MEVs tagged by GWAS Catalog variants. Empirical *P* of one-sided permutation test is shown.

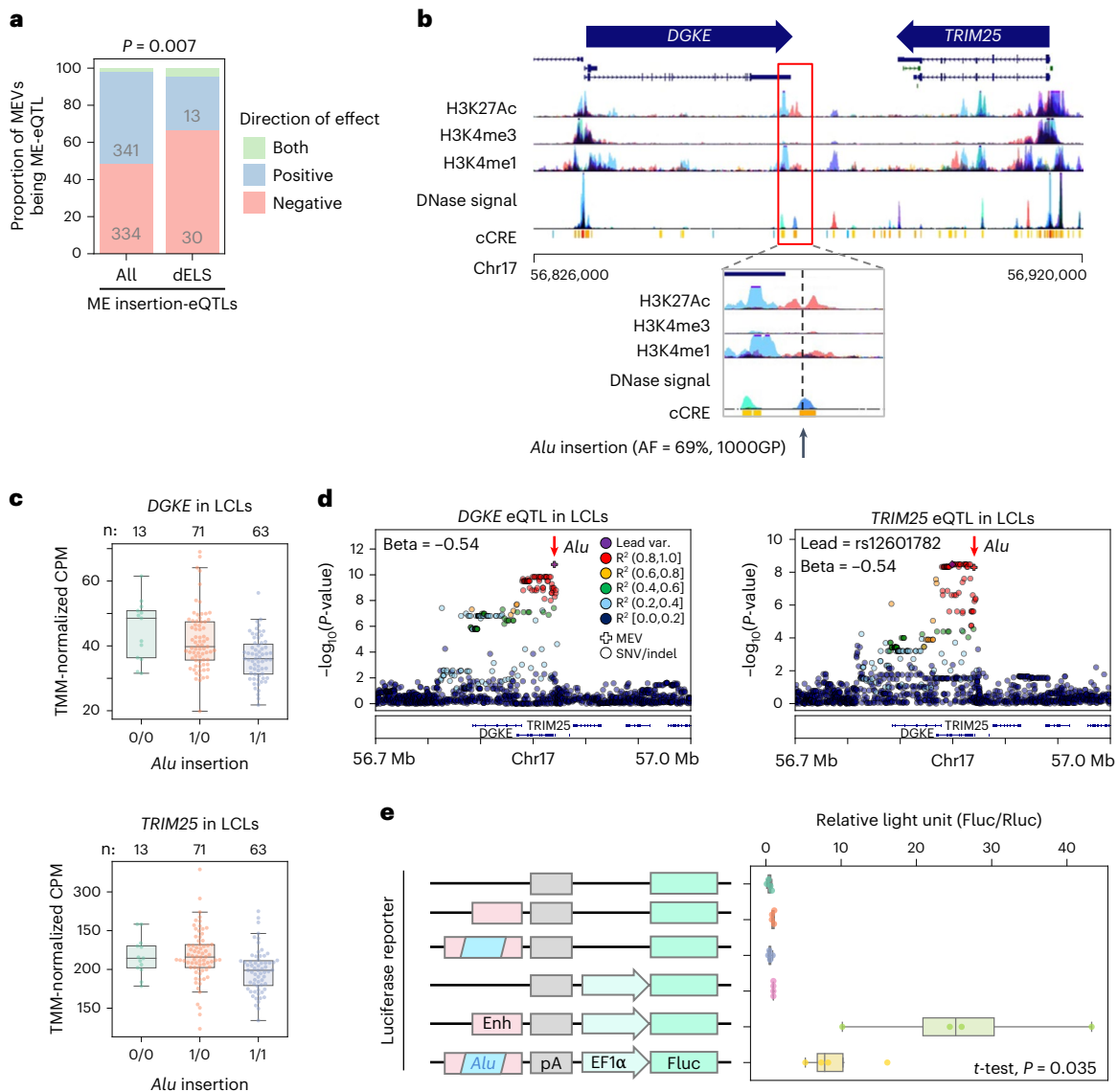


Fig. 4 | *Alu* insertions in regulatory elements. **a**, Comparison between MEVs detected as ME-eQTLs genome-wide and those inserted in distal enhancer-like signature (dELS) candidate *cis*-regulatory elements (cCRE). Gray numbers in the bar plot show the counts of MEVs used for analysis. *P* of Fisher exact test of whether ME-eQTL variants in dELS more often have negative effect is shown. **b**, UCSC genome browser view showing position of an *Alu* insertion in an enhancer-like sequence near *DGKE* and *TRIM25* genes. The position of the *Alu* insertion is shown with an arrow and a vertical dashed line. **c**, Expression levels of *DGKE* and *TRIM25* in LCLs. Trimmed mean of M value (TMM)-normalized CPM grouped by genotypes of *Alu* insertion are shown. Numbers of data points are shown in figures. CPM, count per million. **d**, Regional association plots showing *DGKE*-eQTL and *TRIM25*-eQTL. MEVs and SNPs are shown as plus marks and

circles, respectively. The *Alu* insertions are highlighted with red arrows. *P* values were calculated by fastQTL. **e**, Dual luciferase reporter assay of the enhancer with or without *Alu* insertion. The illustration shows the structure of Firefly luciferase reporter plasmids. The enhancer and the *Alu* insertion found near *DGKE* and *TRIM25* are drawn in red and blue. Plasmids were transfected into a lymphoblastoid cell line (LCL; GM12878). *P* of two-sided *t*-test between the activities of the bottom two constructs is shown. *n* = 4 independent experiments. **c, e**, Middle line of box plot represents median, and lower and upper whiskers represent the lowest data point above $Q1 - 1.5 \times IQR$ and highest below $Q3 + 1.5 \times IQR$, respectively, where $Q1$, $Q3$, IQR are the first and third quartiles and interquartile range, respectively.

Although *Alu*-ADIPOQ was not detected as an ME-eQTL in LCLs, it is detected as an eQTL in all tissues in which *ADIPOQ* is highly expressed. On the other hand, *Alu*-MAP3K21 increased reporter expression in oligodendroglia cells and basal neuroectoderm-like NT2/D1 cells, but not in LCLs (Fig. 5d–f). This is consistent with the ME-eQTL mapping results; although *MAP3K21* is expressed in other tissues, *Alu*-MAP3K21 is an eQTL only in brain tissues. This suggests that factors specific to the brain are required for this particular *Alu* MEV to exert its influence on gene expression. Including singletons, 628 MEVs in the 1000GP datasets were observed in 3'UTRs of protein-coding genes. Although only 71 were used for eQTL mapping due to low allele frequency in GTEx, which

is biased towards European ancestry, these also have the potential to influence gene expression. An East Asian-specific *Alu* insertion in 3'UTR of the pleiotropic gene *EGFR* decreases the expression of the reporter gene (Fig. 5g). Further assessment of the phenotypic consequences of this MEV is warranted; among the 42 diseases tested so far (see below), this variant is modestly associated with asthma (Supplementary Fig. 41; $P = 0.00018$, OR = 1.44).

The *Alu* sequence may recruit factors such as RNA-binding proteins or nucleases that stabilize or destabilize the RNA within which it is transcribed. If so, the expression levels of these factors may correlate with the effect of *Alu* on steady-state RNA. In other words, the effect of

Alu may be dependent on the expression of other genes, and such genes can be considered as proxies of the *Alu*-eQTL effect (proxy genes). To detect such potential factors, we generated an across-tissue regression model with an interaction term relating *Alu* genotype with proxy gene expression and checked for proxy genes for the 20 *Alu*-eQTLs. The most often-detected proxy gene was *FAM120A*, which was inferred to be associated with the effect of 11 *Alu* variants (Fig. 5h). The previously reported *Alu*-binding protein, HNRNPK²⁸, was also detected as a proxy of 4 *Alu* variants. Factors related to RNA degradation, such as *CNOT7* and *EDC3*, and trafficking, such as *XPO7*, were also detected as proxies of more than 6 *Alu* variants. Proxy genes, which can be considered as candidate RNA-binding factors/complexes involved in 3'UTR *Alu*-mediated gene regulation, are enriched for RNA-related processes, such as mRNA processing and RNA splicing (Fig. 5i). To validate this approach, we tested the effect of *FAM120A* overexpression on the regulatory influence of a 3' UTR *Alu* polymorphism (*Alu*-ADIPOQ) for which it was detected as a proxy. *Alu*-dependent downregulation of reporter gene expression was augmented by the overexpression of *FAM120A* in a dose-dependent manner (Fig. 5j), consistent with the effect of *Alu*-ADIPOQ being altered by *FAM120A*. Together, these results show MEVs' propensity to influence gene expression via shared patterns and mechanisms based on context and ME family²⁹.

Trait association and GWAS including MEVs

As MEVs cause gene expression differences (see above), they may also underlie trait associations. We surveyed the LD between MEVs and trait-associated variants identified by GWAS in BBJ and UK Biobank (Pan-UKB). Out of 4,369 lead variants in 172 GWAS in BBJ, 54 lead variants were in high LD with ME polymorphisms (Supplementary Fig. 42a, $r^2 > 0.8$). In Pan-UKB, 833 out of 169,822 lead variants in 7,221 GWASs tagged MEVs; 147 of these lead variants associated with clinically relevant measurements (Supplementary Fig. 42b and Supplementary Table 13). MEVs tag a similar number of GWAS Catalog variants as harmonized SNVs (Supplementary Fig. 43).

To demonstrate that MEVs genotyped by MEGANe can be integrated in GWAS to pinpoint putative genetic causes of disease risk, we performed GWAS including MEVs. MEV, SNV, and indel genotypes were imputed using an imputation reference panel based on 1000GP haplotypes, and all imputed variants were associated with 42 diseases studied in BBJ. We identified 54 MEVs associated with traits with P below the genome-wide significance threshold. After serial conditioning on lead variants, five MEVs associated with three diseases (Fig. 6a–e, Supplementary Fig. 45 and Supplementary Table 15); one is detected as a lead variant and four tagged lead variants. Absence of a reference L1 insertion 11-kb upstream of the transcription start site of *EVI2A* (L1-EVI2A, AF = 0.42 in 1000GP) is detected as a new lead variant in GWAS of type 2 diabetes (T2D), replacing the SNV that previously served as the sentinel of this haplotype (Fig. 6b). Whereas this locus has previously been linked to *NFI* as the likely candidate gene³⁰, the

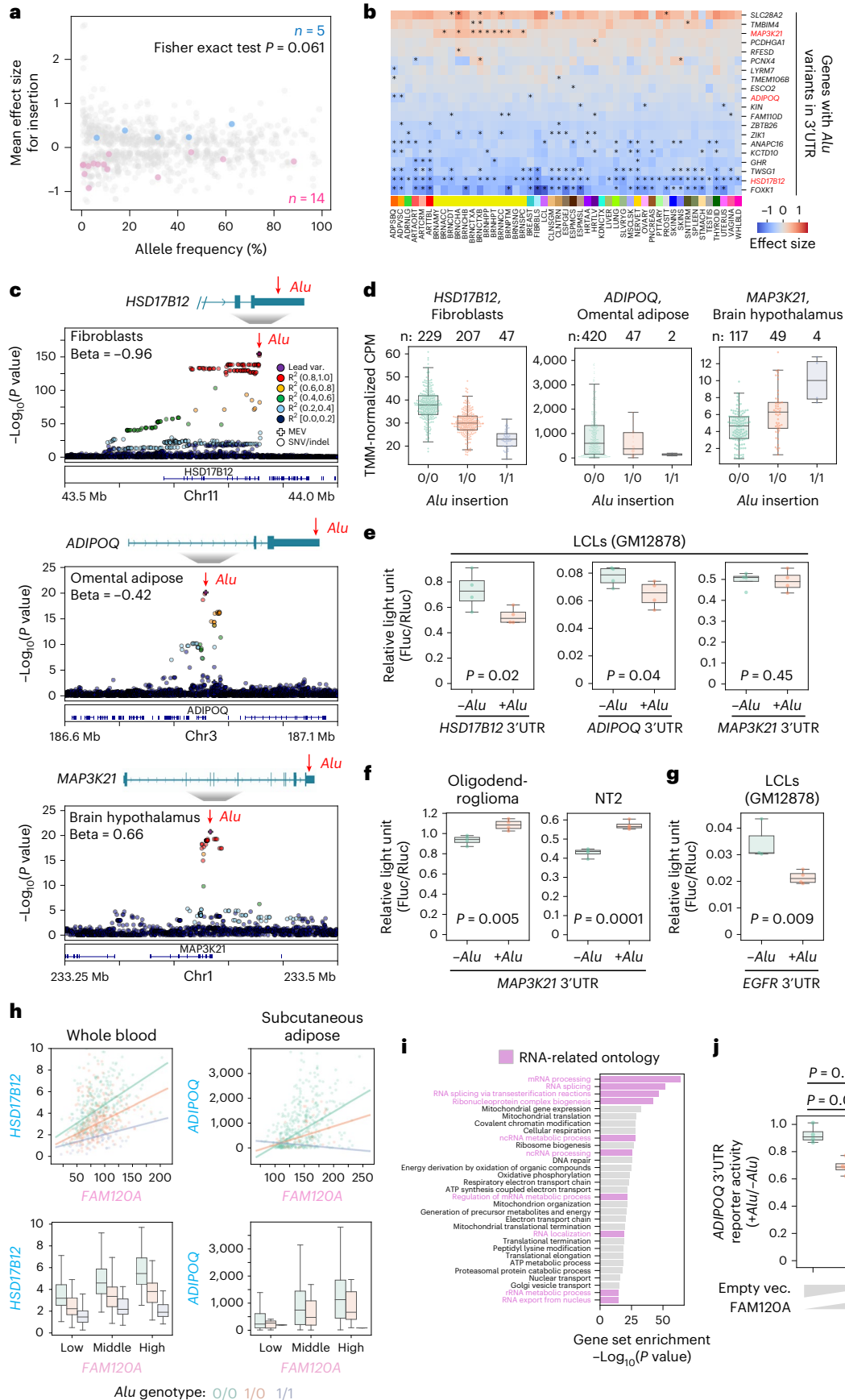
L1-EVI2A is also the lead variant of an eQTL of *EVI2A* (encoded from an *NFI* intron) in omental adipose tissue (Fig. 6b). L1-EVI2A also tags a lead SNV rs12943365 in sex hormone-binding globulin (SHBG) protein GWAS in Pan-UKB ($r^2 = 0.86$) associated with decreased SHBG, which often inversely correlates with BMI³¹. Also in T2D GWAS, an *Alu* insertion tagged a lead variant ($r^2 = 0.94$) of a locus on chromosome 19 within a cluster of zinc finger proteins (Fig. 6c). This insertion is predominantly found in East Asians; the MAF in JPT and EAS is 2.4% and 4.7%, respectively, whereas the MAF in other populations is 0.15% or lower, suggesting that MEVs can underlie population-specific risk haplotypes.

An L1 insertion in an intron of *NEDD4* (L1-NEDD4) associates with keloid, tagging a lead SNV rs16976600 ($r^2 = 0.85$, 1000GP EAS) (Fig. 6d,e) of a known risk locus³². L1-NEDD4 is also in high LD with variants associated with increased *NEDD4* expression in GTEx fibroblasts that colocalize with keloid GWAS (Supplementary Fig. 46a, coloc PP4 = 93%). *NEDD4* has two promoters, expressing long and short transcript variants (Fig. 6f). The short variant is highly expressed in keloid scars and reportedly activates inflammatory pathways³³. To test whether L1-NEDD4 associates with increased expression of this shorter transcript, we performed exon-eQTL analysis. The expression of exon 9, which is specific to the short variant, is strongly associated with the presence of L1-NEDD4, whereas exon 1, the long variant-specific exon, is not (Fig. 6g,h). Because L1 often functions as an enhancer, we hypothesized that L1-NEDD4 may enhance expression of the short variant and impact keloid through the activity of this transcript variant on inflammation. Notably, L1-NEDD4 tags lead variants of Dupuytren's disease and fasciitis GWAS in Pan-UKB (rs8032158 and rs59912282, $r^2 = 0.93$ and 0.85, respectively), suggesting a shared genetic mechanism in several diseases featuring fibroblast inflammation.

To test the influence of this L1 polymorphism directly, we knocked out L1-NEDD4 in iPSCs derived from a healthy Japanese individual carrying two copies of L1-NEDD4 (Fig. 6i). We obtained 9 knockout (KO) and 11 wild-type (WT) clones and differentiated them into fibroblasts. In cells with biallelic knockout of L1-NEDD4, the expression of *NEDD4* decreased (Fig. 6j). Although expression of both variants decreased in KO clones, the effect on the short variant was more pronounced; the ratio of the short variant to the long variant decreased in KO clones. This demonstrates that the L1 insertion functions as an enhancer of *NEDD4*, particularly for the short variant previously implicated in keloid pathogenesis. Because the short variant of *NEDD4* is involved in inflammation³³, L1-NEDD4 genotype may explain heterogeneity in the clinical presentation of keloid. Indeed, L1-NEDD4 increases the odds of developing keloid due to acne, but not after surgery, among BBJ participants (Fig. 6k, Supplementary Fig. 47 and Supplementary Table 17). L1-NEDD4 also increases the odds of clinical indicators of keloid severity, including contracture and spontaneous pain, as well as history of keloid treatment by radiation or surgery. Thus, the molecular pathways activating, and activated by, L1-NEDD4 are rational targets for developing genotype-guided drugs against severe keloid.

Fig. 5 | *Alu* insertions in 3'UTRs. **a**, Distribution of allele frequencies and effect sizes of ME-eQTLs. Nineteen *Alu* insertions in 3'UTRs detected as ME-eQTLs of the genes are highlighted with blue and red dots. One MEV associated with increased or decreased gene expression depending on tissues was excluded. Effect sizes for presence of ME insertion are shown. **b**, Heatmap showing the effects of *Alu* insertions in 3'UTR. Significant associations (local false sign rate < 0.05) are flagged. Color bar corresponds to tissue. **c**, *HSD17B12* eQTL regional association plot in fibroblasts (top), *ADIPOQ* in omental adipose (middle), and *MAP3K21* in hypothalamus. MEVs and SNPs are shown as plus marks and circles, respectively. *Alu* insertions are highlighted with arrows. P calculated by linear regression test. **d**, *HSD17B12* expression in fibroblasts (left), *ADIPOQ* in omental adipose (middle), and *MAP3K21* in hypothalamus (right). Numbers of data points are shown in figures. **e**, Reporter assays of the *HSD17B12* (left), *ADIPOQ* (middle), and *MAP3K21* (right) 3'UTRs with or without *Alu* insertion (GM12878 cells). **f**, Reporter assays of the *MAP3K21* 3'UTR with or without *Alu* insertion. Plasmids

were transfected into Oligodendrogloma (left) and NT2/D1 cells (right). **e**, **f**, $n = 4$ independent experiments. **g**, Reporter assays of the *EGFR* 3'UTR with or without *Alu* insertion (GM12878 cells). $n = 3$ (-*Alu*), 4 (+*Alu*) independent experiments. **h**, The distributions of expression of eGenes, *HSD17B12* (left) and *ADIPOQ* (right), compared to that of a proxy gene, *FAM120A*. Colored lines display linear regression of the data grouped by *Alu* genotype. Bottom: individuals divided into tertiles based on the *FAM120A* expression. $n = 311$ (0/0), 285 (1/0), 74 (1/1). **i**, Gene-set enrichment analysis for proxy genes. P values were calculated by permutation test. **j**, Ratio of reporter activity of the *ADIPOQ* 3'UTR with or without *Alu*, titrating *FAM120A*-flag (GM12878 cells). $n = 4$ independent experiments. **e–g**, **j**, P of two-sided t -test is shown. **d–h**, **j**, Middle line of box plot represents median, and lower and upper whiskers represent the lowest data point above $Q1 - 1.5 \times IQR$ and highest below $Q3 + 1.5 \times IQR$, respectively, where $Q1$, $Q3$, IQR are the first and third quartiles and interquartile range, respectively.



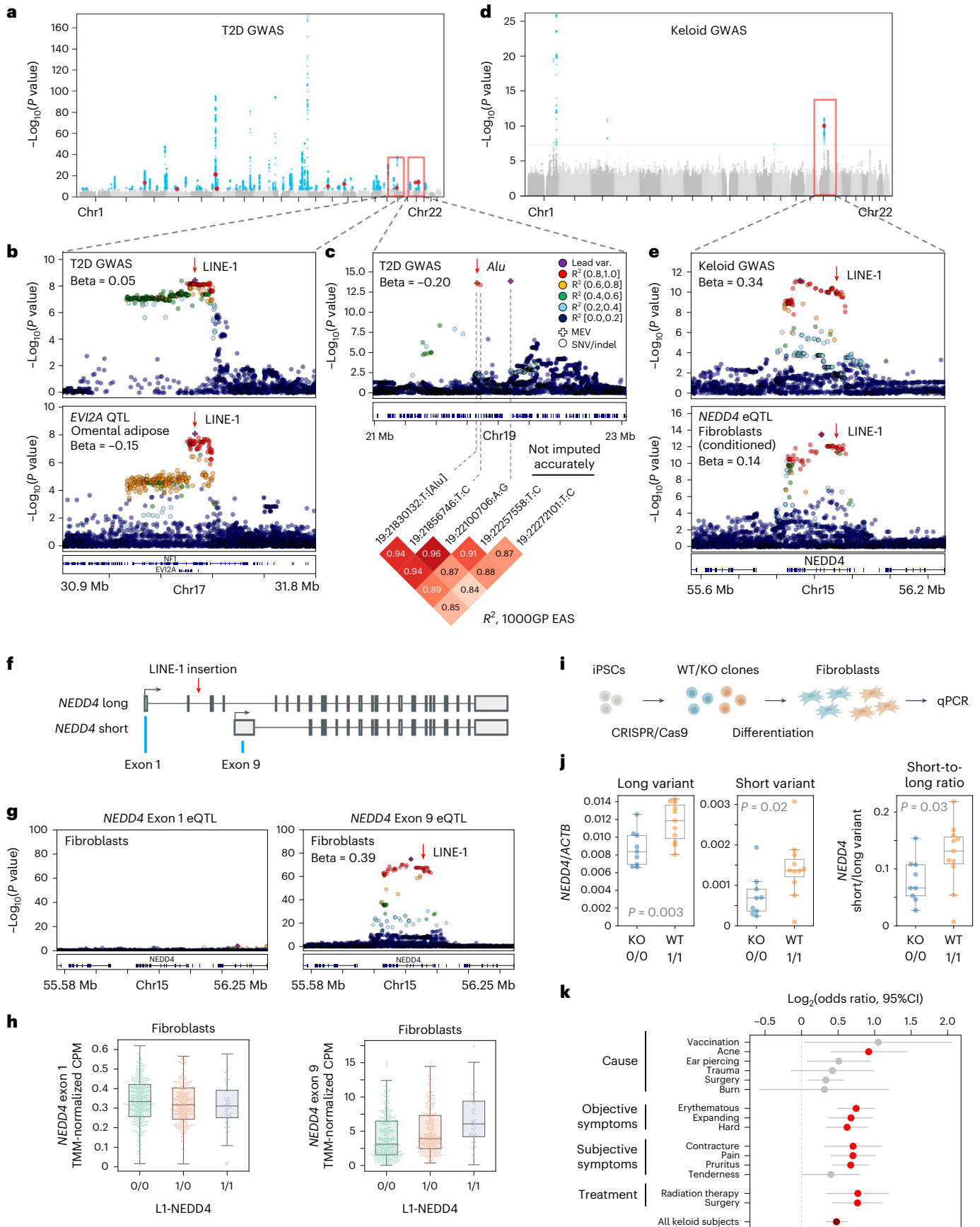


Fig. 6 | MEVs associate with disease. **a**, Manhattan plot of type 2 diabetes (T2D) GWAS in Japanese. **b**, Regional association plots showing haplotypes associated with T2D (top) and *EV12A* expression in omental adipose tissue (bottom). **c**, Regional association plots showing haplotypes associated with T2D. The bottom heatmap shows LD between variants. Two variants, 19:22257558:C:T and 19:22272101:C:T, are not shown because these were not accurately imputed ($\text{INFO} < 0.7$). **d**, Manhattan plot of keloid GWAS in Japanese. **a, d**, Red and blue dots represent MEVs and SNPs, respectively, with logistic mixed regression test $P < 5 \times 10^{-8}$. **e**, Regional association plots showing LDs associated with keloid (top) and *NEDD4* expression in fibroblasts (bottom). **f**, Illustration of the long and short *NEDD4* transcript variants. Location of L1-*NEDD4* is depicted with a red arrow. **g**, Regional association plots showing no association between variants and exon 1 (left) and 9 (right) expressions in fibroblasts. **b, c, e, g**, MEVs and SNPs are shown as plus marks and circles, respectively. MEVs are highlighted with red arrows. *P* calculated by linear regression test. **h**, Expression of *NEDD4* exon 1 (left) and

9 (right) in fibroblasts. $n = 235$ (0/0), 193 (1/0), 55 (1/1). **i**, Experimental design for L1-*NEDD4* knockout (KO). **j**, Expression levels of the *NEDD4* long variant (left), short variant (middle), and short-to-long variant ratio (right). *NEDD4* gene expression was normalized by the expression of *ACTB*. *P* of one-sided *t*-test is shown. **h, j**, Middle line of box plot represents median, and lower and upper whiskers represent the lowest data point above $Q1 - 1.5 \times \text{IQR}$ and highest below $Q3 + 1.5 \times \text{IQR}$, respectively, where $Q1$, $Q3$ and IQR are the first and third quartiles and interquartile range, respectively. **k**, Odds ratios that patients carry L1-*NEDD4* based on disease characteristics, including cause of keloid development, signs and symptoms, and treatment history. Red and dark red points show odds ratios significantly above 1 (two-sided Fisher exact test $P < 0.05$ after Bonferroni correction accounting for additional tests as shown in Supplementary Fig. 46). Odds ratios and these 95% confidence intervals are shown. Numbers of individuals are summarized in Supplementary Table 16.

Discussion

Here, we interrogated the consequences of recent ME activity on human genomes and phenotypes. Accurate detection of MEVs in diverse human populations allowed us to resolve population-specific patterns of recent genome diversification accounted for by ME insertions. These may reflect different active ME copies³⁴ or differences in the repertoire of factors repressing MEs. Although *Alu* insertions tend to be observed in late-replicating domains, this trend was mitigated in East Asians and even reversed in Japanese. This finding suggests that the insertion preference of *Alu* has shifted as humans have populated the earth. Previous work suggested a similar drift in insertion preference occurred during primate radiation; older, nonpolymorphic *Alu* are known to be enriched in early-replicating domains, whereas recent polymorphic ones show the opposite trend³⁵. The factors besides ORF2p that regulate the insertion preferences of human MEs are unknown; changes to the spatiotemporal regulation of transposition-competent ribonucleoproteins could result from accumulation of population-specific mutations in these factors or in active MEs themselves.

Our ME-eQTL analyses shed light on the complex but coherent regulatory logic encoded by MEVs. Although 3'UTR *Alu* are often detected as multi-tissue eQTLs, some are clearly tissue-specific, such as *Alu*-MAP3K21 specific to the brain. Context (for example surrounding sequence and co-expressed genes) is decisive in licensing *Alu* polymorphisms to exert post-transcriptional regulation. Consistent with this concept, we identified FAM120A as a co-regulator of 3'UTR *Alu*. Disruption of interactions like that of FAM120A could represent a new target for multipurpose precision medicines. The 3'UTR *Alu* MEV in *HSD17B12* causes changes in reporter gene expression and associates with a number of biometric traits and basal metabolic rate (highlighted in Supplementary Table 13); this variant can thus be considered to causally influence human weight, and blocking this *Alu*'s regulatory effect can be predicted to be tolerated. Similarly, a 3'UTR *Alu* in the SARS-CoV-2 host factor and dementia-linked gene *TMEM106B*^{36,37}, detected as an ME-eQTL in several tissues, is associated with a number of mental health phenotypes (highlighted in Supplementary Table 13). It will be of great interest to define additional class-specific regulatory effects of MEVs, as these will advance the interpretability of non-coding genomic variation.

Inclusion of MEVs in GWAS bridges the gap between known risk loci and underlying genetic causes, demonstrating a new path to overcome the challenge of connecting GWAS signals in non-coding regions to causal variants, especially in non-European populations. By accurately genotyping MEVs and determining their linkage with SNVs, we identified hundreds of MEVs present on known risk haplotypes. These include an L1 insertion we show is causal for altered gene expression and potentially mediates the increased keloid risk associated with this haplotype. The mechanism demonstrated in the case of L1-*NEDD4*—that an intronic L1 insertion observed as an ME-eQTL enhances gene/isoform expression to potentially drive pathogenesis—represents

an attractive hypothesis for a class of ME-trait associations we document. For example, an L1 insertion in an intron of the gene encoding thyroid stimulating hormone receptor (TSHR) and also detected as a *TSHR* ME-eQTL is associated with Graves' disease (characterized by TSHR-reactive autoantibodies), and an L1 insertion intronic to and associated with *ULK4* expression is associated with diastolic blood pressure and pulse pressure, among other examples (highlighted in Supplementary Table 13). Extending these analyses using more WGS data will allow the integration of more, and rarer, MEVs in GWAS of additional phenotypes, leading to the discovery of additional disease-causing MEs and motivating development of ME-targeting drugs. The observation that a human-specific ME insertion substantially predisposes to keloid, which has not been observed in other primates³⁸, also supports the utility of this approach to infer genetic origins of other traits characteristic of our species³⁹.

By improving detection and prioritization of a type of variants difficult to assess at genome-wide scale, our tool and results are applicable to medical genetics. Even so, a major limitation remains: confident prediction of which MEVs alter phenotype requires additional data integration and statistical testing. However, our results also demonstrate that ME ontology relates coherently to MEV effect. Here, we infer putative effects of several MEVs at the level of disease, providing important information for personalized medicine; MEVs impact many traits plausibly entangled with fitness in our varied landscapes, but we have not explicitly addressed beneficial variants or those with antagonistic pleiotropy. Still, our work provides comprehensive backing to the assertion that MEs are drivers of diversification of genome sequence and function, classic concepts of genome evolution. In addition, we highlight MEs as a source of biased mutation, invoked to account for neutral evolution of complexity⁴⁰. As the direction and pace of diversification can be modified by MEs, differences in ME-derived mutation patterns may potentiate differential genome plasticity between lineages.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-023-01390-2>.

References

- Wells, J. N. & Feschotte, C. A field guide to eukaryotic transposable elements. *Annu. Rev. Genet.* **54**, 539–561 (2020).
- Serrato-Capuchina, A. & Matute, D. R. The role of transposable elements in speciation. *Genes (Basel)* **9**, 254 (2018).
- Payer, L. M. & Burns, K. H. Transposable elements in human genetic disease. *Nat. Rev. Genet.* **20**, 760–772 (2019).

4. Kobayashi, K. et al. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature* **394**, 388–392 (1998).
5. Hancks, D. C. & Kazazian, H. H. J. Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016).
6. Kagawa, T. et al. Recessive inheritance of population-specific intronic LINE-1 insertion causes a rotor syndrome phenotype. *Hum. Mutat.* **36**, 327–332 (2015).
7. Goubert, C., Zavallos, N. A. & Feschotte, C. Contribution of unfixed transposable element insertions to human regulatory variation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **375**, 20190331 (2020).
8. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
9. Payer, L. M. et al. Alu insertion variants alter gene transcript levels. *Genome Res.* **31**, 2236–2248 (2021).
10. Cao, X. et al. Polymorphic mobile element insertions contribute to gene expression and alternative splicing in human tissues. *Genome Biol.* **21**, 185 (2020).
11. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
12. Payer, L. M. et al. Structural variants caused by Alu insertions are associated with risks for many human diseases. *Proc. Natl Acad. Sci. USA* **114**, E3984–E3992 (2017).
13. Jacques, P.-É., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet.* **9**, e1003504 (2013).
14. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
15. Trizzino, M. et al. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res.* **27**, 1623–1633 (2017).
16. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
17. Peter, E. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
18. Scott, A. J., Chiang, C. & Hall, I. M. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res.* **31**, 2249–2257 (2021).
19. Ito, J. et al. A hominoid-specific endogenous retrovirus may have rewired the gene regulatory network shared between primordial germ cells and naïve pluripotent cells. *PLoS Genet.* **18**, e1009846 (2022).
20. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
21. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
22. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
23. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
24. Jabbari, K. & Bernardi, G. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochores families. *Gene* **224**, 123–128 (1998).
25. Soumillon, M. et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3**, 2179–2190 (2013).
26. van Bree, E. J. et al. A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci. *Genome Res.* **32**, 656–670 (2022).
27. Vialle, R. A., de Paiva Lopes, K., Bennett, D. A., Crary, J. F. & Raj, T. Integrating whole-genome sequencing with multi-omic data reveals the impact of structural variants on gene regulation in the human brain. *Nat. Neurosci.* **25**, 504–514 (2022).
28. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107–111 (2018).
29. Fueyo, R., Judd, J., Feschotte, C. & Wysocka, J. Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.* **23**, 481–497 (2022).
30. Spracklen, C. N. et al. Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**, 240–245 (2020).
31. Hautanen, A. Synthesis and regulation of sex hormone-binding globulin in obesity. *Int. J. Obes.* **24**, S64–S70 (2000).
32. Ogawa, R. et al. Associations between keloid severity and single-nucleotide polymorphisms: importance of rs8032158 as a biomarker of keloid severity. *J. Invest. Dermatol.* **134**, 2041–2043 (2014).
33. Fujita, M. et al. NEDD4 is involved in inflammation development during keloid formation. *J. Invest. Dermatol.* **139**, 333–341 (2019).
34. Gardner, E. J. et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
35. Costantini, M., Auletta, F. & Bernardi, G. The distributions of ‘new’ and ‘old’ Alu sequences in the human genome: the solution of a ‘mystery’. *Mol. Biol. Evol.* **29**, 421–427 (2012).
36. Baggen, J. et al. Genome-wide CRISPR screening identifies TMEM106B as a proviral host factor for SARS-CoV-2. *Nat. Genet.* **53**, 435–444 (2021).
37. Van Deerlin, V. M. et al. Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat. Genet.* **42**, 234–239 (2010).
38. Marttala, J., Andrews, J. P., Rosenbloom, J. & Uitto, J. Keloids: Animal models and pathologic equivalents to study tissue fibrosis. *Matrix Biol.* **51**, 47–54 (2016).
39. Linker, S. B., Marchetto, M. C., Narvaiza, I., Denli, A. M. & Gage, F. H. Examining non-LTR retrotransposons in the context of the evolving primate brain. *BMC Biol.* **15**, 68 (2017).
40. Yampolsky, L. Y. & Stoltzfus, A. Bias in the introduction of variation as an orienting factor in evolution. *Evol. Dev.* **3**, 73–83 (2001).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

¹Genome Immunobiology RIKEN Hakubi Research Team, RIKEN Center for Integrative Medical Sciences and RIKEN Cluster for Pioneering Research, Yokohama, Japan. ²Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ³Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. ⁴Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA. ⁵Graduate School of Medical Life Science, Yokohama City University, Yokohama, Japan. ⁶Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ⁷Laboratory for Pharmacogenomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ⁸Graduate School of Advanced Science and Engineering, Waseda University, Tokyo, Japan. ⁹Computational Bio Big-Data Open

Innovation Laboratory (CBBDOIL), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan. ¹⁰Cell Engineering Division, BioResource Research Center, RIKEN, Tsukuba, Japan. ¹¹Laboratory of Complex Trait Genomics, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. ¹²Laboratory for Statistical and Translational Genetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ¹³RIKEN-IFOM Joint Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ¹⁴Institute for the Advanced Study of Human Biology, Kyoto University, Kyoto, Japan. ¹⁵IFOM ETS - the AIRC Institute of Molecular Oncology, Milan, Italy. ¹⁶Laboratory for Human Immunogenetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ¹⁷Clinical Research Center, Shizuoka General Hospital, Shizuoka, Japan. ¹⁸The Department of Applied Genetics, The School of Pharmaceutical Sciences, University of Shizuoka, Shizuoka, Japan. ⁴¹Present address: Next-Generation Precision Medicine Development, Integrative Genomics Laboratory, Graduate School of Medicine, Department of Medical Science, The University of Tokyo, Tokyo, Japan. ⁴²Present address: Paleovirology Lab, Department of Biology, University of Oxford, Oxford, UK. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: shohei.kojima@riken.jp; nicholas.parrish@riken.jp

Biobank Japan Project Consortium

Koichi Matsuda^{19,20}, Yuji Yamanashi²¹, Yoichi Furukawa²², Takayuki Morisaki²³, Yoshinori Murakami²⁴, Kaori Muto²⁵, Akiko Nagai²⁵, Wataru Obara²⁶, Ken Yamaji²⁷, Kazuhisa Takahashi²⁸, Satoshi Asai^{29,30}, Yasuo Takahashi³⁰, Takao Suzuki³¹, Nobuaki Sinozaki³¹, Hiroki Yamaguchi³², Shiro Minami³³, Shigeo Murayama³⁴, Kozo Yoshimori³⁵, Satoshi Nagayama³⁶, Daisuke Obata³⁷, Masahiko Higashiyama³⁸, Akihide Masumoto³⁹ & Yukihiko Koretsune⁴⁰

¹⁹Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ²⁰Laboratory of Clinical Genome Sequencing, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. ²¹Division of Genetics, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ²²Division of Clinical Genome Research, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ²³Division of Molecular Pathology, IMSUT Hospital, Department of Internal Medicine, Project Division of Genomic Medicine and Disease Prevention, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ²⁴Department of Cancer Biology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ²⁵Department of Public Policy, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ²⁶Department of Urology, Iwate Medical University, Iwate, Japan. ²⁷Department of Internal Medicine and Rheumatology, Juntendo University Graduate School of Medicine, Tokyo, Japan. ²⁸Department of Respiratory Medicine, Juntendo University Graduate School of Medicine, Tokyo, Japan. ²⁹Division of Pharmacology, Department of Biomedical Science, Nihon University School of Medicine, Tokyo, Japan. ³⁰Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, Japan. ³¹Tokushukai Group, Tokyo, Japan. ³²Department of Hematology, Nippon Medical School, Tokyo, Japan. ³³Department of Bioregulation, Nippon Medical School, Kawasaki, Japan. ³⁴Tokyo Metropolitan Geriatric Hospital and Institute of Gerontology, Tokyo, Japan. ³⁵Fukujuji Hospital, Japan Anti-Tuberculosis Association, Tokyo, Japan. ³⁶The Cancer Institute Hospital of the Japanese Foundation for Cancer Research, Tokyo, Japan. ³⁷Center for Clinical Research and Advanced Medicine, Shiga University of Medical Science, Shiga, Japan. ³⁸Department of General Thoracic Surgery, Osaka International Cancer Institute, Osaka, Japan. ³⁹Izuka Hospital, Fukuoka, Japan. ⁴⁰National Hospital Organization Osaka National Hospital, Osaka, Japan.

Methods

Overview of the algorithm of MEGAnE

MEGAnE finds ME insertions and absences and genotypes the discovered MEVs. It searches for discordantly mapped reads and finds potential breakpoints from clipped reads. It uses BLASTn to search for similarity between the overhangs of clipped reads and ME insertions. It makes breakpoint pairs that represent the upstream and downstream breakpoints of an ME insertion or absence, or, in most cases, the start and end positions of a target site duplication (TSD). It then extracts breakpoints that are highly likely to derive from ME insertions or absences and fits a Gaussian mixture model, which models homozygosity and heterozygosity of the input sample. Based on the modeled distribution, MEGAnE removes likely false positives. After discovering ME insertions and absences, it genotypes the polymorphic MEs based on the number of reads providing evidence of each breakpoint, evidence of breakpoint absence and read depth of the TSD. It outputs discovered ME insertions and absences in VCF format (Supplementary Fig. 1).

After MEV discovery and genotyping of multiple samples, MEGAnE can merge them to make a joint callset. It first merges the breakpoint positions in multiple VCF files, then searches for reads providing evidence of the merged breakpoints. If sufficient reads support a breakpoint, discrete genotypes (that is '0/1' or '1/1') are assigned. If there are no reads supporting a breakpoint, it assigns genotypes as '0/0'. If there is weak evidence of the breakpoint, it leaves the genotype as missing, that is './0'.

MEV discovery from 1000GP GRCh38 datasets

The 30× WGS data from 3,202 individuals mapping to GRCh38DH were downloaded from the 1000GP website (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/). Throughout this paper, we refer to this dataset as '1000GP GRCh38 datasets'. MEs were discovered and genotyped using MEGAnE's call_genotype_38 command. The joint callset was generated using MEGAnE's joint_calling_hs command. We also generated a joint callset from 2,503 individuals, which does not include relatives, using from the same dataset. We generated a separate joint callset for 34 individuals who were sequenced using PacBio in the 1000GP HGVC project. The HGVC sequenced 35 individuals by PacBio; however, we excluded one individual, HG002, from our joint callset, because the individual was not included in the 3,202 individuals who were sequenced in the 1000GP 30× WGS. In 2,503 individuals analyzed here, MEGAnE detected 48,248 MEVs with the filter 'PASS' flag. Of those, 8,609 (18% of total) were common variants (AF > 1%).

MEV discovery from 1000GP GRCh37 datasets

The raw fastq reads of the 2,504 individuals in the 1000GP 30× GRCh38 datasets were downloaded from the 1000GP website (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/). The fastq reads were mapped on the human reference genome build, human_g1k_v37 by BWA MEM using the same options as used by 1000GP to map on GRCh38DH (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/20190405_NYGC_b38_pipeline_description.pdf). In brief, we used the -Y option with the -K 100000000 option. Throughout this paper, we refer to this dataset as '1000GP GRCh37 datasets'. The output alignment was converted to CRAM format and analyzed using MEGAnE's call_genotype_37 command. The joint callset was generated using MEGAnE's joint_calling_hs command. In 2,504 individuals analyzed here, MEGAnE detected 48,360 MEVs with the filter 'PASS' flag. Of those, 8,665 (18% of total) were common variants (AF > 1%) (Fig. 1e).

MEV discovery from 25× WGS datasets in BBJ

We applied MEGAnE to the 25× WGS (either 160 or 150 bp paired-end) from 1,235 individuals in BBJ^{41,42}. We mapped the raw fastq reads to the

human reference genome hs37d5 by BWA MEM using the same option as we used for mapping of 1000GP dataset and saved as CRAM format. We did not perform further individual-level QC, because the dataset was already subjected to QC. The output CRAM files were analyzed by the MEGAnE's call_genotype_37 command. The joint callset was generated by the MEGAnE's joint_calling_hs command. In 1,235 Japanese individuals analyzed here, it detected 10,996 MEVs with the filter 'PASS' flag. Of those, 4,943 (45% of total) were common variants (AF > 1%). This callset was used for evaluating LD between MEVs and SNVs.

MEV discovery from 25× and 15× WGS datasets in BBJ

To find rare insertions in Japanese individuals, we generated a joint callset by merging data from as many Japanese individuals as possible. To this end, we analyzed additional 30× WGS (either 125 or 124-nt paired-end) from 256 individuals and 15× WGS (150 bp paired-end) from 3,389 individuals by MEGAnE and merged with the MEVs detected from 1,235 individuals described above. When analyzing 15× WGS, we used the 'lowdep' option of MEGAnE, which assumes non-Gaussian distributions of supporting read counts in heterozygous and homozygous insertions. In total, we merged MEVs from 4,480 Japanese individuals using the joint_calling_hs command. In 4,880 Japanese individuals analyzed here, MEGAnE detected 24,933 MEVs with the filter 'PASS' flag. Of those, 5,452 (22% of total) were common variants (AF > 1%) (Fig. 1e). This joint callset was used to investigate ME insertion preferences in Japanese.

Haplotype estimation for MEGAnE callset 1000GP GRCh38

First, we merged the MEI and ME absence callsets from MEGAnE. We used MEGAnE's reshape_vcf command to merge these two callsets and remove multi-allelic ME variants. To estimate haplotypes of 2,503 individuals in 1000GP phase3, we merged the ME callset with SNVs. For quality control (QC), we first split the ME callset into individuals belonging to each of five superpopulations and evaluated Hardy-Weinberg equilibrium. Variants that violated Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$) in at least one superpopulation were removed. SNVs that overlap with polymorphic MEs were removed. Singleton SNVs and MEs were also removed. Then, the QC-ed ME callset was merged with the SNV callset (1000GP, GRCh38_v1a) without variants violating Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$). Each chromosome of the merged callset was saved in VCF format and phased using SHAPEIT4 software with default genetic maps. The phased haplotypes were converted to an imputation reference panel using Minimac3 software. Due to the unavailability of SNVs on sex chromosomes, we estimated the haplotypes for MEs only on autosomes and PARs.

Haplotype estimation for MEGAnE callset 1000GP GRCh37

First, we merged the MEI and ME absence callsets from MEGAnE using the same MEGAnE command described in the previous section. To estimate haplotypes of 2,504 individuals in 1000GP phase3, we merged the ME callset with SNVs and indels. We first split the ME callset into individuals belonging to each of five superpopulations and evaluated Hardy-Weinberg equilibrium. Variants that violated Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$) in at least one superpopulation were removed. SNVs and indels that overlap with polymorphic MEs were removed. Singleton SNVs, indels, and MEs were also removed. Then, the QC-ed ME callset was merged with the SNV and indel callset (1000GP, v5a) without variants violating Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$). Each chromosome of the merged callset was saved in VCF format and phased by SHAPEIT4 software with default genetic maps. An imputation reference panel was made using Minimac3 software. Due to the unavailability of SNVs on the Y chromosome, we estimated haplotypes for MEs only on autosomes and the X chromosome.

Genotype imputation for GTEx individuals

To impute ME genotypes in 838 individuals recruited in the GTEx v8, we used the 5,006 haplotypes in 1000GP. We used the phased

SNVs and indels provided from GTEx (GTEx_Analysis_2017-06-05_v8_WholeGenomeSeq_838Indiv_Analysis_Freeze.SHAPEIT2_phased.vcf.gz) as target haplotypes. Variants violating Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$) were removed before imputation. ME genotypes on autosomes and PARs were imputed using Minimac3 software with the imputation reference panel generated from the 1000GP GRCh38 callset. After imputation, ME genotypes were extracted and merged with the original SNV and indel calls. MEs violating Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$) and/or having Minimac R^2 lower than 0.5 were removed. Variants with allele frequency lower than 0.5% were removed, leaving 9,836 MEVs for use in eQTL analysis.

Genotype imputation in BBJ

To impute ME genotypes of participants in BBJ, we used the 5,008 haplotypes in the 1000GP GRCh37 dataset. We used phased SNVs genotyped by SNV array as target haplotypes. ME genotypes on autosomes were imputed using Minimac3 software with the imputation reference panel generated from the 1000GP GRCh37 callset. After imputation, variants violating Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$) and those with Minimac R^2 lower than 0.7 were removed. All variants with minor allele count lower than 10 were removed, and the remaining variants were used for GWAS.

PC analysis of MEVs

The PCs of ME polymorphisms called from 1000GP GRCh37 datasets and the SFARI cohort were calculated by Plink2 software. We first removed MEVs violating Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$), those with minor allele frequency lower than 1%, and those in regions of long-range high LD ([https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_\(LD\)](https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD))). The variants were then pruned by Plink2 software with '-indep-pairwise 500 5 0.2' option. The top 10 PCs were calculated using the plink2-pca command.

Intersections between MEVs and gene annotations

To compile MEVs that intersect with exons, CDS, and promoters, we first reshaped gene annotation files downloaded from GenCode using a script provided in the GTEx pipelines (https://github.com/broadinstitute/gtex-pipeline/blob/master/gene_model/collapse_annotation.py). We defined the 1-kb regions upstream from transcription start sites as promoters. All gene annotations in the GTF file were used for this analysis. To see intersection with MEVs called from 1000GP, we used 48,241 MEVs with the filter 'PASS' flag called from 1000GP GRCh38 datasets. For this analysis, we used a GenCode GTF version 26. To see intersection with MEVs called from BBJ, we used 10,997 MEVs with the filter 'PASS' flag called from 1,235 individuals sequenced at 25× depth WGS. For this analysis, we used a GenCode GTF version 26lift37.

Correlations between ME insertions and genomic features

To evaluate the characteristics of genome features found to have insertions of MEs, the correlation between the number of ME insertions and genomic features was calculated. We calculated the genomic features for nonoverlapping 100-kb windows (see the 'Preparation of genomic features' section). Because L1 and SVA insertions are sparse, we first resized the window size to 1 Mb and 5 Mb, respectively. To this end, the average values were calculated for each nonoverlapping window. Then, 1-Mb and 5-Mb windows that contain one or more 100-kb window(s) with missing value and ones with at least one 'N' character in the human genome assembly, GRCh38DH, were excluded from the analysis. The Spearman correlation coefficients were calculated using the SciPy module in Python.

eQTL analysis in 49 tissues

We performed eQTL mapping using MEVs. We followed the eQTL mapping method used in GTEx v8. As for GTEx v8, we excluded 5 tissues out of the 54 tissues (Bladder, Cervix_Ectocervix, Cervix_Endocervix,

Fallopian_Tube, and Kidney_Medulla) from analysis due to the few available RNA-sequencing samples. First, expression profiles of the 49 tissues were prepared. The count per million matrices provided from GTEx (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz) were normalized across samples by TMM normalization using the script provided from GTEx (https://github.com/broadinstitute/gtex-pipeline/blob/master/qlt/src/eqlt_prepare_expression.py), then the genes that are expressed (≥ 0.1 TPM in $\geq 20\%$ samples and ≥ 6 reads in $\geq 20\%$ samples) were retained for eQTL mapping (38,471 genes in total of 49 tissues). Each gene was then inverse-normal transformed across samples. Next, we performed eQTL mapping by fastQTL software⁴³ with the same analysis options as for the previous eQTL mapping (<https://github.com/broadinstitute/gtex-pipeline/tree/master/qlt>). We also used the same covariates as those used for QTL mapping in GTEx: 5 genetic PCs, PEER factors, library preparation methods, sequencing platforms and sex. Genetic variants within 1 Mb from a gene were tested for associations. The 9,836 and 13,498,030 quality controlled ME and non-ME (that is SNVs and indels) variants, respectively, were used for eQTL mapping.

Across-tissue meta-analysis

After the eQTL mapping in each tissue, we performed across-tissue meta-analysis using the same method as performed in GTEx v8. First, we formatted the fastQTL results for MASH software⁴⁴. Then, the MASH model was trained by the same protocol as GTEx v8 performed (https://github.com/stephenslab/gtexresults/blob/master/workflows/fastqtl_to_mash.ipynb). The trained model was applied to ME-eGene pairs.

Detection of ME-eQTL

We defined ME-eQTLs as those which satisfy these criteria: (1) in the fastQTL output, an MEV is either the lead variant or has $r^2 > 0.95$ to the lead variant in at least one tissue, and (2) in the result of across-tissue meta-analysis, the MEV has local false sign rate < 0.05 in at least one tissue (Supplementary Table 9).

ME-GWAS of 42 diseases in BBJ

GWAS for 42 diseases were done using 179,660 individuals in BBJ using methods similar to those used in Ishigaki et al.⁴⁵. The MEV genotypes in 179,660 individuals were imputed by Minimac3 software using the imputation reference panel generated from the 1000GP GRCh37 datasets. After imputation, variants violating Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$), those with Minimac R^2 lower than 0.7, and those with a minor allele count lower than 10 were removed. The associations were calculated using a generalized linear mixed model implemented in SAIGE (version 0.44.5)⁴⁶ with the leave-one-chromosome-out approach. We used age, sex and the first five genetic PCs as covariates. For each disease, we defined a significantly associated locus as a genomic region within 3 Mb from the lead variants. Based on the methodology used in Ishigaki et al.⁴⁵, we used 9.58×10^9 as a genome-wide significance threshold and 5×10^8 as a threshold of suggestive association.

Knockout of L1-NEDD4 in iPSCs

We designed two sgRNAs cleaving upstream and downstream of L1-NEDD4 insertion. To reconstruct the allele without the L1-NEDD4, we amplified the L1-flanking regions (703 bp upstream and 787 bp downstream) and connected them at the TSD using overlap-extension PCR. The connected fragment was used as a template for homology-directed repair. The sgRNA-Cas9 complex and homology-directed repair template DNA were transfected to iPSCs derived from a healthy Japanese individual (60 s, male) found to carry two copies of L1-NEDD4 by electroporation using the NEON transfection system. After electroporation, cells were cultured for 2 weeks, and single cell-derived clones were obtained by limiting dilution. Deletion of L1-NEDD4 was checked by the same primers as used for PCR validation in 70 Japanese (Supplementary Fig. 44b).

Differentiation of iPSCs into fibroblasts

iPSC clones were first differentiated to mesenchymal stem cells (MSCs) using STEMdiff Mesenchymal Progenitor Kit according to the manufacturer's protocol. iPSC-derived MSCs were then differentiated to fibroblasts based on the protocol published in Lee *et al.*⁴⁷. MSCs were cultured in DMEM containing 100 ng ml⁻¹ CTGF, 50 ng/ml ascorbic acid, 1× penicillin/streptomycin, and 10% FBS for at least 3 weeks. Fully differentiated fibroblasts were maintained in the same medium used for MSC to fibroblast differentiation.

qRT-PCR of *NEDD4* transcripts

To measure the expression levels of *NEDD4* in fibroblasts, we collected LI-NEDD4 KO and WT clones differentiated into fibroblasts and extracted total RNA. Polyadenylated RNA was reverse-transcribed using oligo-dT primer. To measure the expression level of the long transcript variant of *NEDD4*, we designed primers in the long-variant-specific exons (exon 1 and 8). To measure expression of the short transcript variant of *NEDD4*, we designed primers amplifying the junction of the short-variant-specific exon (exon 9) and an exon that are shared in both short and long variants (exon 14), because exon 9 is the only exon that is specific to the short variant. Beta-actin transcript was used as an internal control. We also measured the expression of *GAPDH*, and the linearity between beta-actin and *GAPDH* expressions across samples was confirmed. The relative expression levels of the *NEDD4* transcripts were calculated by $\Delta\Delta C_t$ method. We serially diluted cDNA to confirm that the qPCR conditions used resulted in exponential amplification. qPCR was performed on ViiA7 Real-Time PCR System using SYBR Green reagent. The sequences of the primers are listed in Supplementary Table 16.

Ethics approval

For all participating studies, we obtained informed consent from all participants by following the protocols approved by their institutional ethical committees. We obtained approval from the ethics committee of the RIKEN Center for Integrative Medical Sciences. We have complied with all the relevant ethical regulations.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

MEVs identified from 1000GP and summary statistics of eQTL analysis were uploaded in Zenodo (<https://doi.org/10.5281/zenodo.7703708>). The positions of and allele frequencies of MEVs identified from Japanese recruited by BBJ are available from National Bioscience Database Center (<https://humandbs.biosciencedbc.jp/en/>, accession ID: hum0014.v28) without any access restrictions. Summary statistics of GWAS are publicly available from our website (JENGER, <http://jenger.riken.jp/en>). Human reference genomes, human_g1k_v37, hs37d5, and GRCh38DH, are available from 1000GP repository (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz, and http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa, respectively). The 30× WGS data from 3,202 individuals recruited by 1000GP were downloaded from the 1000GP website (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/). SV callset generated by the phased assembly variant caller (PAV) in part of 1000GP is available from 1000GP repository (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/variants_freeze4_sv_insdel.tsv.gz). SV callset generated by Panenien in part of 1000GP is available from 1000GP repository (<http://ftp.1000genomes.ebi.ac.uk/>

[vol1/ftp/data_collections/HGSVC2/release/v1.0/PanGenie_results](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/PanGenie_results)). GATK-SV callset generated by Panenien in part of 1000GP is available from 1000GP repository (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20210124.SV_illumina_Integration/1KGP_3202.Illumina_ensemble_callset.freeze_V1.vcf.gz). Human repeat library is available from RepBase (<https://www.girinst.org/repbase>). Gene models, GenCode GTF v26 and v26lift37, are available from GenCode (<https://www.encodegenes.org/human>). SNVs found from participants in 1000GP are available from 1000GP repository (<http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502>). ENCODE cCRE dataset is available from ENCODE repository (<https://screen.encodeproject.org>). Accession numbers of histone ChIP-seq data from ENCODE are summarized in Supplementary Table 7. Other datasets of ESCs we used are available from ENCODE repository (<https://screen.encodeproject.org>) and deposited under accession numbers ENCFF601NBW (CpG methylation), ENCFF524BMX (CpG methylation), ENCFF379ZXG (CHG methylation), ENCFF086MMC (CHG methylation), ENCFF417VRB (CHH methylation), ENCFF918PML (CHH methylation), ENCFF000KUF (Repli-Chip), ENCFF000KUG (Repli-Chip), ENCFF000KUK (Repli-Chip), ENCFF905XDS (DNase-seq), ENCFF574LKL (DNase-seq), ENCFF338KTY (DNase-seq), ENCFF821AQO (CTCF ChIP-seq), ENCFF418QVJ (phospho-Pol-II A ChIP-seq), ENCFF422HDN (Pol-II ChIP-seq), and ENCFF834UVX (EP300 ChIP-seq). Annotations of DHS were obtained from data deposited under accession number ENCFF503GCK. Methylation data of iPSCs and ESCs taken by genome tiling array deposited under number accession [GSE60821](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60821) was used. Hi-C data of H1-hESCs deposited under accession numbers [GSM5057489](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5057489) and [GSM5057481](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5057481) was used. Gene expression profiles during spermatogenesis and early embryo deposited under accession numbers [GSE120508](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120508) and [GSE36552](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36552) were used. Read count tables of RNA-sequencing done by GTEx are available from GTEx repository (https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_exon_reads.parquet and https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz). Covariates for eQTL analysis and phased SNVs and indels in GTEx (that is PEER factors, genetic PCs, library preparation methods, sequencing platforms, and sex) are available from NCBI under dbGaP accession number phs000424. The script used to collapse GenCode GTF file is available from the URL below: https://github.com/broadinstitute/gtex-pipeline/blob/master/gene_model/collapse_annotation.py. The script used to apply TMM normalization is available from the URL below: https://github.com/broadinstitute/gtex-pipeline/blob/master/qlt/src/eqlt_prepare_expression.py. MEVs identified in Cao *et al.* are available from the following URLs: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02101-4>. Gene sets used for gene-set enrichment analysis is available from msigdb (<https://www.gsea-msigdb.org/gsea/msigdb/>). 7221 summary statistics of GWAS done by Pan-UKB were downloaded from the URLs listed in a file: https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_release/phenotype_manifest.tsv.bgz. The 'tophit' variants in BBJ were downloaded from http://jenger.riken.jp:8080/top_hits and <https://phweb.jp>.

Code availability

MEGANe is deposited in Zenodo (<https://doi.org/10.5281/zenodo.7703696>) and available from GitHub and Dockerhub (<https://github.com/shohei-kojima/MEGANe>, <https://hub.docker.com/repository/docker/shoheikojima/megane>). A complete environment including MEGANe and other required software is available from Dockerhub.

References

- Hirata, M. *et al.* Cross-sectional analysis of BioBank Japan clinical data: a large cohort of 200,000 patients with 47 common diseases. *J. Epidemiol.* **27**, S9–S21 (2017).
- Nagai, A. *et al.* Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

43. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
44. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
45. Ishigaki, K. et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).
46. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
47. Lee, C. H., Moioli, E. K. & Mao, J. J. Fibroblastic differentiation of human mesenchymal stem cells using connective tissue growth factor. *Conf. Proc. IEEE Eng. Med Biol. Soc.* **2006**, 775–778 (2006).

Acknowledgements

We are grateful to all of the participants in BBJ, as well as the staff of BBJ for their assistance. BBJ is supported by the Japan Agency for Medical Research and Development (AMED) (grant JP19km0605001). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. The GTEx data used for the analyses described in this study were obtained from dbGaP accession number phs000424.v8.p2 We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren and E. Wijsman). We appreciate obtaining access to genetic and pedigree data on SFARI Base. We acknowledge the resources of the 1000 Genomes Project and HGDP-CEPH Human Genome Diversity Cell Line Panel. We are grateful to the UK Biobank participants contributing to the results made public via the Pan-UK Biobank Resource and acknowledge the Pan-UKBB team (<https://pan.ukbb.broadinstitute.org/team>). Super-computing resources were provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo (SHIROKANE), and the Office for Information Systems and Cybersecurity, RIKEN (HOKUSAI General Use project G20021 and Q21537). We acknowledge H. Yoshida and T. Suzuki of RIKEN Center for Integrative Medical Sciences for providing plasmids, X. Chen and G. Bourque of Kyoto University and N. Sasa and Y. Okada of Osaka University for testing prerelease versions

of MEGAnE, K. Sato and J. Ito of University of Tokyo for helpful discussion, and M. Yoshioka for outstanding administrative support. S. Kojima, M.N. and A.J.K. acknowledge funding from the Incentive Research Projects of RIKEN, supported in part by Resona Bank. S. Kojima acknowledges a funding from Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Early-Career Scientists 22K15385. S. Kojima and S.M.H. acknowledge the RIKEN Special Postdoctoral Researcher Program. K. Ito acknowledges fundings from JSPS KAKENHI Grant-in-Aid for Scientific Research(B) 21H02919, AMED (JP22ek0210164, JP21tm0724601, JP20km0405209 and JP20ek0109487) and Research Funding for Longevity Sciences from the NCGG. N.F.P. acknowledges funding from JSPS KAKENHI Grants-in-Aid for Scientific Research(S) 20H05682, JSPS KAKENHI Grants-in-Aid for Scientific Research(B) 21H02972, RIKEN-McGill International Collaborative grant, Gout and Uric Acid Foundation of Japan, Cluster for Pioneering Research under the Hakubi fellowship program and from the discretionary budget of the Director of the RIKEN Center for Integrative Medical Sciences, K. Yamamoto.

Author contributions

S. Kojima and N.F.P. designed this study. S. Kojima, A.T., M.H., and N.F.P. developed MEGAnE. S. Kojima, S. Koyama, and N.F.P. analyzed data. N.F.P., K.H., K. Ishigaki, K. Ito, C.T., and Y.K. organized data in BBJ. S. Kojima, Y.S., M.E., M.M., S.T., Y.M., Y. Momozawa, and N.F.P. performed targeted deep sequencing. S.M.H., A.J.K., Y.M., Y. Murakawa, C.T., M.N., and Y.N. managed cell culture. S. Kojima, E.H.P., M.K., A.F.G., and R.K. performed wet experiments. S. Kojima and N.F.P. wrote and all authors checked the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-023-01390-2>.

Correspondence and requests for materials should be addressed to Shohei Kojima or Nicholas F. Parrish.

Peer review information *Nature Genetics* thanks Clement Goubert and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection We used a custom software, MEGAnE v0.1.1, for data collection. All original code is available at <https://github.com/shohei-kojima/MEGAnE>.

Data analysis We used publicly available software for data analysis. The software used are Python 3.7.4, pandas 1.3.1, numpy 1.19.2, seaborn 0.11.2, matplotlib 3.3.2, scipy 1.7.3, statmodels 0.12.2, scikit-learn 0.22.1, pystan 2.19.1.1, arviz 0.11.2, cooltools 0.5.0, hdf5 1.10.2, biopython 1.74, pysam 0.15.2, R 3.6.1, lme4 1.1.27.1, mashr 0.2.38, fgsea 1.19.3, coloc 4.0.4, SMR 1.03, SAIGE 0.44.5, BWA 0.7.17, Minimac3 2.0.1, Minimac4 1.0.2, SHAPEIT4 4.1.3, phenogram 1.2.1, samtools 1.10 and 1.14, bedtools v2.29.2, bcftools 1.9, PLINK v1.90b6.17, PLINK v2.00a2.3LM, RepeatMasker 4.0.9 and 4.1.0, MELT 2.1.5, HOMER (downloaded on Dec 29, 2021), fastQTL v2.184_gtex, mfepriemer 3.2.0, primer3-py 0.6.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

MEVs identified from 1000GP and summary statistics of eQTL analysis were uploaded in Zenodo (DOI: 10.5281/zenodo.7703708). The positions of and allele frequencies of MEVs identified from Japanese recruited by BBJ are available from National Bioscience Database Center (<https://humandbs.biosciencedbc.jp/en/>, accession ID: hum0014.v28) without any access restrictions. Summary statistics of GWAS are publicly available from our website (JENGER, <http://jenger.riken.jp/en/>).

In the course of benchmarking the tool presented here, we generated individual-level mobile element variant genotyping data from subjects enrolled in the Biobank Japan project. This is potentially-identifiable information and cannot be shared with anonymous reviewers, subject to ethical considerations and Japanese law. It will be made available to approved investigators upon application to the Biobank Japan consortium.

We used previously-published data as described below.

Human reference genomes, human_g1k_v37, hs37d5, and GRCh38DH, are available from 1000GP repository (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz, and http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa, respectively). The 30x WGS data from 3,202 individuals recruited by 1000GP were downloaded from the 1000GP website (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/). SV callset generated by the phased assembly variant caller (PAV) in part of 1000GP is available from 1000GP repository (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/variants_freeze4_sv_insdels.tsv.gz). SV callset generated by Panenier in part of 1000GP is available from 1000GP repository (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/PanGenie_results). GATK-SV callset generated by Panenier in part of 1000GP is available from 1000GP repository (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20210124.SV_illumina_Integration/1KGP_3202.Illumina_ensemble_callset.freeze_V1.vcf.gz). Human repeat library is available from RepBase (<https://www.girinst.org/repbase>). Gene models, GenCode GTF v26 and v26lift37, are available from GenCode (<https://www.gencodegenes.org/human>). SNVs found from subjects in 1000GP are available from 1000GP repository (<http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502>). ENCODE cCRE dataset is available from ENCODE repository (<https://screen.encodeproject.org>). Accession numbers of histone ChIP-seq data from ENCODE are summarized in Supplementary Table 7. Other datasets of ESCs we used are deposited under accession numbers ENCF601NBW (CpG methylation), ENCF524BMX (CpG methylation), ENCF379ZYG (CHG methylation), ENCF086MMC (CHG methylation), ENCF417VRB (CHH methylation), ENCF918PML (CHH methylation), ENCF000KUF (Repli-Chip), ENCF000KUG (Repli-Chip), ENCF000KUK (Repli-Chip), ENCF905XDS (DNase-seq), ENCF574LKL (DNase-seq), ENCF338KTY (DNase-seq), ENCF821AQO (CTCF ChIP-seq), ENCF418QVJ (phospho-Pol-II A ChIP-seq), ENCF422HDN (Pol-II ChIP-seq), and ENCF834UVX (EP300 ChIP-seq). Annotations of DHS were obtained from data deposited under accession number ENCF503GCK. Methylation data of iPSCs and ESCs taken by genome tiling array deposited under number accession GSE60821 was used. Hi-C data of H1-hESCs deposited under accession numbers GSM5057489 and GSM5057481 was used. Gene expression profiles during spermatogenesis and early embryo deposited under accession numbers GSE120508 and GSE36552 were used. Read count tables of RNA-seq done by GTEx are available from GTEx repository (https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNAseQCv1.1.9_exon_reads.parquet and https://storage.googleapis.com/gtex_analysis_v8/gene_reads.gct.gz). Covariates for eQTL analysis and phased SNVs and indels in GTEx (i.e. PEER factors, genetic PCs, library preparation methods, sequencing platforms, and sex) are available from NCBI under dbGaP accession number phs000424. The script used to collapse GenCode GTF file is available from the URL below: https://github.com/broadinstitute/gtex-pipeline/blob/master/gene_model/collapse_annotation.py. The script used to apply TMM normalization is available from the URL below: https://github.com/broadinstitute/gtex-pipeline/blob/master/ctl/src/eqlt_prepare_expression.py. MEVs identified in Cao et al. are available from the following URLs: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02101-4>. Gene sets used for gene-set enrichment analysis is available from msigdb (<https://www.gsea-msigdb.org/gsea/msigdb/>). 7221 summary statistics of GWAS done by Pan-UKB were downloaded from the URLs listed in a file: https://pan-ukb-us-east-1.s3.amazonaws.com/sumstats_release/phenotype_manifest.tsv.bgz. The "tophit" variants in BBJ were downloaded from the following URLs: http://jenger.riken.jp:8080/top_hits and <https://pheweb.jp>.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

We did not perform sex- and gender-based analysis, because the influences of mobile genetic elements on disease is our main focus rather than influences of sex and gender on disease.

None of our findings applies to only one sex or gender.

Population characteristics

BBJ is a prospective biobank that collaboratively collected DNA and serum samples from 12 medical institutions in Japan and recruited approximately 200,000 participants, mainly of Japanese ancestry. Mean age of participants at recruitment was 63.0 years old, and 46.3% were female. All study participants had been diagnosed with one or more of 47 target diseases by physicians at the cooperating hospitals.

Recruitment

All study participants in BBJ had been diagnosed with one or more of 47 target diseases by physicians at the cooperating hospitals. Participants were registered to the cohort from June 2003 to March 2008, and their clinical information was collected annually via interviews and medical record reviews until 2013.

Ethics oversight

All the participants provided written informed consent approved from ethics committees of the Institute of Medical Sciences, the University of Tokyo and RIKEN Center for Integrative Medical Sciences.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All sample size in GWASs in this study is summarized in the previous study, Ishigaki et al, 2020, Nature Genetics.
Data exclusions	All samples were selected based on quality-control criteria in each cohort, which is summarized in the previous report, Ishigaki et al, 2020, Nature Genetics.
Replication	<p>We confirmed all the attempts of replication were successful.</p> <p>We compared the GWAS hits identified in BBJ with independent GWASs in UK Biobank. Detection of LD in keloid GWAS was replicated in GWAS in UK Biobank.</p> <p>All the cell-based experiments were replicated. Number of replications are summarized below. Enhancer reporter assay: finding was replicated by four independent experiments. 3'UTR reporter assay: findings were replicated by at least three independent experiments. CRISPR-Cas9 KO experiments: findings were replicated by at least nine independent cell clones.</p>
Randomization	Randomization is not applicable, because we only performed population-based retrospective case-control study and cell-based experimental analyses (i.e. reporter assays and KO experiments).
Blinding	We did not apply blinding of the samples because this is a genotype-phenotype association study and no intervention was conducted in our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	Cell lines obtained from Coriell Institute: GM12878, GM18954, GM19088, GM20787, GM18944, GM18999 Cell lines purchased from ATCC: NT2/D1 (CRL-1973) The human iPSCs used was established by RIKEN BRC from PBMC of a healthy male.
Authentication	None of cells were used for authentication.
Mycoplasma contamination	None of the cells were tested for mycoplasma infection.
Commonly misidentified lines (See ICLAC register)	No commonly used misidentified cell lines were used in this study.