Computational and Structural Biotechnology Journal 30 (2025) 1–11

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

The topography of nullomer-emerging mutations and their relevance to human disease

Candace S.Y. Chan [a,b,c], Ioannis Mouratidis [c], Austin Montgomery [c], Georgios Christos Tsiatsianis [c], Nikol Chantzi [c], Martin Hemberg [d], Nadav Ahituv [a,b,*], Ilias Georgakopoulos-Soares [c,**]

[a] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA
[b] Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA
[c] Institute for Personalized Medicine, Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA, USA
[d] Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital, Massachusetts General Hospital, and Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Keywords:
Nullomers
CpG Islands
Pathogenicity

ABSTRACT

Nullomers are short DNA sequences (11–18 base pairs) that are absent from a genome; however, they can emerge due to mutations. Here, we characterize all possible putative human nullomer-emerging single base pair mutations, population variants and disease-causing mutations. We find that the primary determinants of nullomer emergence in the human genome are the presence of CpG dinucleotides and methylated cytosines. Putative nullomer-emerging mutations are enriched at specific genomic elements, including transcription start and end sites, splice sites and transcription factor binding sites. We also observe that putative nullomer-emerging mutations are more frequent in highly conserved regions and show preferential location at nucleosomes. Among repeat elements, Alu repeats exhibit pronounced enrichment for putative nullomer-emerging mutations at specific positions. Finally, we find that disease-associated pathogenic mutations are significantly more likely to cause emergence of nullomers than their benign counterparts.

1. Introduction

Nullomers are the shortest absent sequences from a genome and were initially discovered and annotated from the first sequenced human genome [16]. In the human genome, the first nullomers appear at eleven base pairs (bps) and the number of nullomers exponentiates with k-mer length. Even though nullomers are absent from the reference genome, they can be present in the genomes of other individuals. Germline mutations can be linked with the presence of nullomers, originally absent from the reference human genome [13], including rare variants [19]. Mutations that have arisen in the life of a person include private somatic mutations and clonal mutations, such as those that appear during cancer development and those can give rise to nullomers [14,22,38]. For cancer, presence of nullomers has been shown to be a useful biomarker for the early detection of cancer, using liquid biopsies [13,22,38]. Other applications involving nullomers have been described, including cancer cell killing and drug discovery targets [2–4], forensic applications [15], pathogen detection and surveillance [24,25,29,34] and immunogenic compounds [28]. Putative nullomer-emerging mutations have been only examined in a single study to date [14]. In that study, all possible single base pair substitutions and single base pair insertions and deletions that can cause the emergence of nullomers were examined. It was shown that nullomers generated from putative nullomer emerging mutations are enriched in promoters and coding regions, while the subset of nullomers that could emerge at hundreds of genomic loci were primarily found at Alu repeats.

The reasons why nullomers are absent from the human genome is an active area of research. Increased likelihood of mutagenesis at specific contexts, including at CpG sites, has been proposed [1], along with negative selection [13,19]. One such example was the identification of restriction site nullomers at viral genomes, which was suggested to be a mechanism safeguarding them against bacterial endonucleases [19]. In addition, a subset of nullomers was found to be shared between different organismal genomes, which could reflect stronger selection constraints

* Corresponding author at: Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA.
** Correspondence to: Institute of Personalized Medicine, Department of Biochemistry and Molecular Biology, Hershey, PA, USA.
E-mail addresses: nadav.ahituv@ucsf.edu (N. Ahituv), izg5139@psu.edu (I. Georgakopoulos-Soares).

https://doi.org/10.1016/j.csbj.2024.12.026
Received 25 October 2024; Received in revised form 20 December 2024; Accepted 22 December 2024
Available online 25 December 2024

[14,16,26,7]. Nevertheless, a thorough association between nullomer-emerging and functional genomic elements, or between nullomer-emerging and pathogenicity has yet to be investigated.

Here, we perform a systematic examination of nullomer-emerging mutations across the human genome. We analyze all possible one-base pair mutations, which we term putative nullomer emerging mutations throughout the manuscript. We find that putative nullomer-emerging mutations are enriched at early-replicating regions and at specific genomic sites, including transcription factor binding sites (TFBSs), CpG islands and relative to transcription start sites (TSSs) and splice sites. They are also preferentially positioned relative to nucleosomes. The strongest enrichment patterns for putative nullomer-emerging mutations are observed at CpG methylation sites, consistent with the increased mutation rate at these loci. Finally, we show their clinical relevance using disease causing mutations and pathogenic mutation sites, which are significantly more likely to cause nullomer emergence. In summary, we provide evidence for the mechanisms that cause nullomer formation and the selection constraints against them.

2. Results

For our analyses we used all human nullomers between the lengths of 11 and 13 bps, as described previously [13]. The shortest putative nullomer emerging length studied was the minimal length at which nullomers appear (11 bps). The upper length (13 bp) was selected as the largest k-mer length for which the number of putative nullomer emerging mutations is less than the number of bps of the human genome, which is 13 bps. For larger lengths, the excess of putative nullomer emerging mutations makes it harder to characterize the subset of putative nullomer-emerging mutations that are biologically relevant. The thirteen bp length limit was selected because for longer k-mer lengths the majority of loci generate nullomer-emerging mutations and therefore stochastic effects increase (40.09 nullomer mutations/ 1 kb). We examined all possible substitution and base-pair insertions or deletions for their potential for the emergence of nullomers. For 13 bp sequences, this analysis generated 271,432,758 putative nullomer-emerging mutations, categorized into insertions, deletions, and six types of substitutions (Fig. 1**a-b,** see **Methods**). On average, each putative nullomer-emerging mutation resulted in 2.03, 2.36 and 3.43 nullomers for 11 bp, 12 bp and 13 bp respectively. For each putative nullomer-emerging mutation, we generated a simulated mutation matched for trinucleotide context and located within 1 kb of the original mutation. These matched simulated mutations were used as controls for multiple comparisons to assess statistical significance.

3. Putative nullomer-emerging mutations are enriched in early replicating, genic regions and in *cis*-regulatory elements

To investigate the degree of putative nullomer-emerging mutation clustering in close genomic proximity to each other, we examined the distance between consecutive putative nullomer-emerging mutations relative to the expected distance from the simulations. We found a significantly different inter-mutation distance across nullomer lengths and mutation types, with putative nullomer-emerging mutations showing a 5-fold higher degree of clustering than expected by chance (p-value=0, Mann-Whitney U-test, Fig. 1**d**, Supplementary Figure 1a**-b,** Supplementary Table 1). This result suggests the presence of putative nullomer-emerging-mutation clusters in the human genome.

We split the genome into 1kB, 50kB or 500kB bins and calculated the number of unique k-mers per bin as well as the number of putative nullomer emerging mutations. We observe that in all cases, bins with nullomer emerging mutations have more unique k-mers than bins without (p-value=0, Mann-Whitney U-test, Supplementary Figure 1c, Supplementary Table 2). We conclude that genomic loci with putative nullomer emerging mutations are more likely to be in information-rich sequences of the human genome.

Next, we examined if putative mutations that cause the emergence of nullomers are differentially distributed within the human genome and within functional genomic elements. We separated mutations into coding and non-coding and observed a higher density of coding relative to non-coding putative nullomer-emerging mutations across k-mer lengths (p-value=0, Mann-Whitney U-test, Fig. 1**c,** Supplementary Figure 1d**,** Supplementary Table 3) and across mutation categories (Supplementary Figure 1e). Importantly, the mutational density was 8.59-fold higher in coding relative to non-coding regions, indicating that putative nullomer-emerging mutations are preferentially located at coding sites.

Replication timing stratifies multiple genomic features, which include gene organization, histone modifications, DNA methylation, heterochromatization and likelihood of mutagenesis [35,36,5]. Repli-Seq is a method used to infer the replication timing of different regions across the genome in a cell type. We used Repli-Seq data from fourteen human cell lines (BG02ES, BJ, GM06990, GM12801, GM12812, GM12813, GM12878, HeLaS3, HepG2, HUVEC, IMR90, K562, MCF7 and NHEK cell lines) [8] to study the distribution of putative nullomer-emerging mutations relative to replication timing. We separated the data into deciles, based on the replication timing of genomic regions and examined the density of putative nullomer-emerging mutations in each decile. We observed that early replicating regions had an excess of putative nullomer-emerging mutations (Fig. 1**e,** Pearson correlation r = 0.97, p-value = 1.94E-06, Supplementary Table 4). Separation of putative nullomer-emerging mutations by mutation category revealed that the largest difference between early and late replicating regions in mutation density was observed for insertions and substitutions relative to deletions (Supplementary Figure 1**f**). We also examined how the mutational density of each substitution type for putative nullomer-emerging mutations changed across the replication timing deciles and found that G>C mutations (or equivalently C>G) showed the most pronounced differences across replication timing deciles (Supplementary Figure 1**g**). These results indicate that the emergence of nullomers is more likely to occur in early replicating and coding regions, genomic regions with higher GC content which are under stronger selection constraints.

4. Putative nullomer-emerging mutations are enriched at functional genic sites in transcribed regions

We next examined the distribution of putative nullomer-emerging mutations relative to functional genomic sites at base-pair resolution. Promoter sequences are composed of specific regulatory elements such as the TATA-box, the INR element and TFBSs, which tend to be under evolutionary constraint. We therefore reasoned that putative nullomer-emerging mutations, which could impair transcriptional activity, would be enriched relative to transcription start sites (TSSs). Indeed, we found an enrichment of 1.79-fold immediately upstream of the TSS for mutations that cause the emergence of 13 bp nullomers (Fig. 1**f**). We also adjusted for the simulated mutations, finding an enrichment of 1.53-fold of putative nullomer-emerging mutations around the TSS (Kolmogorov-Smirnov test, p-value=4.07E-59, Supplementary Figure 2a**,** see **Methods**). In particular, the enrichment levels for substitutions, insertions and deletions were 1.79-fold, 1.76-fold and 1.99-fold (Supplementary Figure 2b), with the substitution type and indel type influencing the likelihood of putative nullomer emergence. Similar results were obtained for 11 bp and 12 bp putative nullomer-emerging mutations (Supplementary Figure 2a**-c**) and when adjusting for the simulated mutation controls. These results indicate that putative nullomer-emerging mutations are enriched relative to the TSSs.

We replicated this methodology relative to 3' splice sites (3'ss) and 5' splice sites (5'ss) as well as relative to Transcription End Sites (TESs). We found that the enrichment levels ranged across these elements with 3'ss, 5'ss and TES showing enrichments of 2.07-fold, 2.10-fold and 1.23-fold respectively relative to surrounding regions (Fig. 1**f-g**). We also adjusted for the simulated mutations, finding adjusted enrichments of 1.50-fold,

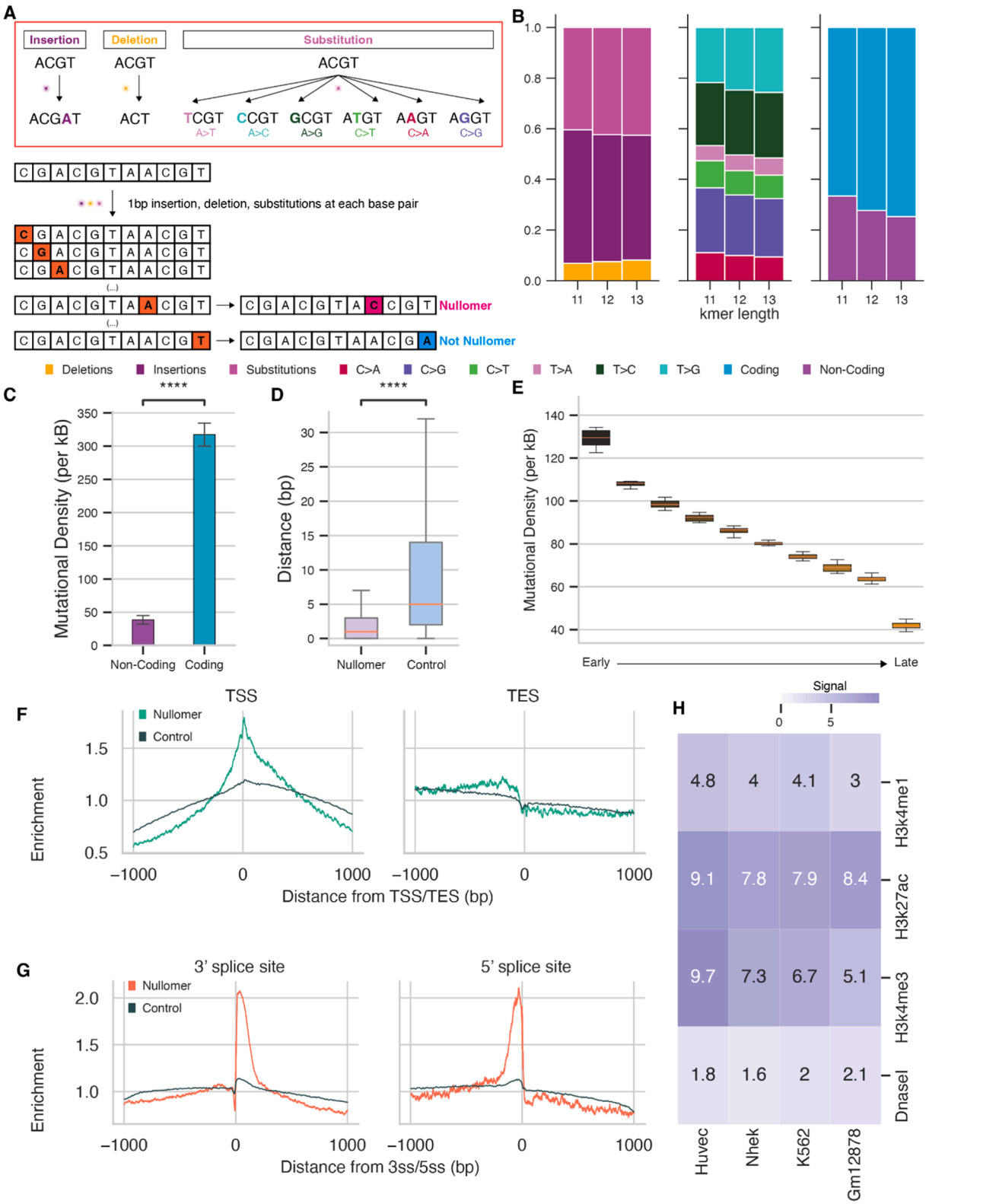

Fig. 1. Putative nullomer-emerging mutations show clustering patterns and are enriched in early-replicating regions, promoters and coding sequences. a) Schematic of simulated mutations to generate putative nullomer-emerging mutations and control mutations. b) Proportion of putative nullomer-emerging mutations by mutation type for k-mer lengths of 11 bp, 12 bp and 13 bp. c) Distance distribution between consecutive putative nullomer-emerging mutations and simulated putative nullomer-emerging mutations. d) putative nullomer-emerging mutational density at coding and non-coding regions. e) Density of putative nullomer-emerging mutations across replication timing deciles. Early replicating regions display a higher nullomer emergence density (Pearson correlation r = 0.97, p-value = 2.4e-06). Mean mutational density across all fourteen cell lines are shown. Error bars indicate standard deviation of mutational density between cell lines. f-h) Enrichment of putative nullomer-emerging mutations in: f) TSS, TES and g) 3'ss, 5'ss, and h) at DNase footprinting sites and histone modifications. In f-g, the fold enrichment for putative nullomer-emerging mutations was calculated as the ratio of the number of mutations found in a given position over the mean number of mutations across the whole window. The dark grey lines in f-g represent the distribution of putative nullomer-emerging mutations for simulated controls.

1.72-fold and 1.05-fold at the 3'ss, 5'ss and TES respectively (Supplementary Table 5). The 3'ss and 5'ss mutation types with lowest and highest adjusted enrichments were insertions and substitutions with 1.70-fold and 1.56-fold enrichments and were replicated for 11 bp and 12 bp with high consistency (Supplementary Figure 2d-**f**). The enrichment of putative nullomer-emerging mutations at TSS and splice sites indicates that these mutations are prone to affect transcriptional activity.

We also investigated if putative nullomer-emerging mutations are enriched for specific epigenetic modifications, including histone modifications and open chromatin marks. To that end, we analyzed H3K4me3, H3K27ac, H3K4me1 and DNaseI data across four human cell

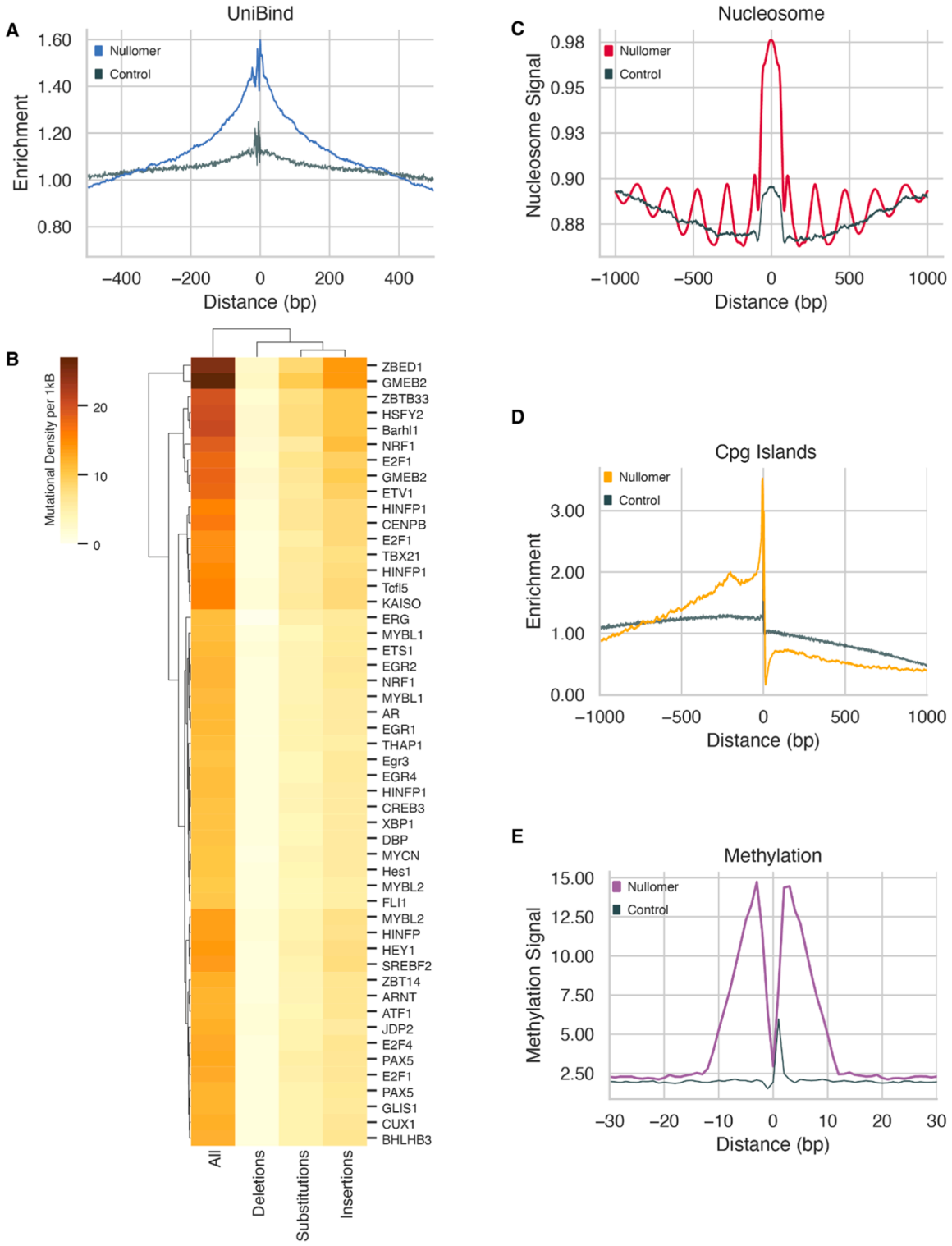

Fig. 2. Association between putative nullomer-emerging mutations and open epigenetic marks and methylation. Putative nullomer-emerging mutation sites are enriched at a) TF-DNA interaction sites based on UniBind inferred data, b) transcription factor binding sites, c) nucleosome core positions, d) CpG islands, and e) methylation sites from WGBS.

lines (Fig. 1**h,** Supplementary Figure 2g-**h**). We find consistently that H3K27ac and H3K4me3 are most enriched for putative nullomer-emerging mutations (mean enrichments of 8.2-fold and 7.1-fold). We conclude that epigenetic marks are linked to differences in putative nullomer emergence frequencies.

5. Transcription factor binding sites show an excess of putative nullomer-emerging mutations

We next examined whether putative nullomer-emerging mutations are associated with certain TFBSs. First, we used collapsed consensus motifs from genome-wide DNase footprinting data generated from 263 cell and tissue types [40](see **Methods**), which have been previously shown to reflect bound TFBSs [11]. We observed that 39.89 % of DNase footprints overlapped one or more putative nullomer-emerging mutations representing a 2.13-fold enrichment (Supplementary Figure 3a). When adjusting based on our simulated controls we find an enrichment of putative nullomer-emerging mutations of 1.56-fold, suggesting that nucleotide composition accounts for a subset of the pattern. We observed consistent patterns when we replicated this analysis across nullomer lengths and separating by mutation category (Supplementary Figure 3a-**c**), with deletions observed to have the strongest enrichment (1.70-fold) and substitutions (1.55-fold) showing the weakest enrichment.

To verify these findings, we used a collection of TFBSs derived from the UniBind database [30]. In this dataset, for every ChIP-seq peak the corresponding TFBS is inferred and therefore it reflects high-confidence transcription factor bound TFBSs. We analyzed this dataset to examine potential enrichment of putative nullomer-emerging mutations within actively-bound TFBSs. We found that the results obtained were highly consistent with those obtained using DNase footprinting (Fig. 2**a,** Supplementary Figure 3d-**f**), with 40.68 % of TFBSs overlapping one or more putative nullomer-emerging mutation, representing an overall enrichment of 1.18-fold over background rates and providing additional support that nullomer emergence occurs more frequently at TFBSs. We were also interested in investigating potential differences in the enrichment of putative nullomer-emerging mutations by transcription factor category. We examined the density of putative nullomer-emerging mutations across the TFBSs of individual transcription factors; we observed that a subset of transcription factors had a high nullomer-emerging mutation density, with the highest densities occurring with ZBED1 and GMEB2 among others (Fig. 2**b**). These findings suggest an enrichment of putative nullomer-emerging mutations at TFBSs across the human genome.

6. Putative nullomer-emerging mutations are preferentially positioned relative to nucleosomes

We reasoned that transcription factor binding site accessibility during transcriptional activity can be influenced by chromatin organization. Thus, we examined if nucleosome positioning influences the likelihood of nullomer emergence. Micrococcal Nuclease (MNase) data are generated by MNase digestion, in which exposed DNA regions are digested which in turns enables the derivation of nucleosome positioning. We used available MNase data for K562 and GM12878 cell lines from the ENCODE Consortium [9] to identify whether putative nullomer-emerging mutations show a preference for nucleosome core or linker regions. We found that 13 bp putative nullomer-emerging mutations show an 1.08-fold enrichment at nucleosome core sequences, with a periodicity that is approximately the size of the inter-nucleosome distance [32] (Fig. 2**c**). The results were also replicated for 11 bp and 12 bp mutations with very similar results obtained (Supplementary Figure 4a). However, when we separated by mutation type and repeated the same analysis we found significant differences. Overall, between substitutions, insertions and deletions we found largely consistent results with nucleosome cores displaying an enrichment for putative nullomer-emerging mutations (Supplementary Figure 4b); however when we separated by substitution type we observed that G>A (and C>T) and G>T (and C>A) mutations were more likely to be found at the linker regions, whereas all other substitution types were enriched for the nucleosome core sequence (Supplementary Figure 4c). Our findings indicate that nullomer emergence, as examined using all putative mutations, is influenced by nucleosome positioning and by the mutation type.

7. CpG sites are the primary determinants of nullomer emergence in the human genome

Previous reports have suggested that CpG sites are hypermutable; we therefore examined the association between nullomer emergence and presence of CpG dinucleotides [37,39]. We found that putative nullomer-emerging mutations are highly enriched in CpG dinucleotides, with 100 % (104), 100 % (44,287) and 99.996 % (2347,572) of them harboring one or more CpG dinucleotides for 11 bp, 12 bp and 13 bp respectively. This is particularly unexpected given that the percentage of k-mers that harbor CpG dinucleotides in all possible 11,12, and 13 bp k-mers are 69 %, 72 %, and 75 %. This result suggests that the primary driver of nullomer generation is presence of CpG dinucleotides. We also examined if putative nullomer-emerging mutations are enriched at CpG islands (binomial test, p-value = 0, Fig. 2**d**, Supplementary Table 6**,** see **Methods**). We found that 22.4 %, 95.8 % and 96.9 % of CpG islands contain one or more putative nullomer-emerging mutation for 11 bp, 12 bp and 13 bp nullomers respectively, with 1.09-fold, 1.10-fold and 2.33-fold enrichments at substitutions, insertions and deletions respectively when compared to simulated mutations (Supplementary Figure 4d-**f**). We then examined putative nullomer-emerging mutations at CpG islands, at CpG shelves defined as within 2kB distance regions from a CpG island and CpG shelves defined as 2–4kB distance regions from CpG islands. We observe, that in all cases putative nullomer-emerging mutations are enriched relative to simulated controls; however the highest enriched is observed within CpG islands (1.4-fold enrichment), followed by CpG shores (0.45-fold enrichment) and CpG shelves (0.46-fold enrichment) (Supplementary Figure 7).

Next, we used whole-genome bisulfite sequencing (WGBS) from the ENCODE Consortium [8] on six different tissues (adrenal gland, esophagus squamous epithelium, gastroesophageal sphincter, stomach, small intestine, spleen) to examine methylation of DNA at the fifth position in cytosine (5mC), across putative nullomer-emerging mutation sites. Across all putative nullomer-emerging mutations, we observed an enrichment of 2.31-fold, directly at the 5mC sites (Mann Whitney U-test, p-value=0, Fig. 2**e,** Supplementary Figure 3**g,** Supplementary Table 5), findings that were consistent across mutation categories (Supplementary Figure 3h-**i**). These results indicate that putative nullomer-emerging mutations occur at CpG sites, which are the most frequently mutated dinucleotides in the human genome [10]; therefore these enrichments likely reflect hyper-mutation at these sequences. We conclude that a driver of nullomer generation in the human genome is the CpG mutation rate.

8. Alu repeats display positional enrichment for putative nullomer-emerging mutations

We examined the distribution of putative nullomer-emerging mutations across transposable elements. We analyzed emerging nullomers in Long interspersed nuclear elements (LINE), Short interspersed nuclear elements (SINE) and Long Terminal Repeats (LTR) transposable element families and found that the highest nullomer-emerging density from all putative one bp mutations was observed at SINE repeats with median of 60 mutations per kB (Fig. 3**a**). In particular, we observed this for insertions relative to substitutions and deletions and the findings were consistent across k-mer lengths (Fig. 3**a;** Supplementary Figure 5). We also separated the transposable repeat element families into sub-types

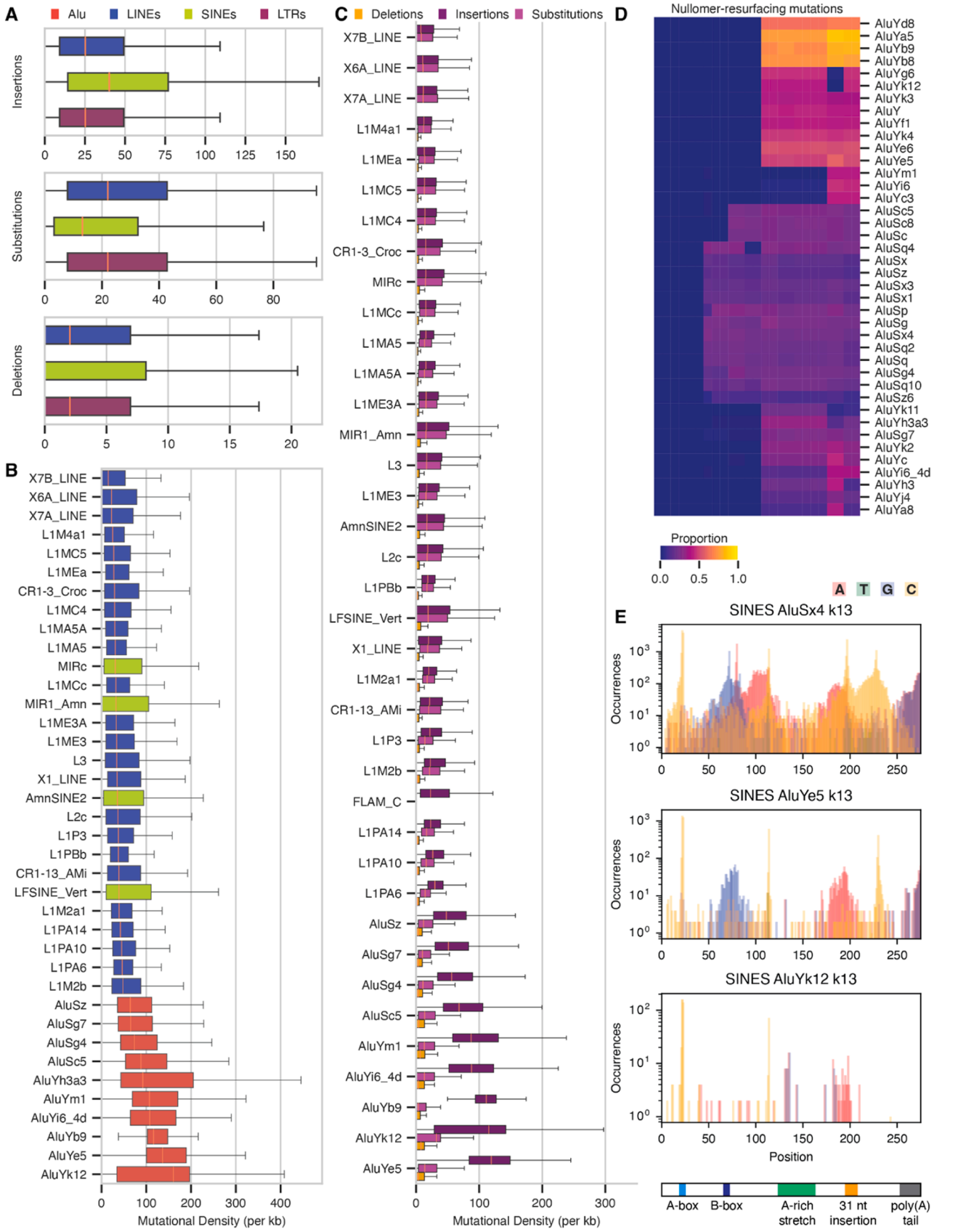

Fig. 3. Nullomer emergence is pronounced at Alu repeat elements. a) Mutational density at LINE, SINE and LTR repeats across mutation types. b) Mutational density at transposable element repeat sub-families. c) Mutational density at transposable element repeat sub-families for the mutation subtypes. d) Heatmap of proportion of putative nullomer-emerging mutations appearing at Alu repeat elements. e) Distribution of putative nullomer-emerging mutations occurrences across Alu repeat elements AluYe5, AluYk12, AluSx4.

and found that the most recently active Alu repeats in the human genome, including AluSx, AluJ and AluY repeats displayed the highest putative nullomer-emerging mutational density (Fig. 3**b**), and specifically for insertions (Fig. 3**c**). The repeat elements with the highest putative nullomer-emerging mutation density were AluYe5, AluYk12 and AluYb9 (mean densities 155.4, 145.0, 139.5 elements per kb, Fig. 3**c**).

We next examined the subset of nullomers that had the highest density of putative nullomer-emerging mutations for single base-pair substitutions, insertions and deletions in the human genome. We find that the recurrent putative nullomer-emerging mutations can be explained by their presence at recently evolved Alu repeats, particularly for insertions (Fig. 3**d**). Additionally, we observe that the mutations are inhomogeneously distributed across the Alu repeat elements (Fig. 3**e**), with specific hotspots, notably at regions with AluYa5. These findings provide support for putative nullomer-emerging mutational hotspots in the human genome, likely reflecting selection constraints and silencing mechanisms that have been operative during recent human evolution.

9. Nullomer-emergence in human population variants

Previous work has showcased the utility of nullomers in forensic applications [15]. We have also previously shown an association between population variants and nullomer emergence [14]. Here, we further investigate the relationship between the likelihood of nullomer emergence and pathogenicity. We first examined the likelihood of nullomer emergence due to human population variants derived from dbSNP. Interestingly, we observe a negative correlation between SNP variant allele frequency and the likelihood of nulloemr emergence (Fig. 4**a**). Nullomer-emerging germline mutations were more likely to be found in the rarest population variants with minor allele frequencies (MAFs) below 0.01. To understand whether rare germline mutations that can lead to nullomers emergence may have a deleterious effect, we then examined the pathogenicity of these variants. Using scores derived from CADD [31] and population variants derived from gnomAD [18], we examined the predicted pathogenicity of nullomer-emerging mutations. We find that across mutation types, nullomer-emerging germline mutations found in gnomAD yielded a higher CADD score than in simulated mutations, with insertions yielding the strongest enrichment (Figs. 4**b,** 1.58-fold**,** Mann-Whitney *U* test, p-value<0.01, Supplementary Figure 6a**,** Supplementary Table 7**)**. These findings suggest that nullomer emerging germline mutations are rare in the human population and associated with pathogenicity.

10. Predicted pathogenic mutations are more likely to result in nullomer emergence

Assuming that some of the causes of nullomer absence include selection constraints and pathogenicity, we examined the pathogenicity of nullomer-emerging mutations. We analyzed the deleteriousness of all possible putative nullomer-emerging substitutions relative to all putative substitutions that do not cause nullomer emergence, throughout the human genome. We found that putative nullomer-emerging substitutions on average display a higher CADD score for 13 bp nullomer emerging mutations compared to controls (1.11-fold, Mann-Whitney U, p-value<0.01, Supplementary Table 7). We also separated putative substitutions into ten quantiles based on the deleteriousness of each substitution. We found that the most pathogenic putative substitution mutations are also the most likely to cause nullomer-emergence (Fig. 4**c**, Spearman correlation = 0.92, p-value = 1.86E-5). Furthermore, when separating putative nullomer-emerging mutations and mutations that do not cause nullomer emergence into deciles based on the CADD score, across the deciles putative nullomer-emerging mutations have a higher pathogenicity (Fig. 4**d**).

We next separated the mutation types based on their effect into: i) 3'UTR, ii) 5'UTR, iii) canonical splice, iv) stop gained, v) stop loss, vi) synonymous and vii) non-synonymous. We examined which of these showed the largest discrepancy in pathogenicity when they caused nullomer-emerging versus when they did not cause nullomer-emerging. We observed that putative mutations in 5'UTR showed the strongest enrichment (1.17-fold) while the 3'UTR were under-enriched (0.97-fold) (Fig. 4**e**, Supplementary Table 7). We obtained consistent results across nullomer lengths (Supplementary Figure 6b).

We then examined if putative nullomer-emerging mutations are enriched in amino acid substitutions that are predicted to affect protein function. Using SIFT scores we show that putative nullomer-emerging mutations are slightly more likely to affect protein function (Figs. 4**f,** 1.02-fold, Mann-Whitney U, p-value=0, Supplementary Table 8). These findings provide evidence for the higher likelihood of pathogenicity for putative nullomer-emerging substitutions across mutation types, for mutations that cause protein sequence changes and across functional genomic compartments.

11. Pathogenic mutations cause the emergence of nullomers in the human genome

ClinVar is a database that encompasses manually curated mutations that have been annotated relative to their pathogenicity status as "Benign" or "Likely Benign" and "Pathogenic" or "Likely Pathogenic" [20]. We examined if clinically relevant nullomer-emerging mutations show an enrichment for Pathogenic or Likely Pathogenic relative to Benign or Likely Benign mutations. We find that across nullomer lengths, pathogenic mutations are more likely to cause the emergence of nullomers (Supplementary Figure 6c**,** chi-square test, p-value=3.26E-243, Supplementary Table 9). We find that pathogenic clinical variants show a 1.34-fold enrichment in causing 13 bp nullomer emergence over the frequency of benign and likely benign clinical variants, in particular for deletions and insertions relative to substitutions (Fig. 4**g**). Across substitution types, T > C mutations are the most enriched type (Supplementary Figure 6d). We also examined the frequency of nullomer-emerging mutations in expression quantitative trait loci (eQTL) relative to common SNPs and found a 1.18-fold enrichment at eQTLs for 13 bp nullomers (Fig. 4**h,** Supplementary Table 10). This was replicated across mutation types (Supplementary Figure 6e**).** These results indicate that nullomer emergence is associated with pathogenicity and human disease.

12. Discussion

Here, we performed a thorough genomic characterization of nullomer-emerging mutations across the human genome and provided new evidence that increased mutation rate in methylated cytosines [27] are the primary driver of nullomer emergence. We find a remarkable 16.12-fold enrichment of nullomer-emerging mutations in 5-methylcytosines. We also find almost every nullomer encompassing one or more CpG dinucleotides, suggesting that hypermutation of CpG loci is the principal force of nullomer formation, consistent with the proposition by [1].

We also find that putative nullomer-emerging mutations are inhomogeneously distributed across the human genome, with clear enrichment patterns across functional genetic elements and *cis*-regulatory sequences. Therefore, putative nullomer-emerging mutations provide evidence for negative selection constraints within those elements, which is the second force driving nullomer formation. Selection constraints against nullomers are highest at TFBSs, splice sites and the TSS. Previously, we and others have shown that selection constraint can be detected for nullomers in humans as well as in other species [13,19].

Selection constraints against nullomer emergence are evident at functional genomic sites. In addition, the intriguing putative nullomer-emerging mutation enrichment at the most recently evolved Alu repeats, might indicate repeat silencing mechanisms. Specifically, we find that only the most recently evolved Alu repeats, including the small subset of Alu repeat sub-families that are still active in humans [6,17],

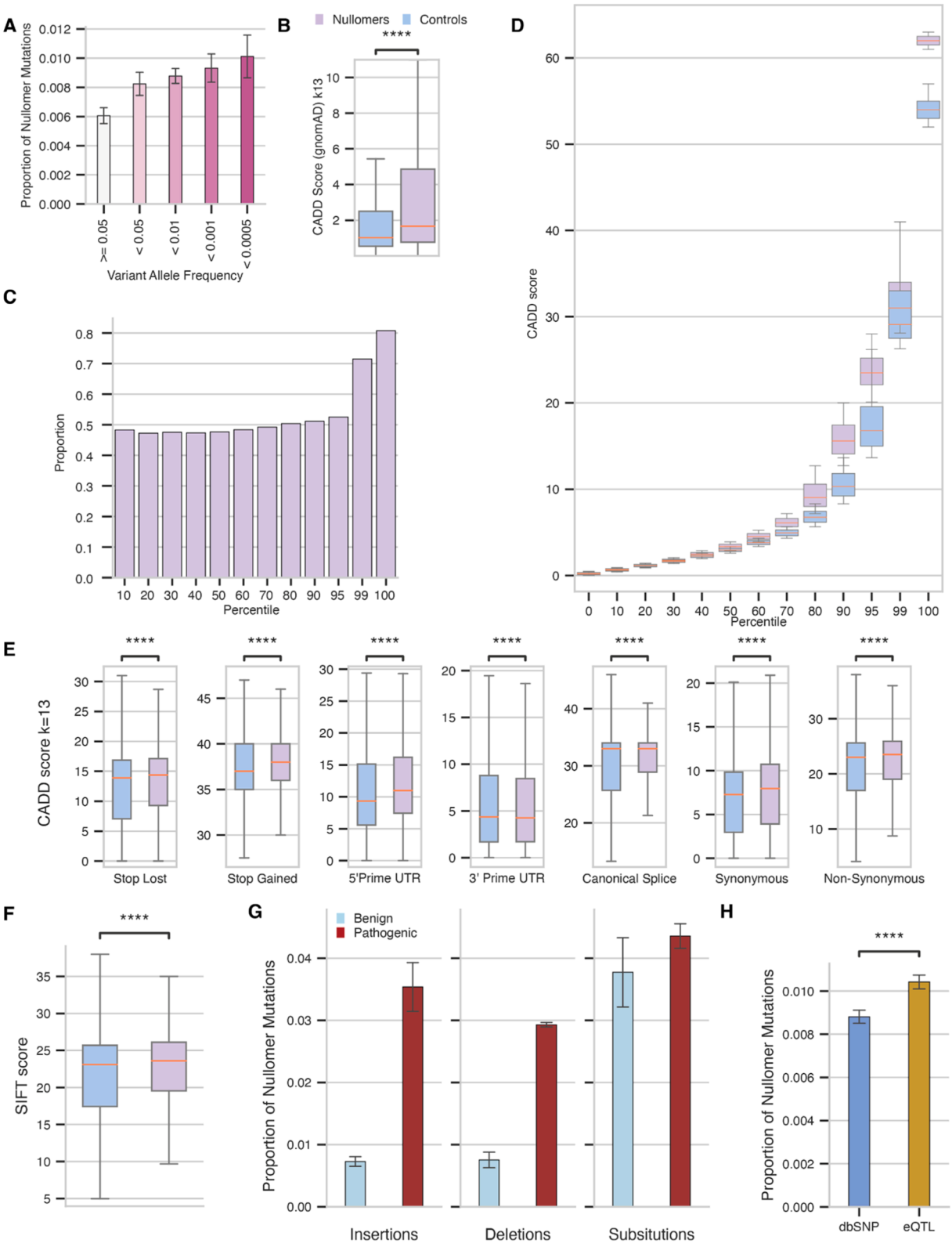

Fig. 4. Pathogenicity of nullomer-emerging sequences in the human genome. a) Variant allele frequencies of nullomer-emerging population variants. b) CADD scores of mutations that cause or not cause the emergence of nullomers in population variants from gnomAD. c) Proportion of nullomer-emerging mutations across CADD score percentiles. d) Distribution of CADD scores across percentiles in nullomer-emerging mutations and non-nullomer emerging mutations. e) Association of CADD score and nullomer emergence for mutations at the stop codon mutation loss, stop stop codon mutation gain, 5'UTR, 3'UTR, canonical splice sites, synonymous and nonsynonymous mutations. f) SIFT scores from nullomer emerging mutations and non-nullomer emerging mutations. g) ClinVar pathogenic mutations are more likely to cause nullomer emergence than their benign counterparts. h) Enrichment of eQTLs relative to common SNPs for nullomer-emerging mutations. In panels a-f, controls are based on simulations controlling for trinucleotide context and proximity to the original nullomer emerging mutation.

show the pronounced nullomer emergence hotspots. The mutation type we observed at these hotspots is primarily insertions, suggesting that the silencing mechanisms which are operative could likely involve deletion events. Therefore, future work is required to investigate the hypothesis that nullomer emergence at these sites can cause an increased Alu repeat activity.

Finally, we showcase how disease-causing and pathogenic mutations are more likely to create nullomer-emerging mutations. Also, rare germline mutations, which are more likely to be pathogenic, are also strongly associated with nullomer emergence. This highlights the potential of nullomers and their emergence, a topic which has been severely understudied, and how it can provide breakthroughs in the understanding of human diseases. Nullomer emergence could be used to find loci that are more likely to be associated with human diseases. It could also be useful in other species, for which disease annotations and identification of pathogenic variants is still lacking. Future work is required to elucidate the mechanisms of pathogenicity at individual nullomer-emerging hotspots and to deconvolute them from the increased mutagenicity of nullomers.

13. Materials and methods

Nullomer extraction and putative nullomer-emerging mutation map generation Nullomer extraction and putative nullomer-emerging mutation map generation was performed as described previously [14] for k-mer lengths of eleven to thirteen base pairs. At each genomic position of the reference genome, we systematically changed each nucleotide to all three other possibilities and compared the resulting k-mer to a list of nullomers. We categorized the mutation types into three classes: substitutions, insertions and deletions. Substitutions were subdivided into six subtypes based on the reference and alternate allele.

Simulated mutations For each putative nullomer-emerging mutation, we identified a position with a distance less than 1000 bp away that had the same trinucleotide context but did not lead to nullomer emergence. Simulated mutations were generated using a custom Python script that is found in [12]. Using this methodology we created a set of simulated mutations that matched the nullomer-emerging mutations, for each nullomer length.

14. K-mer enrichment analysis

A sliding window was used to find sites of nullomer emergence in each genomic region. The distribution of mutations were calculated by scoring the number of mutations at each genomic region, within 1kB. The enrichment was calculated as the number of nullomer-emerging mutation or simulated mutation occurrences at a position over the mean number of occurrences across a window of 1kB. The corrected enrichment was calculated as the ratio of the real enrichment over the background enrichment of simulated mutations.

K-mer analysis across genomic bins To investigate the association between k-mer diversity and nullomer emerging mutations we split the genome into 1kB, 50kB or 500kB bins and calculated the number of unique k-mers per bin and the number of putative nullomer emerging mutations in each bin. From the analysis, we removed the first and last 50kB regions as they are highly repetitive due to telomeric sequences. To estimate statistical significance we performed Mann Whitney U tests.

Genomic and genic annotation data. The reference human genome assembly GRCh38 (hg38) was used. Genic analyses were performed using the GENCODE v40 annotation, for which coding and non-coding regions, TSS, TES, 3'ss and 5'ss annotations were derived. Locations of CpG islands were obtained from the UCSC genome browser. The enrichment was calculated as the number of occurrences at a position over the mean number of occurrences across the window of 1kB. The corrected enrichment was calculated as the ratio of the real enrichment over the background enrichment of simulated mutations.

Repli-seq data. Repli-seq data for fourteen cell lines were derived from [8] and analyzed as previously described in [23]. Repli-seq data were binned into deciles relative to early and late replicating regions. The density per 1kB window of nullomer-emerging mutations was calculated at each decile across mutation categories for all cell lines. Pearson correlation was calculated between the decile number and the mean mutational density at each decile.

CpG analysis CpG island hg38 coordinates were downloaded from UCSC Genome Browser. Coordinates for CpG shores and shelves were found by expanding coordinates by 2 kb and 4 kb respectively from islands using bedtools slop. The island regions were excluded from shores and shelves using bedtools subtract.

15. TFBS datasets

TFBSs at DNase footprints from 243 human cell and tissue types and states were obtained from [40] and ChIP-seq bound TFBSs were obtained from UniBind [30]. We measured the distribution of nullomer or simulated mutations across 1kB windows. The enrichment was calculated as the number of occurrences at a position over the mean number of occurrences across the window of 1kB. The corrected enrichment was calculated as the ratio of the real enrichment over the background enrichment of simulated mutations. The density of TFBS motifs was calculated as the number of motif occurrences over the total number of base pairs.

MNase datasets. MNase-seq data were downloaded from the ENCODE [8] portal for GM12878 and K562. Significance of difference in nucleosome density signal was calculated using scores from nullomer or control mutations extracted with bedtools map function, followed by Mann-Whitney *U* test. Nucleosome signal was calculated as mean score at each loci in the 1kB window of nullomer-emerging mutation and simulated control mutation.

ClinVar, dbSNP mutation and eQTL datasets. Population variants were derived from dbSNP [33] and gnomAD [18]. Clinical variants were derived from the ClinVar database [20] and were subdivided based on pathogenicity in "Benign", "Likely Benign", "Likely Pathogenic" and "Pathogenic". The frequency of nullomer-emergence was compared between mutations of different pathogenicity. eQTLs were derived from GTEx Portal (GTEx_Analysis_v8_eQTL.tar)[21].

Measurement of deleteriousness using CADD. The CADD tool was used for scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome. All single nucleotide variants across the human genome were derived from: https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/whole_genome_SNVs.tsv.gz. All gnomAD substitutions were derived from: https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/gnomad.genomes.r3.0.snv.tsv.gz. All gnomAD insertion and deletion mutations were derived from: https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/gnomad.genomes.r3.0.indel.tsv.gz. All SIFT scores were derived from: https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/whole_genome_SNVs_inclAnno.tsv.gz. Nullomer extraction was performed for each mutation for nullomers of lengths between eleven and thirteen base pairs. Mutations were separated into those that caused nullomer emergence and those that did not and the significance of difference in deleteriousness between the two groups was calculated using Mann-Whitney U tests.

Whole Genome Bisulfite Sequencing Analysis. WGBS data were downloaded from ENCODE for six different human tissues, the adrenal gland (ENCFF524MTO), esophagus squamous epithelium (ENCFF283YAZ), the gastroesophageal sphincter (ENCFF441OSB), stomach (ENCFF896GOF), small intestine (ENCFF537NCQ) and spleen (ENCFF865OXJ) tissues. Methylation signal was calculated as the mean score at each loci in the 1kB window of nullomer-emerging mutation and simulated control mutation.

Author contributions

C.C. N.A., and I.G.S. conceived the study. C.C, I.M. A.M. and I.G.S., wrote the code, C.C., A.M, and I.G.S., performed the analyses and generated the visualizations. N.A. and I.G.S. supervised the research, I. G.S., and C.C wrote the manuscript with input from all authors.

Code availability

All code to perform case study analysis is provided at https://github.com/Georgakopoulos-Soares-lab/nullomer_topography.

CRediT authorship contribution statement

Candace S.Y. Chan: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Austin Montgomery:** Formal analysis. **Ioannis Mouratidis:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Nikol Chantzi:** Formal analysis. **Georgios Christos Tsiatsianis:** Formal analysis. **Nadav Ahituv:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Martin Hemberg:** Writing – review & editing, Supervision. **Ilias Georgakopoulos-Soares:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Competing Interest

N.A. is a cofounder and on the scientific advisory board of Regel Therapeutics.

Acknowledgements

I.G.S, I.M, A.M, C.S.Y.C, and N.C. were funded by startup funds provided by the Penn State College of Medicine. I.G.S. was also funded in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM155468. C.S.Y.C was funded in part by UCSF Hillblom Center for the Biology of Aging and Bakar Aging Research Institute Graduate Fellowship. N.A. was funded in part by the National Human Genome Research Institute grant number 1UM1HG011966 and National Institute of General Medical Sciences grant number R01GM142112. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.12.026.

References

[1] Acquisti Claudia, Poste George, Curtiss David, Kumar Sudhir. Nullomers: really a matter of natural selection? PLoS One 2007;2(10):e1022.
[2] Alileche Abdelkrim, Goswami Jayita, Bourland William, Davis Michael, Hampikian Greg. Nullomer derived anticancer peptides (NulloPs): differential lethal effects on normal and cancer cells in vitro. Peptides 2012. https://doi.org/10.1016/j.peptides.2012.09.015.
[3] Alileche Abdelkrim, Hampikian Greg. The effect of nullomer-derived peptides 9R, 9S1R and 124R on the NCI-60 panel and normal cell lines. BMC Cancer 2017;17(1):533.
[4] Ali Nilufar, Wolf Cody, Kanchan Swarna, Veerabhadraiah Shivakumar R, Bond Laura, Turner Matthew W, et al. 9S1R nullomer peptide induces mitochondrial pathology, metabolic suppression, and enhanced immune cell infiltration, in triple-negative breast cancer mouse model. Biomed Pharmacother = Biomedecine Pharmacother 2024;170((January):115997.
[5] Aran Dvir, Toperoff Gidon, Rosenberg Michael, Hellman Asaf. Replication timing-related and gene body-specific methylation of active human genes. Hum Mol Genet 2011;20(4):670–80.
[6] Bennett EAndrew, Keller Heiko, Mills Ryan E, Schmidt Steffen, Moran John V, Weichenrieder Oliver, et al. Active Alu retrotransposons in the human genome. Genome Res 2008;18(12):1875–83.
[7] Chantzi, Nikol, Ioannis Mouratidis, Manvita Mareboina, Maxwell A. Konnaris, Austin Montgomery, and Ilias Georgakopoulos-Soares. 2023. "The Determinants of the Rarity of Nucleic and Peptide Short Sequences in Nature." bioRxiv. https://doi.org/10.1101/2023.09.24.559219.
[8] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57–74.
[9] ENCODE Project Consortium, Moore Jill E, Purcaro Michael J, Pratt Henry E, Epstein Charles B, Shoresh Noam, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 2020;583(7818):699–710.
[10] Fryxell Karl J, Moon Won-Jong. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol 2005;22(3):650–8.
[11] Galas DJ, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 1978;5(9):3157–70.
[12] Georgakopoulos-Soares Ilias, Morganella Sandro, Jain Naman, Hemberg Martin, Nik-Zainal Serena. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. Genome Res 2018;28(9):1264–71.
[13] Georgakopoulos-Soares, Ilias, Ofer Yizhar-Barnea, Ioannis Mouratidis, Rachael Bradley, Ryder Easterlin, Candace Chan, and , . 2021a. "Leveraging Sequences Missing from the Human Genome to Diagnose Cancer." medRxiv.
[14] Georgakopoulos-Soares Ilias, Yizhar-Barnea Ofer, Mouratidis Ioannis, Hemberg Martin, Ahituv Nadav. Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. Genome Biol 2021;22(1):245.
[15] Goswami Jayita, Davis Michael C, Andersen Tim, Alileche Abdelkrim, Hampikian Greg. Safeguarding forensic DNA reference samples with nullomer barcodes. J Forensic Leg Med 2013;20(5):513–9.
[16] Hampikian Greg, Andersen Tim. Absent sequences: nullomers and primes. Pac Symp Biocomput Pac Symp Biocomput 2007:355–66.
[17] Häsler Julien, Strub Katharina. Alu elements as regulators of gene expression. Nucleic Acids Res 2006;34(19):5491–7.
[18] Karczewski Konrad J, Francioli Laurent C, Tiao Grace, Cummings Beryl B, Alföldi Jessica, Wang Qingbo, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581(7809):434–43.
[19] Koulouras Grigorios, Frith Martin C. Significant non-existence of sequences in genomes and proteomes. Nucleic Acids Res 2021;49(6):3139–55.
[20] Landrum Melissa J, Lee Jennifer M, Benson Mark, Brown Garth, Chao Chen, Chitipiralla Shanmuga, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016;44(D1):D862–8.
[21] Lonsdale John, Thomas Jeffrey, Salvatore Mike, Phillips Rebecca, Lo Edmund, Shad Saboor, et al. The genotype-tissue expression (GTEx) project. Nat Genet 2013;45(6):580–5.
[22] Montgomery, Austin, Georgios Christos Tsiatsianis, Ioannis Mouratidis, Candace S. Y. Chan, Maria Athanasiou, Anastasios D. Papanastasiou, , et al. 2023. "Utilizing Nullomers in Cell-Free RNA for Early Cancer Detection." medRxiv. https://doi.org/10.1101/2023.06.10.23291228.
[23] Morganella Sandro, Alexandrov Ludmil B, Glodzik Dominik, Zou Xueqing, Davies Helen, Staaf Johan, et al. The topography of mutational processes in breast cancer genomes. Nat Commun 2016;7(May):11383.
[24] Mouratidis, Ioannis, Fotis A.Baltoumas, Nikol Chantzi, Candace S.Y. Chan, Austin Montgomery, Maxwell A. Konnaris, , et al. 2023a. "kmerDB: A Database Encompassing the Set of Genomic and Proteomic Sequence Information for Each Species." bioRxiv. https://doi.org/10.1101/2023.11.13.566926.
[25] Mouratidis Ioannis, Chan Candace SY, Chantzi Nikol, Tsiatsianis Georgios Christos, Hemberg Martin, Ahituv Nadav, et al. Quasi-prime peptides: identification of the shortest peptide sequences unique to a species. NAR Genom Bioinforma 2023;5(2):lqad039.
[26] Mouratidis, Ioannis, Maxwell A.Konnaris, Nikol Chantzi, Candace S.Y. Chan, Austin Montgomery, Fotis A. Baltoumas, , et al. 2023c. "Nucleic Quasi-Primes: Identification of the Shortest Unique Oligonucleotide Sequences in a Species." bioRxiv. https://doi.org/10.1101/2023.12.12.571240.
[27] Mugal Carina F, Ellegren Hans. Substitution rate variation at human CpG Sites correlates with non-CpG divergence, methylation level and GC content. Genome Biol 2011;12(6):R58.
[28] Patel Ami, Dong Jessica C, Trost Brett, Richardson Jason S, Tohme Sarah, Babiuk Shawn, et al. Pentamers not found in the universal proteome can enhance antigen specific immune responses and adjuvant vaccines. PLoS One 2012;7(8):e43802.
[29] Pratas Diogo, Silva Jorge M. Persistent minimal sequences of SARS-CoV-2. Bioinformatics 2021;36(21):5129–32.
[30] Puig Rafael Riudavets, Boddie Paul, Khan Aziz, Castro-Mondragon Jaime Abraham, Mathelier Anthony. UniBind: maps of high-confidence direct TF-DNA interactions across nine species. BMC Genom 2021;22(1):482.
[31] Rentzsch Philipp, Witten Daniela, Cooper Gregory M, Shendure Jay, Kircher Martin. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47(D1):D886–94.
[32] Sasaki Shin, Mello Cecilia C, Shimada Atsuko, Nakatani Yoichiro, Hashimoto Shin-Ichi, Ogawa Masako, et al. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. Science 2009;323(5912):401–4.
[33] Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 1999;9(8):677–9.

[34] Silva Raquel M, Pratas Diogo, Castro Luísa, Pinho Armando J, Ferreira Paulo JSG. Three minimal sequences found in ebola virus genomes and absent from human DNA. Bioinformatics 2015;31(15):2421–5.
[35] Stamatoyannopoulos John A, Adzhubei Ivan, Thurman Robert E, Kryukov Gregory V, Mirkin Sergei M, Sunyaev Shamil R. Human mutation rate associated with DNA replication timing. Nat Genet 2009;41(4):393–5.
[36] Suzuki Masako, Oda Mayumi, Ramos María-Paz, Pascual Marién, Lau Kevin, Stasiek Edyta, et al. Late-replicating heterochromatin is characterized by decreased cytosine methylation in the human genome. Genome Res 2011;21(11):1833–40.
[37] Sved J, Bird A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci 1990. https://doi.org/10.1073/pnas.87.12.4692.
[38] Tsiatsianis Georgios Christos, Chan Candace SY, Mouratidis Ioannis, Chantzi Nikol, Tsiatsiani Anna Maria, Yee Nelson S, et al. Peptide absent sequences emerging in human cancers. Eur J Cancer 2024;196(January):113421.
[39] Vergni Davide, Santoni Daniele. Nullomers and high order nullomers in genomic sequences. PLoS One 2016;11(12):e0164540.
[40] Vierstra Jeff, Lazar John, Sandstrom Richard, Halow Jessica, Lee Kristen, Bates Daniel, et al. Global reference mapping of human transcription factor footprints. Nature 2020;583(7818):729–36.