GBE

Evolutionary Balancing of Genetic Consequence and Innovation in Mammals Through Variable Number Tandem Repeats

Petar Pajic [1,*], Omer Gokcumen [2,*]

[1]Department of Chemistry, Yale University, New Haven, CT 06511, USA

[2]Department of Biological Sciences, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA

*Corresponding authors: E-mails: petar.pajic@yale.edu; omergokc@buffalo.edu.

Accepted: December 17, 2025

Abstract

Understanding genomic function has historically relied on sequence conservation across evolutionary time. However, advances in genomics have revealed that functional innovations often arise from rapidly evolving, nonconserved elements that are frequently overlooked by conservation-based approaches. Among these, variable number tandem repeats (VNTRs) act as engines of both functional innovation and phenotypic consequence. VNTRs are repetitive genomic sequences whose copy numbers can vary significantly between individuals and species, influencing gene regulation, protein structure, and eventually, phenotypic diversity. Recent long-read assemblies and pangenomes now resolve VNTR loci accurately, enabling robust evolutionary reconstruction and functional associations. Here, we synthesize emerging insights into the functional and evolutionary impact of VNTRs in mammals. Specifically, we outline pressing questions on the mutational mechanisms driving VNTR evolution in humans, the selective forces maintaining their structural heterogeneity, and propose a theoretical framework for their persistence through evolutionary tradeoffs.

Key words: tradeoffs, mucins, function, selection, disease susceptibility.

Significance Statement

Variable number tandem repeats are highly mutable regions of the genome that have remained largely hidden due to their repetitive structure and the limitations of earlier sequencing technologies. Recent advances in long-read genomics now reveal that these sequences can shape gene regulation, protein function, and even the emergence of new biological traits. This review brings together growing evidence that such repeats are key contributors to genetic novelty and play a central role in balancing consequence and innovation throughout evolution.

Introduction

Since the sequencing of the human genome (Lander et al. 2001; Venter et al. 2001), a central focus in evolutionary genetics has been on function. Traditional definitions of functional genomic elements have centered on evolutionary conservation, under the assumption that sequences critical to fitness are preserved by purifying selection across species. This view suggests that only a small fraction (~5% to 7%) (Ponting and Hardison 2011) of the mammalian genome is functionally important. In contrast, advances through transcriptomic and epigenomic methods have offered a broader perspective, arguing that a much larger portion of the genome (up to 80% (ENCODE Project Consortium 2012)) may be functionally relevant. Alternative frameworks have also emerged, suggesting a "twilight zone" of functional evolution, where genomic

regions may acquire functionality through variation, gene turnover, or lineage-specific innovation (Ponting 2017).

Adding to this debate, recent long-read sequencing technologies have uncovered previously inaccessible regions of the genome. These advances, led by consortia such as the Vertebrate Genomes Project (Rhie et al. 2021), the Human Genome Structural Variation Consortium (Ebert et al. 2021), and the Human Pangenome Reference Consortium (Liao et al. 2023), have enabled near telomere-to-telomere (T2T) resolution of genomes across and within species (O'Donnell et al. 2023; Wu et al. 2024; Yoo et al. 2025; Zhang et al. 2025). T2T human assemblies have revealed nearly 200 megabases (Mb) of novel sequence, particularly within pericentromeric and subtelomeric regions (Nurk et al. 2022; Vollger et al. 2022; Logsdon et al. 2025). Many of these newly haplotype-resolved regions are shaped by structural variants (SVs) and are enriched for highly repetitive elements, including centromeres (Logsdon et al. 2025), sex chromosomes (Hallast et al. 2023; Rhie et al. 2023; Makova et al. 2024), and complex loci associated with disease (Olson et al. 2023; Yilmaz et al. 2023).

In this emerging era of T2T assemblies, SVs have gained increasing recognition for their significant contributions to evolution, species diversity, and genome function (Chaisson et al. 2015; Huddleston et al. 2017 ; Pollard et al. 2018). Traditionally, SVs have included chromosomal rearrangements such as insertions, inversions, and translocations, as well as copy number variations like duplications and deletions (Conrad et al. 2010; Mills et al. 2011; Ho et al. 2020; Vollger et al. 2022; Aqil et al. 2023). Fine-scale exploration into SVs has led to several works that show their dynamic evolution and functional implications (Gökçümen and Lee 2009 ; Pajic et al. 2016 , 2019; Karageorgiou et al. 2024; Yılmaz et al. 2024; Pajic et al. 2025; Scheer et al. 2025).

Most recently, attention has turned to a new frontier in understanding genome variation, centered on variable number tandem repeats (VNTRs) (Chaisson et al. 2023; Lu et al. 2023; Cui et al. 2024; Readman et al. 2024; Zhang et al. 2024; Ziaei Jam et al. 2024; English et al. 2025; Schloissnig et al. 2025). VNTRs are DNA segments composed of tandemly repeated motifs (historically defined as >6 base pairs per motif [shorter motifs are termed Short Tandem Repeats (STRs)]), where the copy number of the repeating units varies between individuals or species (Fig. 1a). With the application of improved sequencing technologies to assemble VNTR loci and the inclusion of more diverse populations by major human genome consortia, the detection and the known extent of VNTR heterogeneity have increased more than tenfold in the last decade alone (Fig. 1b).

This dramatic increase in VNTR detection is primarily a technical feat. Traditional short-read sequencing has fundamental limitations: most VNTR tracts exceed typical Illumina read lengths (Fig. 1c), leading to collapsed assemblies and widespread allele dropout (Treangen and Salzberg 2012). Even comprehensive resources such as the 1KGP Phase 1 and Phase 3 releases contain only sparse and biased documentation of VNTRs (Fig. 1b). Read-depth–based strategies and dedicated algorithms like VNTRseek (Gelfand et al. 2014) could recover only a minority of loci, systematically undercalling medium- and longer-length VNTRs. Thus, short-read-based catalogs captured only a small and highly biased subset of true VNTR diversity.

Recent advances in PacBio HiFi long-read sequencing have greatly improved VNTR resolution by providing longer, high-accuracy reads (Javadzadeh et al. 2025). In addition, emerging ultra-long-read platforms can span the largest and most complex VNTR arrays in their entirety, together with their flanking sequence, enabling accurate copy number estimation, sequence-level reconstruction, and haplotype phasing (Jain et al. 2018) (Fig. 1c). These technologies have dramatically expanded the known landscape of VNTR diversity, leading to the discovery of new rare and common disease-associated VNTRs (Cui et al. 2024; Dolzhenko et al. 2024; English et al. 2025). Nevertheless, defining a truly comprehensive tandem repeat catalog remains a substantial computational challenge, and recent work aggregating multiple detection approaches shows that each method contributes unique loci to the combined callset (Weisburd et al. 2025).

VNTRs are becoming increasingly focal in evolutionary genetics, as they are the third most common type of mutation (behind single-nucleotide polymorphisms [SNPs] and InDels) when comparing two human genomes (Collins and Talkowski 2025), and are among the most mutable loci when looking across generations (Porubsky et al. 2025). Specifically, VNTRs are emerging as promising candidates for understanding missing heritability, regulatory complexity, and the genetic basis of biological variation in general (Gymrek and Goren 2021; Ichikawa et al. 2023; Mukamel et al. 2023; Lamkin and Gymrek 2024; Manigbas et al. 2024; Tanudisastro et al. 2025). In this review, we synthesize recent literature highlighting the evolutionary and functional importance of VNTRs in mammals, with a particular emphasis on their implications in humans. We explore how these repeat regions that were largely undetectable in the earlier genomics era are increasingly recognized as dynamic elements that drive disease susceptibility, phenotypic diversity, adaptation, and genomic innovation by affecting exons, regulatory regions, and even giving rise to de novo functional elements. We argue that incorporating VNTRs challenges conventional conservation-based views of function and reveals a broader, more nuanced understanding of genomic variation and opportunities to uncover missing heritability.

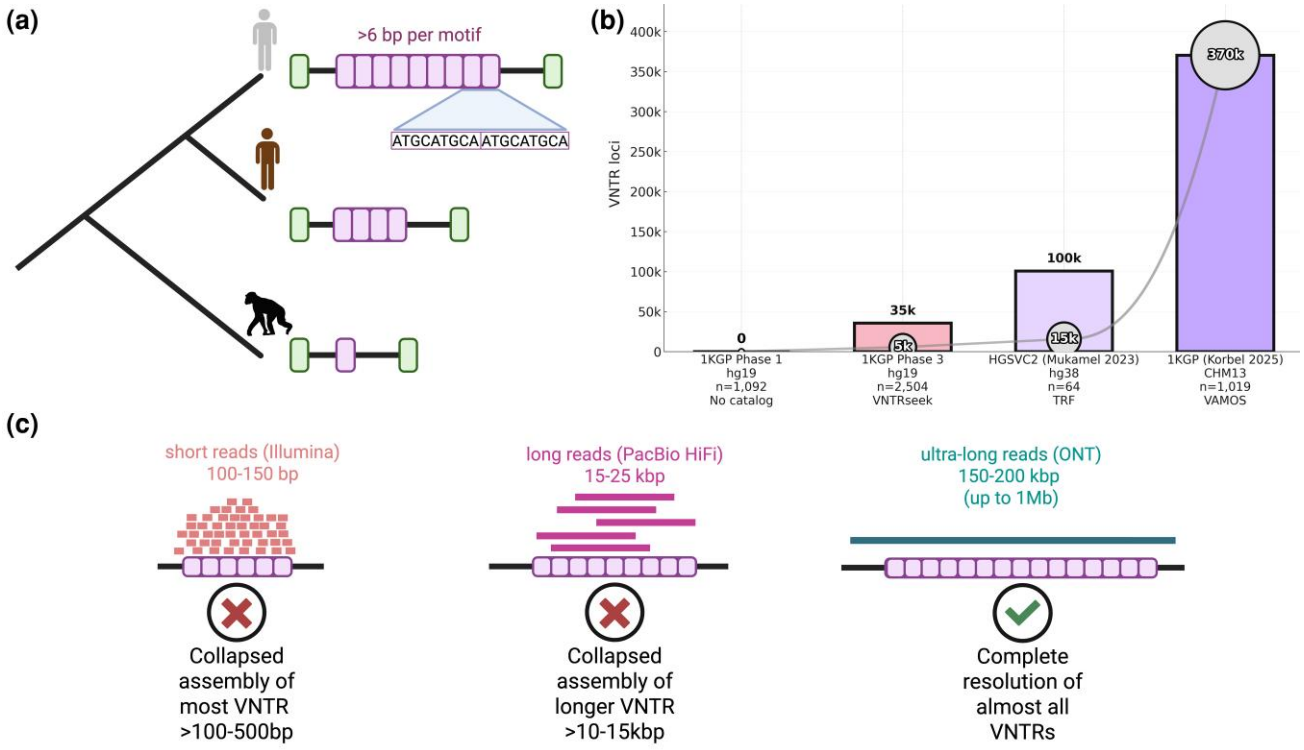

Fig. 1. VNTR loci across major human genome sequencing efforts. a) Cladogram depicts two human individuals from different populations and a chimpanzee (silhouettes). Gene models to the right illustrate the emergence of a human-specific VNTR (magenta rectangles) within an exon. The inset shows two representative repeat units (>6 bp each), highlighting the motif structure that defines VNTRs. b) Bar plots show the total number of VNTR loci identified across major sequencing or analysis projects, with bars indicating total loci detected and bubbles denoting the subset of loci found to be commonly polymorphic. X-axis labels list the sequencing project, reference genome, human sample size, and method used. The left-most gray bar represents the negligible VNTR catalog from the 1000 Genomes Project (1KGP) Phase 1 Pilot ($n = 1,092$; hg19), which did not systematically characterize VNTRs (1000 Genomes Project Consortium et al. 2012). The following pink bar corresponds to 1KGP Phase 3 ($n = 2,504$; hg19, VNTRseek), where 35,638 VNTR loci were identified, 5,676 of which were commonly polymorphic (Eslami Rasekh et al. 2021). The lavender bar shows results from the Human Genome Structural Variation Consortia 2nd release: "HGSVC2" ($n = 64$; hg38, TRF), where 100,844 total loci and 15,653 commonly polymorphic VNTRs were reported (Mukamel et al. 2023). The right-most purple bar represents 1KGP sequenced with Oxford Nanopore and genotyped with VAMOS (Ren et al. 2023; Gu and Chaisson 2024) using the CHM13 reference genome ($n = 1,019$; 1KGP), revealing 370,468 total VNTR polymorphic loci (Schloissnig et al. 2025). c) The left image depicts short-read-based Illumina sequencing (light red rectangles; 100 to 150 bp), which cannot accurately map to the VNTR (pink array), causing collapsed assemblies for nearly all VNTRs. The middle image depicts PacBio HiFi long reads (magenta rectangles; 15 to 25 kbp), which span many but not all arrays, leading to collapsed or partially resolved assemblies for larger and more complex VNTRs. In contrast, the right image shows that ultra-long Oxford Nanopore reads (green rectangle; 150 to 200 kbp, up to ~1 Mb) span entire repeat regions including the flaking regions (black line), enabling complete resolution of almost all VNTRs.

Variable Number Tandem Repeats in Function

Exonic Variable Number Tandem Repeats

Approximately one-third of mammalian proteins harbor repetitive, predominantly tandem segments derived from VNTRs within their coding exons (Schaper et al. 2014). While most of these repeats are conserved in sequence and in copy number, a considerable portion (~3% to 5% (Tanudisastro et al. 2025)) exhibits variation in copy number between individuals or species (Nakamura et al. 1998), hereafter referred to as exonic variable number tandem repeats (exVNTRs). Given that exVNTRs directly affect protein sequence and structure, they are highly relevant to understanding biological variation, gene function, and human health. However, despite their prevalence and direct functional implications, exVNTRs have been challenging to characterize due to the limitations of short-read sequencing. In light of advances in long-read sequencing, which can, in most cases, span the entirety of repetitive loci, accurate investigations into the evolution and function of these regions are now feasible.

exVNTRs, nested within coding exons, can directly modulate gene function by altering the encoded protein's length, domain composition, and biochemical properties (Doege et al. 1997; Gemayel et al. 2010, 2012; Marshall et al. 2021). The evolutionary relevance of these structural changes has been demonstrated through their effects on

post-translational modifications, such as the availability of glycosylation sites in mucins (Higuchi et al. 2002; Reily et al. 2019; Plender et al. 2024) and on protein-protein interactions, as seen in filaggrin-keratin interface (Brown et al. 2012; Mac Donagh et al. 2024) (Fig. 2a). A critical feature observed in functional exVNTRs is that their repeat unit length retains an open reading frame (multiples of three nucleotides) resulting in tandemly repeated amino acid motifs that may be preserved under adaptive evolution (Sabino et al. 2014; Eaaswarkhanth et al. 2016; Xu et al. 2016; Pajic et al. 2022) (Fig. 2a). In contrast, frame-shifting mutations in repeats not divisible by three can truncate or dramatically alter the downstream protein sequence, potentially causing loss-of-function effects and can contribute to disease pathology (Kirby et al. 2013; Wenzel et al. 2018). Given that VNTRs can evolve rapidly, frame-shifts and their subsequent purging by selection may be relatively common. The full extent to which such specific mutations in genes with exVNTRs contribute to human disease or novel functions remains an open question, though insights from locus-specific studies are emerging.

exVNTRs Among Species

Several examples illustrate the functional influence of exVNTRs that contribute to trait variation across mammals. In domestic dogs, variation in polyglutamine/polyalanine encoding tandem repeats of the *RUNX2* gene, a key regulator of osteogenesis, is strongly associated with craniofacial diversity and limb proportions. Specifically, the length of the glutamine-alanine VNTR in the transactivation domain of RUNX2 correlates with short-faced (brachycephalic) versus long-faced (dolichocephalic) breeds, acting as a molecular "tuning knob" for skeletal development (Fondon and Garner 2004). In mice, *PRDM9* contains a rapidly evolving zinc finger exVNTR encoded within a single exon. The number and sequence of zinc finger repeats determine DNA-binding specificity, directing the location of meiotic recombination hotspots (Baudat et al. 2010; Altemose et al. 2017). In hybrids between *Mus musculus* subspecies, incompatibilities in PRDM9 binding caused by divergent repeat profiles contribute to meiotic arrest and hybrid sterility, a striking example of an exVNTR influencing reproductive isolation and speciation (Oliver et al. 2009; Parvanov et al. 2010).

In primates, the DUF1220 VNTR domain in the *NBPF* gene varies among nonhuman primates and is expanded in humans, increasing repeat copy number and encoded protein length. These expansions are hypothesized to contribute to increased brain size and neurodevelopmental complexity (Dumas et al. 2012). Long-read sequencing comparisons across primates have revealed over 1,500 human-specific tandem repeat expansions, 1% to 2% of which are exonic (Sulovari et al. 2019). As such, exVNTRs stand out for their capacity to drive phenotypic effects that can underlie species-specific traits (Course et al. 2021; Mukamel et al. 2021; Chaisson et al. 2023).

exVNTRs Within Species

More recently, exVNTRs have been shown to vary substantially when comparing human individuals and are implicated in a wide range of complex trait associations (Linthorst et al. 2020; Bakhtiari et al. 2021; Mukamel et al. 2021; Lamkin and Gymrek 2024). Mukamel et al. systematically identified 118 protein-coding VNTRs with exceptionally large effect sizes by imputing copy numbers across more than 400,000 UK Biobank individuals (Mukamel et al. 2021; Di Maio et al. 2024). These exVNTRs were found to strongly influence phenotypes such as height, hair texture, kidney function, and blood lipid levels, often with greater effect sizes than SNPs. For example, the *LPA* gene contains an exVNTR known as Kringle IV type 2 (KIV-2), consisting of tandemly repeated domains encoding kringle motifs in the apolipoprotein(a) protein. Variation in repeat number influences both isoform size and expression, where a lower number of KIV-2 repeats is associated with higher plasma lipoprotein(a) concentrations and increased cardiovascular risk due to more efficient secretion of smaller isoforms (Mukamel et al. 2021; Di Maio et al. 2024).

Many exVNTRs are associated with disease, yet these same loci often exhibit signatures of adaptive or balancing selection, suggesting that repeat variation can simultaneously contribute to pathogenic risk and confer context-dependent evolutionary advantages. In the *FLG* gene, the exVNTR encodes tandem filaggrin subunits that aggregate keratin intermediate filaments to form the cornified envelope of the skin barrier (Dale et al. 1978; Simon et al. 1996). Variation in repeat copy number directly modulates filaggrin dosage and, consequently, epidermal integrity. Shorter or truncated alleles compromise barrier function, increasing susceptibility to atopic dermatitis and ichthyosis vulgaris (Brown et al. 2012) (Fig. 2a). Conversely, maintaining variation in exVNTR length may reflect adaptive trade-offs, where reduced filaggrin expression could enhance evaporative cooling or alter microbiome composition under certain environmental conditions (Angelova-Fischer et al. 2011) (Fig. 2a). *FLG* also shows evidence of a hitchhiking selective sweep at the nearby HRNR locus, suggesting that complex coding repeats can evolve under indirect selection while remaining tightly linked to trait-associated variation (Eaaswarkhanth et al. 2016).

In the *ACAN* gene, a 57-bp exVNTR within the first chondroitin sulfate (CS1) domain strongly influences height, with longer alleles producing taller stature by increasing the number of glycosaminoglycan attachment sites that enhance cartilage hydration and growth-factor binding within the extracellular matrix. The exVNTR alone explains up to 0.6% of height variance in individuals of African ancestry

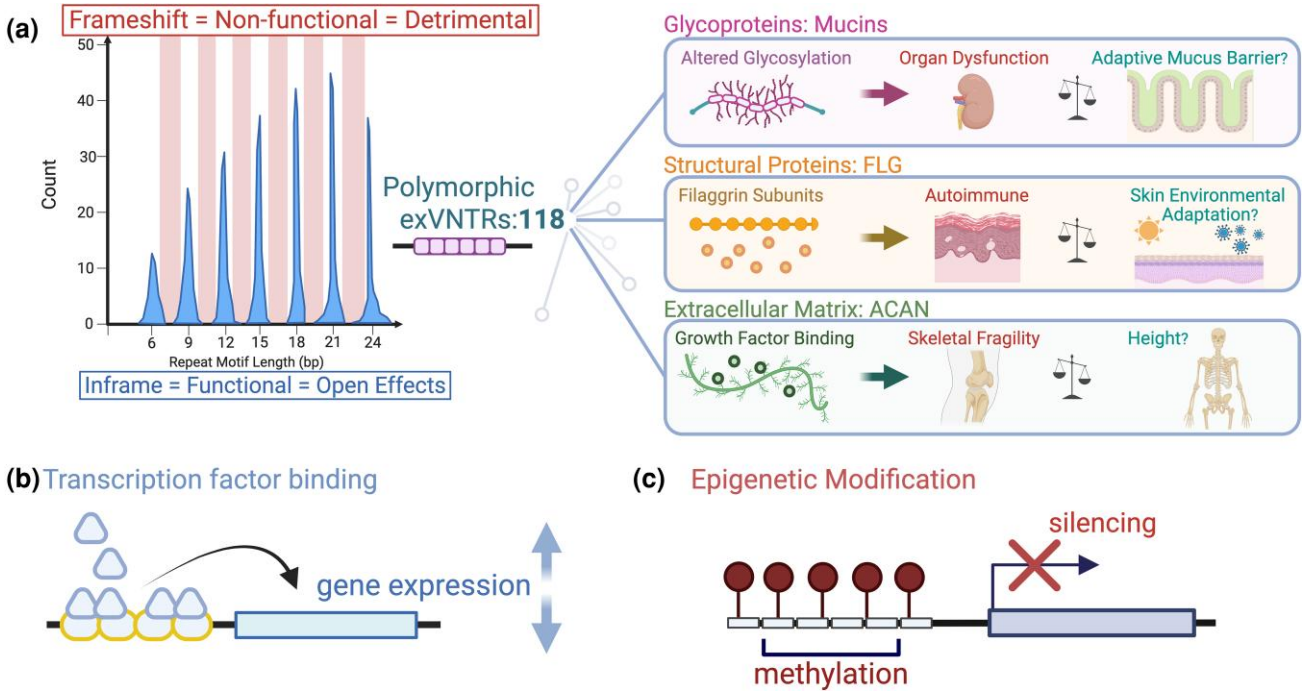

Fig. 2. Function in exonic and noncoding VNTRs. a) Exonic VNTRs frequently occur at repeat unit lengths that preserve the reading frame (multiples of three), enabling "open" effects on protein properties, whereas frameshifted exVNTRs (red shading, non-3n motifs) typically disrupt coding potential and are removed by purifying selection. Among the 118 polymorphic exVNTRs identified with large effect sizes (Mukamel et al. 2021), many are found in glycoproteins such as mucins, where repeat copy number variation alters glycosylation and mucus barrier function in different organs, or in structural proteins like *FLG*, where variation in filaggrin subunit number can affect keratin aggregation, contributing to inflammation in atopic dermatitis or adaptation to environmental stressors. Similarly, in the extracellular matrix protein *ACAN*, exVNTR length modulates the number of chondroitin sulfate attachment sites, influencing growth-factor binding and skeletal properties. These evolutionary tradeoffs are depicted on the right separated by a scale. b) Noncoding VNTRs can influence transcriptional activity by altering the number or spacing of transcription-factor binding sites. c) ncVNTRs also affect epigenetic regulation, where changes in repeat length modify local CpG methylation and lead to allele-specific gene silencing.

and 0.19% in Europeans, with effect sizes exceeding those of nearby SNPs (Mukamel et al. 2021). Beyond stature, shorter *ACAN* VNTR alleles are linked to intervertebral disc degeneration and osteoarthritis, likely due to reduced proteoglycan content and weaker retention of growth-factor gradients (Xu et al. 2012; Cong et al. 2018; Haddadi et al. 2022). Together, these findings suggest a continuum in which VNTR copy number tunes matrix hydration, growth-factor sequestration, and skeletal resilience (Fig. 2a).

Another striking example is the *MUC1* gene, which encodes a heavily glycosylated transmembrane mucin expressed in several epithelial tissues (Dhanisha et al. 2018). Its exVNTR spans exon 2 and consists of 60-bp tandem repeats encoding a 20-amino acid motif rich in serine and threonine, sites for O-glycosylation. Mukamel et al. found that variation in *MUC1* VNTR copy number was associated with urea levels, a marker of kidney function. Higher repeat numbers may increase glycosylation density—altering protein stability, barrier properties, or renal clearance capacity (Fig. 2a). This region is also implicated in autosomal dominant tubulointerstitial kidney disease (ADTKD) through pathogenic frameshifting insertions within the exVNTR that truncate the protein (Kirby et al. 2013; Devuyst et al. 2019). Thus, natural variation in *MUC1* exVNTR length contributes to both normal physiological variation, potentially adaptive barrier function, and disease susceptibility through effects on protein structure and epithelial integrity.

Similarly, the mucin gene, *MUC7*, which harbors an exVNTR rich in serine and threonine residues, shows evidence of local adaptation and retention of glycosylation capacity across primates (Xu et al. 2016). In another mucin, *MUC5AC*, balancing selection has been implicated in maintaining repeat copy number diversity, potentially through heterozygote advantage (Plender et al. 2024). Overall, exVNTRs have emerged as notable contributors to missing heritability and explainable variance in complex traits (Course et al. 2021; Gymrek and Goren 2021) resulting in both detrimental and beneficial effects.

Noncoding VNTRs and Their Effect on Gene Regulation

Though historically overlooked due to their low sequence conservation and position outside of coding regions,

noncoding VNTRs (ncVNTRs) are now increasingly recognized as dynamic regulators of gene expression (Zhang et al. 2024) and genome structure (Gent et al. 2011), contributing to complex phenotypes (Bakhtiari et al. 2021; Marshall et al. 2021; Mukamel et al. 2023). These effects are especially pronounced when ncVNTRs overlap with enhancers and promoters (Bellizzi et al. 2005; Vinces et al. 2009; Gemayel et al. 2012; Quilez et al. 2016; Eslami Rasekh et al. 2021), or reside within introns or untranslated regions (UTRs) (Sulovari et al. 2019; Mukamel et al. 2023). In such contexts, variation in repeat copy number can modulate transcriptional output by affecting transcription factor binding site density (Fig. 2b) (Vasiliou et al. 2012), local methylation patterns (Fig. 2c) (Quilez et al. 2016; Dolzhenko et al. 2024), and the three-dimensional architecture of chromatin (Gent et al. 2011).

Promoters and Enhancers

Recent studies have identified numerous ncVNTRs overlapping regulatory regions, demonstrating widespread impacts on transcriptional regulation and cellular function (Sulovari et al. 2019; Bakhtiari et al. 2021; Tanudisastro et al. 2025). A well-characterized example is the ncVNTR within the promoter of the *MAOA* (monoamine oxidase A) gene, where variation has been linked to behavioral phenotypes and risk for psychiatric conditions (Kunugi et al. 1999). Shorter alleles are associated with reduced transcriptional activity, while longer alleles drive higher expression levels of *MAOA* (Sabol et al. 1998). More recently, a second, more distal ncVNTR upstream of the canonical promoter repeat was identified and shown to regulate *MAOA* mRNA abundance, with both repeats influencing transcript variants in an isoform-specific manner (Manca et al. 2018; Marshall et al. 2021). Moreover, ncVNTR alleles at this locus show differential responsiveness to cellular stimuli, suggesting that gene-by-environment interactions may be mediated through specific repeat haplotypes. This regulatory plasticity is exemplified by studies demonstrating behavioral outcomes in response to early-life adversity: individuals carrying the low-activity (short) allele exhibit increased antisocial behavior particularly under stressful environments, whereas carriers of the high-activity allele are comparatively resilient (Caspi et al. 2002).

Few examples of ncVNTRs affecting enhancer regions have been documented, largely because enhancer landscapes remain incompletely mapped. However, it is plausible that in ncVNTRs where the repeat motif itself constitutes a transcription factor binding site, expansions overlapping with enhancer regions could have a major impact on gene expression by increasing both the number and accessibility of transcription factors bound (Vasiliou et al. 2012) (Fig. 2b). One candidate is a 72-bp VNTR located in intron 5 of the *SIRT3* gene, where both repeat copy number and internal sequence variation modulate allele-specific enhancer activity (Bellizzi et al. 2005). Longer alleles drive higher reporter gene expression, while a single nucleotide substitution within the repeat converts a GATA3 binding site into a DeltaEF1 site, abolishing enhancer function (Bellizzi et al. 2005). This highlights how both the copy number and sequence integrity of ncVNTRs can be critical for regulatory activity.

Introns and UTRs

Recent work has identified intronic ncVNTRs as potent regulators of gene function and disease risk (Mukamel et al. 2023). A striking example is a ncVNTR within an intron of *TMCO1*, where expanded repeat alleles are strongly associated with elevated intraocular pressure and increased risk of primary open-angle glaucoma. This repeat explains more phenotypic variance than surrounding SNPs and likely represents the causal variant underlying previously observed GWAS signals. Similarly, an intronic ncVNTR in *CUL4A* influences multiple red blood cell traits, particularly mean corpuscular hemoglobin. Beyond trait associations, this repeat also modulates splicing of *CUL4A* across multiple tissues, suggesting a direct mechanistic role in transcript regulation (Mukamel et al. 2023). Another example involves the *IL4* gene, where intron repeat variation may influence cytokine regulation (Duan et al. 2014). The shorter ncVNTR has been associated with reduced *IL4* expression and increased susceptibility to tuberculosis (Kulpraneet et al. 2019), whereas the longer allele enhances *IL4* transcription and promotes a Th2-skewed autoimmune response, advantageous in parasite-rich environments (Gyan et al. 2004; Jha et al. 2012). Together, these patterns suggest the possibility of pathogen-driven selection maintaining ncVNTR diversity to fine-tune immune responses under differing infectious pressures.

Several ncVNTRs have also been implicated in neuropsychiatric conditions such as schizophrenia and bipolar disorder (Marshall et al. 2021; Birnbaum 2023; Barlattani et al. 2024). One well-studied example is the dopamine transporter gene *SLC6A3* (DAT1), which harbors a 40-bp ncVNTR in its 3′ untranslated region (MacKenzie and Quinn 1999). The polymorphism commonly exists in 9- or 10-repeat alleles and is associated with differential expression in the striatum and prefrontal cortex, regions of the brain critical for executive function and reward processing (van Dyck et al. 2005). As such, variation in ncVNTR length has been linked to behavioral traits such as impulsivity, attention-deficit/hyperactivity disorder (ADHD), and substance use disorders (Bédard et al. 2010; Ma et al. 2016). Functional studies indicate that the 10-repeat allele generally enhances DAT1 expression and transporter function, resulting in greater dopamine reuptake and reduced synaptic dopamine, though the tissue specificity of this effect

varies across studies (Mill et al. 2005 ; Vasiliou et al. 2012; Reith et al. 2022). Moreover, population-genetic analyses have revealed signatures of balancing selection across the *SLC6A3* locus (Kelada et al. 2006), suggesting long-term maintenance of both 9- and 10-repeat alleles, which may reflect adaptive tuning of dopaminergic signaling and behavioral flexibility across variable social and environmental contexts. Collectively, such findings support the hypothesis that VNTR-mediated regulatory variation contributes to cognitive traits and the evolutionary trajectory of the human brain (Sulovari et al. 2019; Course et al. 2021).

Epigenetic Modifications

In addition to transcriptional regulation, ncVNTRs are increasingly recognized as important mediators of epigenetics, as seen in methylation modifications. Variation in repeat copy number can significantly influence the local DNA methylation landscape, especially when ncVNTRs are positioned within CpG islands (Fig. 2c) or regions bound by histone modifiers (Dolzhenko et al. 2024). Recent high-resolution methylation profiling has begun to reveal that repeat length-dependent methylation patterns are common across the genome, particularly at disease-associated ncVNTRs (Dolzhenko et al. 2024). For example, promoter-proximal ncVNTRs can directly influence local CpG methylation and gene activity, as shown for the *EIF3H* ncVNTR, where length variation is linked to differential methylation and altered transcription (Quilez et al. 2016; Mukamel et al. 2023). These methylation changes are not transient but can be stably maintained across developmental stages and across cell types, contributing to long-term regulation of gene expression (Dolzhenko et al. 2024).

Taken together, these findings establish ncVNTRs as versatile and mostly underappreciated elements of gene regulation. By modulating transcription factor binding, splicing, and epigenetic marks, ncVNTRs influence gene expression through multiple, often interrelated, mechanisms. While most have been characterized through their deleterious associations, some evidence points to ncVNTRs as drivers of adaptive regulatory plasticity. As regulatory genomics, like ATAC-seq (Grandi et al. 2022) and FIBER-seq (Stergachis et al. 2020) continue to improve regulatory element mapping, additional examples of ncVNTRs influencing enhancer activity are likely to emerge.

Mucins as a Model for Evolutionary Innovation

How novel gene functions evolve is a fundamental question in biology. The recent discovery of thousands of non-traditional open reading frames that encode functional proteins has significantly broadened our understanding of genomic plasticity (Deutsch et al. 2024). Researchers have begun to identify novel species-specific genes using bioinformatic methods to detect new routes to adaptive novelty (Luis Villanueva-Cañas et al. 2017) in noncoding regions across eukaryotes (Tautz and Domazet-Lošo 2011; Ruiz-Orera et al. 2014; Van Oss and Carvunis 2019; Weisman 2022; Zhao et al. 2024; Xia et al. 2025). Mobile genetic elements and SVs are well-known drivers of this process (Gemayel et al. 2010; Ji and Salzberg 2024); however, VNTRs are emerging as powerful sources of functional novelty.

Beyond impacting the function of existing genes, VNTRs can contribute to the birth of entirely new functions. A prime example is antifreeze glycoproteins (AFGPs), which evolved independently from noncoding regions in Arctic cod and Antarctic notothenioids to inhibit ice crystal formation in sub-zero waters (Chen et al. 1997b, 1997a; Baalsrud et al. 2018). In Antarctic notothenioids, AFGPs originated de novo from a trypsinogen-like pseudogene, where frameshift mutations and tandem expansions of a Thr-Ala-Ala coding motif created a novel open reading frame. In contrast, Arctic cod evolved AFGPs from a duplicated trypsinogen gene that lost its ancestral function and gained antifreeze activity through similar, albeit a different type of, repeat expansions. In both cases, antifreeze function was gained through VNTRs that serve as O-glycosylation sites, enhancing ice-binding capacity.

A more recent example of genomic repurposing is in the evolution of mucins (Pajic 2025). Mucin genes encode glycoproteins that form a critical component of mucus, providing barrier function through lubrication on organs, and defense from pathogens on epithelial surfaces (Wagner et al. 2018). They exhibit remarkable structural diversity, particularly through the presence of exVNTRs that are densely glycosylated on the encoded mucin protein (Dekker et al. 2002; Lang et al. 2007). Although some mucins have arisen through tandem gene duplications (Desseyn et al. 2000), most have arisen independently rather than through identity-by-descent, and are referred to as "orphans" (Dekker et al. 2002; Lang et al. 2007). Evidence suggests that mucin genes have undergone convergent evolution across different mammalian lineages tailored to specific physiological demands (Dekker et al. 2002; Lang et al. 2007; Pajic et al. 2022).

Paji? et al. explored the dynamic evolution of mucin genes across mammals (Pajic et al. 2022). By analyzing whole-genome assemblies from diverse species, it was demonstrated that mucin genes can evolve de novo by acquiring VNTRs encoding (P) proline-, (T) threonine-, and (S) serine-rich motifs, characteristic of mucin functional domains. Astonishingly, non-VNTR genes encoding proline-rich proteins (PRPs) recurrently evolve into mucins, establishing a new model where PRP genes serve as natural precursors, as their proline content is one mutational step from T/S peptides that enable glycosylation (Fig. 3). Furthermore, PRPs may have vestigial functions or evolve under relaxed selection, making them more permissive to evolutionary

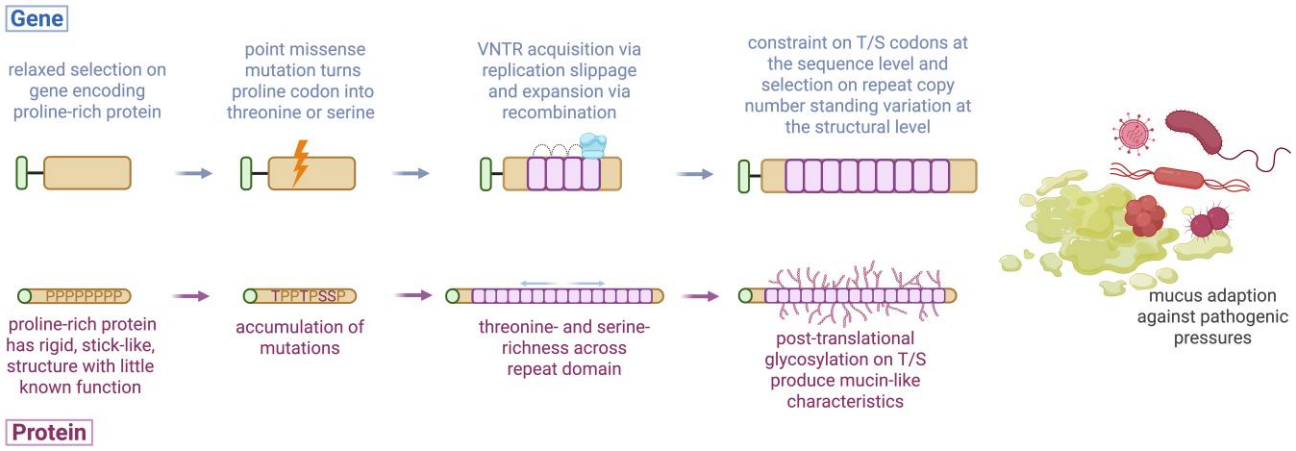

Fig. 3. Evolutionary "mucinization" of proline-rich proteins into functional mucins. Schematic showing how proline-rich proteins can be repurposed into mucins. Top: genetic-level events. A proline-rich gene under relaxed selection (tan, with upstream exon/signal peptide in green) acquires point missense mutations (lightning bolt) converting some proline codons into threonine (T) or serine (S) codons. Replication slippage and recombination generate and expand VNTR in-frame arrays encoding T/S-rich motifs (pink). Sequence constraint maintains T/S-rich functional codons, while standing copy number variation becomes a target of selection in certain environments. Bottom: protein-level consequences. Proline-rich proteins with rigid, low-complexity domains (tan) accumulate T/S amino acids, expand into repeat-TS-rich regions, and undergo dense post-translational O-glycosylation, yielding mucin-like biophysical properties. This transition generates immediate functional heterogeneity, and once a new mucin is recruited, repeat copy number polymorphism arises as a natural consequence, opening functional flexibility that can be acted upon by natural selection, for example, through adaptive modulation of mucus properties in response to pathogenic pressures.

innovation. These VNTR expansions were found to be recurrent and evolving rapidly in the saliva across multiple species, likely shaped by selective pressures related to host-pathogen interactions and mucosal defense involving glycans in the oral cavity (Cross and Ruhl 2018; Wagner et al. 2018; Barnard et al. 2020; Yang et al. 2025).

The de novo birth of mucin genes and the extensive copy number variation within their VNTR domains offer valuable insights into the mechanisms of functional evolution. This variation has immediate consequences for protein size, glycosylation potential, and interactions with the extracellular environment, features that directly impact mucosal biology. As such, mucins serve as powerful models for investigating how VNTRs contribute to phenotypic diversity and adaptive evolution (Gemayel et al. 2010, 2012). Together, these findings reveal how VNTRs not only modify existing genes but also catalyze the emergence of entirely new ones, raising a central question of how such mutable yet functional elements persist through evolutionary time.

Conclusions and Future Perspectives on Evolutionary Maintenance of VNTR Variation

As we have discussed, the apparent lack of conservation across VNTRs reflects an ongoing evolutionary interplay of both phenotypic consequence and functional innovation, often involving the same genes (Fig. 2a). Different parts of the genome experience distinct selection regimes, and VNTRs reflect this at different levels. In regulatory regions, copy number may be selected when altering the density of transcription-factor binding sites confers advantageous expression changes, whereas in coding domains, purifying selection constrains sequence integrity across motifs to preserve biochemical function (Pajic et al. 2022). Interruptions or degenerative repeats, which can modulate mutability by disrupting expansion mechanisms such as replication slippage or nonallelic homologous recombination (NAHR), may also be selectively favored (Sulovari et al. 2019). Selection can therefore act not only on repeat copy number but also on the degree of sequence homogeneity among repeat units. The interplay between these pressures shapes VNTR heterogeneity at multiple levels, framing how different evolutionary forces maintain or erode repeat variation across the genome.

While some VNTRs are under selective constraint, most evolve under **Neutrality.** Neutral regions subject to drift or relaxed purifying selection accumulate mutations without major fitness costs, leading to increased divergence among individuals and species (Hunt et al. 2011). A non-VNTR example is the reduction of olfactory receptor genes in humans relative to other primates, likely reflecting relaxed selection accompanying a decreased reliance on smell as vision became dominant (Gilad et al. 2003; Veilleux et al. 2023). Similarly, neutral alleles can persist if their deleterious effects emerge only after reproductive age, as in late-onset Huntington's disease (Ross and Tabrizi 2011; Handsaker et al. 2025). Such relaxation may also foster evolutionary innovation, as neutral or mildly deleterious mutations occasionally acquire adaptive value

and become targets of positive selection (Wang et al. 2017; Zhao et al. 2023). It stands to reason that the majority of the variation seen in VNTRs may have little functional impact as they occur outside of genes, and thus evolve largely under neutrality. Yet a central question remains: why do VNTRs with clear molecular, cellular, or organismal effects, remain so variable, and which evolutionary forces maintain this diversity?

Positive Selection

Promotes the rapid diversification of sequences that confer adaptive benefits, leading to lineage-specific functional innovations. Antimicrobial peptides such as defensins exemplify this pattern, evolving rapidly in response to species-specific pathogen pressures (Ganz 2003). Another classic example in humans is lactase persistence, where selection on SNPs near the *LCT* gene enabled continued lactase expression into adulthood, aligning genetic change with the cultural adoption of dairying in different geographies (Tishkoff et al. 2007; Kerdoncuff et al. 2025). Other work has discussed inconsistencies in the lactase story (Ségurel and Bon 2017), underscoring that detecting positive selection is difficult, even for SNPs, for which the majority of modern population genetics tools are designed.

Most common human SNPs predate the emergence of anatomically modern humans, reflecting deep coalescent times and a low mutation rate ($\sim 10^{-8}$ per bp per generation). As a result, SNPs rarely introduce new adaptive alleles on timescales relevant to rapid Holocene environmental change (Hernandez et al. 2011; Scally and Durbin 2012). In contrast, VNTRs mutate orders of magnitude faster ($\sim 10^{-5}$ to 10^{-2} per locus per generation; (Porubsky et al. 2025), continually generating novel alleles and maintaining abundant standing variation. Because VNTR alleles can behave neutrally when selection is weak, substantial repeat copy number and sequence diversity can accumulate without major fitness costs, forming a reservoir of alleles that can be co-opted when strong ecological or cultural pressures arise. These properties make VNTRs a more plausible substrate for recent, rapid human adaptation than SNPs and help explain why VNTR loci show population-specific expansions, contractions, and functional effects (Hannan 2018; Sulovari et al. 2019; Mukamel et al. 2021). Yet, despite this evolutionary potential, direct evidence for positive selection on VNTRs remains limited, partly because tools to robustly analyze selection in such loci are not available.

Detecting selection on VNTRs remains challenging, as classical scans rely on biallelic SNPs and therefore miss the multiallelic, rapidly mutating nature of VNTRs. Consequently, locus-specific analyses incorporating haplotype-level information remain the most reliable approach. A recent study of *MUC19* found striking differences in VNTR copy number between global populations and admixed Americans that carry an introgressed archaic haplotype (Villanea et al. 2025). Positive selection was then inferred using SNP-based statistics across the introgressed region, including population-branch statistics (PBS), and demonstrated via demographic simulations that these signals cannot be explained by neutrality. Earlier works on the *DRD4* VNTR used intra-allelic haplotype diversity, linkage disequilibrium (LD) decay, and allele-age estimation to argue for positive selection on the 7-repeat allele (Ding et al. 2002; Wang et al. 2004). These studies have essentially conducted traditional population genetic analyses on SNPs that are on the same haplotype harboring putatively selected VNTR alleles.

Looking forward, the field urgently needs population-genetic tools that can detect selection on VNTRs with confidence. Several gaps are holding us back. First, we need approaches that can robustly infer VNTR copy number variation in both modern and ancient genomes from short-read data. Second, we lack formal tests to evaluate whether the striking absence of frameshift mutations during exVNTR expansion deviates from neutral expectations, and whether this pattern reflects purifying selection preserving the reading frame. Third, LD calculations must be adapted to explicitly model recurrent mutations and homoplasy (van Leeuwen et al. 2015). Finally, we need mutational models that move beyond simple stepwise accumulation and instead incorporate gains and losses of multiple repeat units through recombination events.

Closing these gaps will enable integrated simulation-based and empirical frameworks to evaluate selection on VNTRs. Such approaches will jointly model copy number variation across ancient and present-day populations, reconstruct ancestral states and allele-frequency spectra, quantify temporal shifts in repeat-copy distributions, and detect selection acting both on the motifs/nucleotide content of repeat units and on haplotypes carrying specific VNTR alleles. Together, these advances will allow rigorous tests of classical positive and balancing selection, while also supporting more nuanced models, such as oscillating or frequency-dependent bouts of selection acting on standing VNTR variation in population- and time-specific ways.

Balancing selection and evolutionary tradeoffs

Can sustain genetic variation within populations. Classic examples include the extreme polymorphism of major histocompatibility complex (MHC) genes maintained by pathogen-mediated balancing selection (Hedrick 2006). Recent work shows that even gene-disrupting SVs can persist, such as the growth hormone receptor (*GHR*) exon 3 deletion, which enhances prenatal growth in some contexts but increases risks for metabolic dysfunction in others (Allison 1954; Saitou et al. 2021; Nikkanen et al. 2022; Aqil et al. 2023). These cases illustrate how alleles can be maintained when their advantages and disadvantages are context-dependent. Within this framework, VNTRs emerge

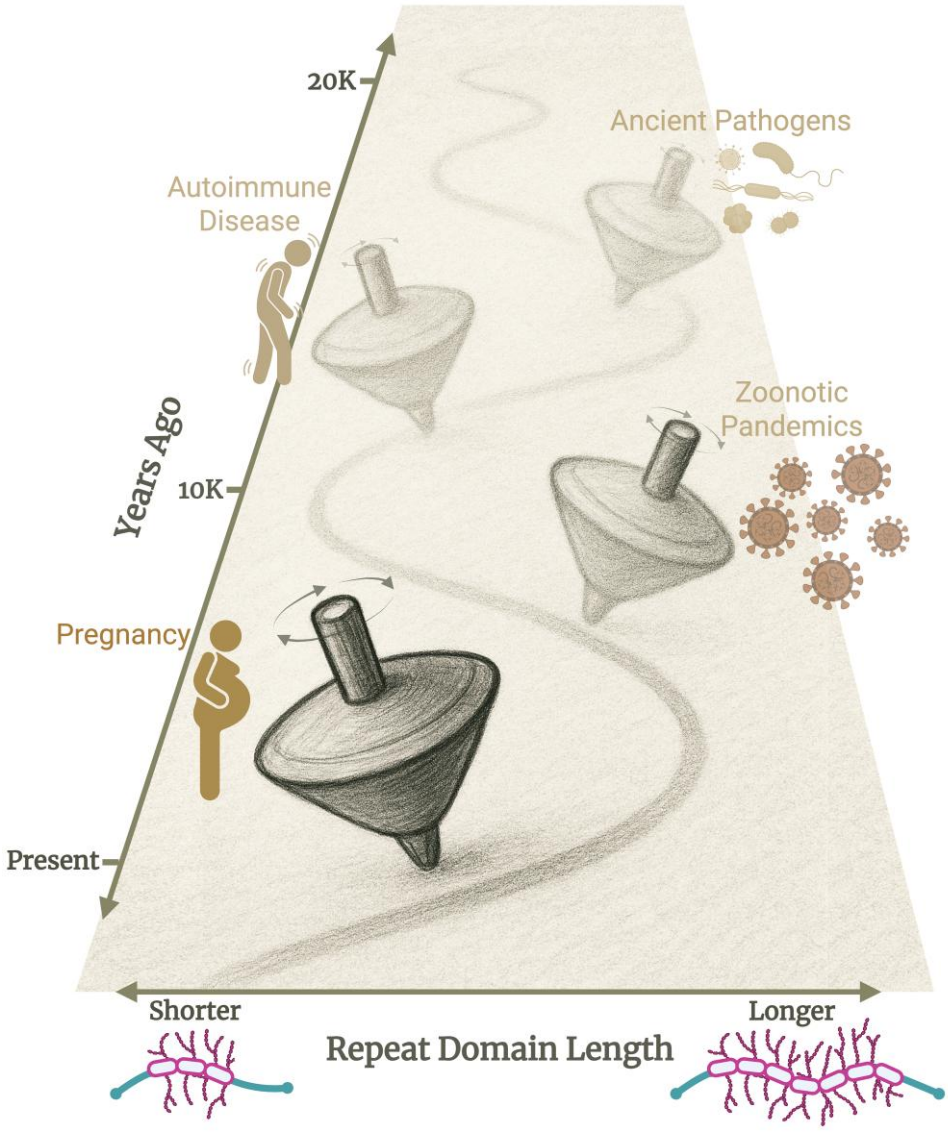

Fig. 4. Spinning-top model of evolutionary tradeoffs. The vertical slanted-axis represents time, with the present at the bottom and the past toward the top. The horizontal-axis denotes VNTR domain length in a mucin, ranging from shorter (left) to longer (right). The spinning top illustrates the dynamic, multidimensional nature of VNTR evolution, oscillating through time and across environments as repeat copy number population allele frequencies shift configurations continuously. Bronze pictograms mark major environmental or physiological transitions (e.g. pathogen exposure, zoonotic immune challenge, pregnancy) that influence the direction and stability of the spinning top's balance. For example, pandemics may drive selection on longer mucin VNTR alleles generating stronger mucus barriers on organs against pathogens, but at the same time, increasing mucosal and immune-related disease susceptibility. Whereas shorter mucin VNTR alleles—driven by reduced autoimmune disease and improved pregnancy success through embryo implantation dependent on the mucus barrier—may be favored under benign exposure contexts.

as particularly informative loci: they are highly mutable, recurrently evolving, and often pleiotropic in function. Their enduring variation likely reflects an inherent capacity to mediate both constraint and innovation, a balance that shapes genomic and phenotypic diversity. Yet, unlike well-characterized loci such as *MHC* or *GHR*, the dual adaptive and deleterious consequences of VNTR variation have rarely been examined in an integrated comprehensive manner across studies.

Evolutionary tradeoffs are often depicted as a simple scale balancing benefit and cost. For pleiotropic VNTRs, however, the balance of effects can shift dramatically across tissues, environments, and historical contexts, and is further complicated by the presence of multiple alleles—each defined by different repeat copy numbers—rather than a simple two-allele system. We envision this using a "spinning-top" moving along a path through time (Fig. 4). The path represents the temporal trajectory

of the allele, while the horizontal position of the top reflects which repeat configuration (shorter or longer) is favored at that moment. The tilt of the top indicates the strength of the selective bias toward that configuration, and the orientation captures which specific pressure (e.g. pathogen exposure, immune activation, pregnancy) is exerting that pull. At any given point, the same VNTR allele may confer a functional benefit in one context while simultaneously imposing a biological consequence in another, such that gain and cost can coexist rather than occur sequentially. As different ecological or physiological forces rise and fall, the top shifts position, tilt, and orientation, illustrating how VNTRs experience a continually reweighted balance of pressures rather than a single, static tradeoff. This shifting, multidimensional equilibrium helps maintain standing VNTR diversity and contributes to their rapid evolutionary trajectories. We hope that this conceptual model provides a blueprint for emerging new perspectives that incorporate the multiallelic nature of VNTRs within a pleiotropic context.

Mucin genes, discussed thoroughly in this review, offer a striking empirical framework for exploring these dynamics, as recurrent emergence and extreme variation (Prodanov et al. 2025) highlight both the instability and adaptive potential of these loci. For example, specific repeat copy numbers have recurrently evolved in the *MUC7* gene in humans and have independently expanded across mammalian lineages, likely in response to pathogenic pressures encountered in the oral cavity (Xu et al. 2016, 2020). Likewise, recent analyses of *MUC5AC* indicate that distinct common copy number alleles are maintained by balancing selection in some human populations (Plender et al. 2024).

Immune-related genes have been shaped by selection during historical pandemics, such as the Black Death (Klunk et al. 2022). It follows that earlier zoonotic pathogens, emerging alongside the agricultural revolution (Jones et al. 2013), may likewise have played major roles in shaping key immune and barrier-related genes. Accordingly, mucins, central to the mucus barrier and pathogen defense (Wagner et al. 2018), are likely to have been subjected to similar selective pressures. Alleles that conferred protection under one pathogen regime may now predispose individuals to immune hyperreactivity or autoimmunity, similar to immune-related genes such as *ERAP2* (Klunk et al. 2022). This continual evolutionary arms race of "push-pull" between pathogen-driven diversification and physiological constraint may explain the extent and potential benefit of standing VNTR variation. Consistent with this, pathogen-binding VNTRs in immune genes such as *CD209L* show population-specific signatures of balancing selection, and reveal that pathogen pressures can directly shape repeat-length diversity (Barreiro et al. 2005). Therefore, mucins, such as MUC1, could possess longer glycosylated repeat domains that strengthen mucosal barriers against pathogens but may affect important processes in different adaptive or consequential directions, such as immune (dis)regulation or embryo implantation during pregnancy (Brayman et al. 2004; Dharmaraj et al. 2009) (Fig. 4).

Future studies leveraging high-resolution genomic, transcriptomic, and proteomic tools will further clarify these dynamics. In particular, comprehensive exploration of the *mucinome* (Malaker et al. 2022; Lowery et al. 2024; Finn et al. 2025; Steigmeyer et al. 2025), linking VNTR diversity in mucin genes to corresponding glycosylation maps, will be essential for understanding how genetic structural variation modulates molecular recognition, barrier function, and microbial interactions. Integrative analyses across molecular layers promise to reveal how VNTR-driven diversity shapes adaptation, constraint, and disease susceptibility, offering a broader view of how repetitive sequences influence mammalian biology and evolution.

Acknowledgments

We thank Dr. Charikleia Karageorgiou for careful reading of the manuscript. All figures were created with BioRender or with specific instructions given to ChatGPT-5.

Funding

P.P. acknowledges support from the NSF (National Science Foundation; PRFB grant no. 2508185). O.G. acknowledges support from the NSF (National Science Foundation; grant nos. 2049947 and 2123284) and the NIH (National Institutes of Health; R35-GM156519).

Conflict of interest

The authors declare that they have no competing interests.

Data Availability

No data was generated for this study.

Literature Cited

1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012:491:56–65. https://doi.org/10.1038/nature11632.

Allison AC. The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria. Trans R Soc Trop Med Hyg. 1954:48:312–318. https://doi.org/10.1016/0035-9203(54)90101-7.

Altemose N, et al. A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. Elife. 2017:6:e28383. https://doi.org/10.7554/eLife.28383.

Angelova-Fischer I, et al. Distinct barrier integrity phenotypes in filaggrin-related atopic eczema following sequential tape stripping and lipid profiling: skin barrier integrity in filaggrin-AD. Exp Dermatol. 2011:20:351–356. https://doi.org/10.1111/j.1600-0625.2011.01259.x.

Aqil A, Speidel L, Pavlidis P, Gokcumen O. Balancing selection on genomic deletion polymorphisms in humans. Elife. 2023:12:e79111. https://doi.org/10.7554/eLife.79111.
Baalsrud HT, et al. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. Mol Biol Evol. 2018:35:593–606. https://doi.org/10.1093/molbev/msx311.
Bakhtiari M, et al. Variable number tandem repeats mediate the expression of proximal genes. Nat Commun. 2021:12:2075. https://doi.org/10.1038/s41467-021-22206-z.
Barlattani T, et al. The influence of PER3 VNTR genotypes on the age of onset in a group of bipolar I disorder patients: an exploratory study. Int J Bipolar Disord. 2024:12:25. https://doi.org/10.1186/s40345-024-00346-7.
Barnard KN, et al. Modified sialic acids on mucus and erythrocytes inhibit influenza A virus hemagglutinin and neuraminidase functions. J Virol. 2020:94:e01567-19. https://doi.org/10.1128/JVI.01567-19.
Barreiro LB, et al. The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. Am J Hum Genet. 2005:77:869–886. https://doi.org/10.1086/497613.
Baudat F, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science. 2010:327:836–840. https://doi.org/10.1126/science.1183439.
Bédard A-C, et al. Dopamine transporter gene variation modulates activation of striatum in youth with ADHD. Neuroimage. 2010:53:935–942. https://doi.org/10.1016/j.neuroimage.2009.12.041.
Bellizzi D, et al. A novel VNTR enhancer within the SIRT3 gene, a human homologue of SIR2, is associated with survival at oldest ages. Genomics. 2005:85:258–263. https://doi.org/10.1016/j.ygeno.2004.11.003.
Birnbaum R. Rediscovering tandem repeat variation in schizophrenia: challenges and opportunities. Transl Psychiatry. 2023:13:402. https://doi.org/10.1038/s41398-023-02689-8.
Brayman M, Thathiah A, Carson DD. MUC1: a multifunctional cell surface component of reproductive tissue epithelia. Reprod Biol Endocrinol. 2004:2:4. https://doi.org/10.1186/1477-7827-2-4.
Brown SJ, et al. Intragenic copy number variation within filaggrin contributes to the risk of atopic dermatitis with a dose-dependent effect. J Invest Dermatol. 2012:132:98–104. https://doi.org/10.1038/jid.2011.342.
Caspi A, et al. Role of genotype in the cycle of violence in maltreated children. Science. 2002:297:851–854. https://doi.org/10.1126/science.1072290.
Chaisson MJP, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015:517:608–611. https://doi.org/10.1038/nature13907.
Chaisson MJP, Sulovari A, Valdmanis PN, Miller DE, Eichler EE. Advances in the discovery and analyses of human tandem repeats. Emerg Top Life Sci. 2023:7:361–381. https://doi.org/10.1042/ETLS20230074.
Chen L, DeVries AL, Cheng CH. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. Proc Natl Acad Sci U S A. 1997a:94:3817–3822. https://doi.org/10.1073/pnas.94.8.3817.
Chen L, DeVries AL, Cheng CH. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. Proc Natl Acad Sci U S A. 1997b:94:3811–3816. https://doi.org/10.1073/pnas.94.8.3811.
Collins RL, Talkowski ME. Diversity and consequences of structural variation in the human genome. Nat Rev Genet. 2025:26:443–462. https://doi.org/10.1038/s41576-024-00808-9.
Cong L, Tu G, Liang D. A systematic review of the relationship between the distributions of aggrecan gene VNTR polymorphism and degenerative disc disease/osteoarthritis. Bone Joint Res. 2018:7:308–317. https://doi.org/10.1302/2046-3758.74.BJR-2017-0207.R1.
Conrad DF, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010:464:704–712. https://doi.org/10.1038/nature08516.
Course MM, Sulovari A, Gudsnuk K, Eichler EE, Valdmanis PN. Characterizing nucleotide variation and expansion dynamics in human-specific variable number tandem repeats. Genome Res. 2021:31:1313–1324. https://doi.org/10.1101/gr.275560.121.
Cross BW, Ruhl S. Glycan recognition at the saliva - oral microbiome interface. Cell Immunol. 2018:333:19–33. https://doi.org/10.1016/j.cellimm.2018.08.008.
Cui Y, et al. A genome-wide spectrum of tandem repeat expansions in 338,963 humans. Cell. 2024:187:2336–2341.e5. https://doi.org/10.1016/j.cell.2024.03.004.
Dale BA, Holbrook KA, Steinert PM. Assembly of stratum corneum basic protein and keratin filaments in macrofibrils. Nature. 1978:276:729–731. https://doi.org/10.1038/276729a0.
Dekker J, Rossen JW, Büller HA, Einerhand AW. The MUC family: an obituary. Trends Biochem Sci. 2002:27:126–131. https://doi.org/10.1016/S0968-0004(01)02052-7.
Desseyn JL, Aubert JP, Porchet N, Laine A. Evolution of the large secreted gel-forming mucins. Mol Biol Evol. 2000:17:1175–1184. https://doi.org/10.1093/oxfordjournals.molbev.a026400.
Deutsch EW, et al. High-quality peptide evidence for annotating non-canonical open reading frames as human proteins, 2024. bioRxiv 2024.09.09.612016. https://doi.org/10.1101/2024.09.09.612016
Devuyst O, et al. Autosomal dominant tubulointerstitial kidney disease. Nat Rev Dis Primers. 2019:5:60. https://doi.org/10.1038/s41572-019-0109-9.
Dhanisha SS, Guruvayoorappan C, Drishya S, Abeesh P. Mucins: structural diversity, biosynthesis, its role in pathogenesis and as possible therapeutic targets. Crit Rev Oncol Hematol. 2018:122:98–122. https://doi.org/10.1016/j.critrevonc.2017.12.006.
Dharmaraj N, Gendler SJ, Carson DD. Expression of human MUC1 during early pregnancy in the human MUC1 transgenic mouse model. Biol Reprod. 2009:81:1182–1188. https://doi.org/10.1095/biolreprod.109.079418.
Di Maio S, et al. Resolving intra-repeat variation in medically relevant VNTRs from short-read sequencing data using the cardiovascular risk gene LPA as a model. Genome Biol. 2024:25:167. https://doi.org/10.1186/s13059-024-03316-5.
Ding Y-C, et al. Evidence of positive selection acting at the human dopamine receptor D4 gene locus. Proc Natl Acad Sci U S A. 2002:99:309–314. https://doi.org/10.1073/pnas.012464099.
Doege KJ, Coulter SN, Meek LM, Maslen K, Wood JG. A human-specific polymorphism in the coding region of the aggrecan gene. Variable number of tandem repeats produce a range of core protein sizes in the general population. J Biol Chem. 1997:272:13974–13979. https://doi.org/10.1074/jbc.272.21.13974.
Dolzhenko E, et al. Characterization and visualization of tandem repeats at genome scale. Nat Biotechnol. 2024:42:1606–1614. https://doi.org/10.1038/s41587-023-02057-3.
Duan Y, Pan C, Shi J, Chen H, Zhang S. Association between interleukin-4 gene intron 3 VNTR polymorphism and cancer risk. Cancer Cell Int. 2014:14:131. https://doi.org/10.1186/s12935-014-0131-7.
Dumas Laura J, et al. DUF1220-domain copy number implicated in human brain-size pathology and evolution. Am J Hum Genet. 2012:91:444–454. https://doi.org/10.1016/j.ajhg.2012.07.016.

Eaaswarkhanth M, et al. Atopic dermatitis susceptibility variants in filaggrin hitchhike hornerin selective sweep. Genome Biol Evol. 2016:8:3240–3255. https://doi.org/10.1093/gbe/evw242.
Ebert P, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021:372: eabf7117. https://doi.org/10.1126/science.abf7117.
ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012:489:57–74. https://doi.org/10.1038/nature11247.
English AC, et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. Nat Biotechnol. 2025:43: 431–442. https://doi.org/10.1038/s41587-024-02225-z.
Eslami Rasekh M, Hernández Y, Drinan SD, Fuxman Bass JI, Benson G. Genome-wide characterization of human minisatellite VNTRs: population-specific alleles and gene expression differences. Nucleic Acids Res. 2021:49:4308–4324. https://doi.org/10.1093/nar/gkab224.
Finn SM, et al. GlycoFASP: a universal method to prepare complex mixtures for O-glycoproteomic analysis. Anal Chem. 2025:97: 23751–23756. https://doi.org/10.1021/acs.analchem.5c02666.
Fondon JW III, Garner HR. Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci U S A. 2004:101: 18058–18063. https://doi.org/10.1073/pnas.0408118101.
Ganz T. Defensins: antimicrobial peptides of innate immunity. Nat Rev Immunol. 2003:3:710–720. https://doi.org/10.1038/nri1180.
Gelfand Y, Hernandez Y, Loving J, Benson G. VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data. Nucleic Acids Res. 2014:42:8884–8894. https://doi.org/10.1093/nar/gku642.
Gemayel R, Cho J, Boeynaems S, Verstrepen KJ. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. Genes (Basel). 2012:3:461–480. https://doi.org/10.3390/genes3030461.
Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. 2010:44:445–477. https://doi.org/10.1146/annurev-genet-072610-155046.
Gent JI, et al. Distinct influences of tandem repeats and retrotransposons on CENH3 nucleosome positioning. Epigenetics Chromatin. 2011:4:3. https://doi.org/10.1186/1756-8935-4-3.
Gilad Y, Man O, Pääbo S, Lancet D. Human specific loss of olfactory receptor genes. Proc Natl Acad Sci U S A. 2003:100:3324–3327. https://doi.org/10.1073/pnas.0535697100.
Gökçümen O, Lee C. Copy number variants (CNVs) in primate species using array-based comparative genomic hybridization. Methods. 2009:49:18–25. https://doi.org/10.1016/j.ymeth.2009.06.001.
Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. Nat Protoc. 2022:17:1518–1552. https://doi.org/10.1038/s41596-022-00692-9.
Gu B, Chaisson MJP. TRCompDB: A reference of human tandem repeat sequence and composition variation from long-read assemblies, 2024. bioRxiv. https://doi.org/10.1101/2024.08.07.607105
Gyan BA, et al. Allelic polymorphisms in the repeat and promoter regions of the interleukin-4 gene and malaria severity in Ghanaian children. Clin Exp Immunol. 2004:138:145–150. https://doi.org/10.1111/j.1365-2249.2004.02590.x.
Gymrek M, Goren A. Missing heritability may be hiding in repeats. Science. 2021:373:1440–1441. https://doi.org/10.1126/science.abl7794.
Haddadi K, Sahebi M, Mahrooz A, ShayestehAzar M, Hashemi-Soteh MB. Association between vitamin D receptor gene polymorphism (rs731236) and aggrecan gene VNTR polymorphism with the risk of lumbar intervertebral disc degeneration. Caspian J Intern Med. 2022:13:418–424. https://doi.org/10.22088/cjim.13.2.418.
Hallast P, et al. Assembly of 43 human Y chromosomes reveals extensive complexity and variation. Nature. 2023:621:355–364. https://doi.org/10.1038/s41586-023-06425-6.
Handsaker RE, et al. Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease. Cell. 2025:188:623–-639.e19. https://doi.org/10.1016/j.cell.2024.11.038.
Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018:19:286–298. https://doi.org/10.1038/nrg.2017.115.
Hedrick PW. Genetic polymorphism in heterogeneous environments: the age of genomics. Annu Rev Ecol Evol Syst. 2006:37:67–93. https://doi.org/10.1146/annurev.ecolsys.37.091305.110132.
Hernandez RD, et al. Classic selective sweeps were rare in recent human evolution. Science. 2011:331:920–924. https://doi.org/10.1126/science.1198878.
Higuchi S, Nakamura Y, Saito S. Characterization of a VNTR polymorphism in the coding region of the CEL gene. J Hum Genet. 2002:47:213–215. https://doi.org/10.1007/s100380200027.
Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020:21:171–189. https://doi.org/10.1038/s41576-019-0180-9.
Huddleston J, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017:27:677–685. https://doi.org/10.1101/gr.214007.116.
Hunt BG, et al. Relaxed selection is a precursor to the evolution of phenotypic plasticity. Proc Natl Acad Sci U S A. 2011:108: 15936–15941. https://doi.org/10.1073/pnas.1104825108.
Ichikawa K, Kawahara R, Asano T, Morishita S. A landscape of complex tandem repeats within individual human genomes. Nat Commun. 2023:14:5530. https://doi.org/10.1038/s41467-023-41262-1.
Jain M, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018:36:338–345. https://doi.org/10.1038/nbt.4060.
Javadzadeh S, et al. Analysis of targeted and whole genome sequencing of PacBio HiFi reads for a comprehensive genotyping of gene-proximal and phenotype-associated Variable Number Tandem Repeats. PLoS Comput Biol. 2025:21:e1012885. https://doi.org/10.1371/journal.pcbi.1012885.
Jha AN, et al. IL-4 haplotype -590T, -34T and intron-3 VNTR R2 is associated with reduced malaria risk among ancestral Indian tribal populations. PLoS One. 2012:7:e48136. https://doi.org/10.1371/journal.pone.0048136.
Ji HJ, Salzberg SL. Upstream open reading frames may contain hundreds of novel human exons. PLoS Comput Biol. 2024:20: e1012543. https://doi.org/10.1371/journal.pcbi.1012543.
Jones BA, et al. Zoonosis emergence linked to agricultural intensification and environmental change. Proc Natl Acad Sci U S A. 2013:110: 8399–8404. https://doi.org/10.1073/pnas.1208059110.
Karageorgiou C, Gokcumen O, Dennis MY. Deciphering the role of structural variation in human evolution: a functional perspective. Curr Opin Genet Dev. 2024:88:102240. https://doi.org/10.1016/j.gde.2024.102240.
Kelada SNP, et al. 5′ and 3′ region variability in the dopamine transporter gene (SLC6A3), pesticide exposure and Parkinson's disease risk: a hypothesis-generating study. Hum Mol Genet. 2006:15: 3055–3062. https://doi.org/10.1093/hmg/ddl247.
Kerdoncuff E, et al. Revisiting the evolution of lactase persistence: Insights from south Asian genomes, 2025. bioRxiv 2025.11.05.686799. https://doi.org/10.1101/2025.11.05.686799
Kirby A, et al. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat Genet. 2013:45:299–303. https://doi.org/10.1038/ng.2543.

Klunk J, et al. Evolution of immune genes is associated with the Black Death. Nature. 2022:611:312–319. https://doi.org/10.1038/s41586-022-05349-x.
Kulpraneet M, et al. Analysis of IL-4 promoter and VNTR polymorphisms in Thai patients with pulmonary tuberculosis. Trop Biomed. 2019:36:874–882.
Kunugi H, et al. A functional polymorphism in the promoter region of monoamine oxidase-A gene and mood disorders. Mol Psychiatry. 1999:4:393–395. https://doi.org/10.1038/sj.mp.4000558.
Lamkin M, Gymrek M. The emerging role of tandem repeats in complex traits. Nat Rev Genet. 2024:25:452–453. https://doi.org/10.1038/s41576-024-00736-8.
Lander ES, et al. Initial sequencing and analysis of the human genome. Nature. 2001:409:860–921. https://doi.org/10.1038/35057062.
Lang T, Hansson GC, Samuelsson T. Gel-forming mucins appeared early in metazoan evolution. Proc Natl Acad Sci U S A. 2007:104:16209–16214. https://doi.org/10.1073/pnas.0705984104.
Liao W-W, et al. A draft human pangenome reference. Nature. 2023:617:312–324. https://doi.org/10.1038/s41586-023-05896-x.
Linthorst J, et al. Extreme enrichment of VNTR-associated polymorphicity in human subtelomeres: genes with most VNTRs are predominantly expressed in the brain. Transl Psychiatry. 2020:10:369. https://doi.org/10.1038/s41398-020-01060-5.
Logsdon GA, et al. Complex genetic variation in nearly complete human genomes. Nature. 2025:644:430–441. https://doi.org/10.1038/s41586-025-09140-6.
Lowery SC, et al. Glycosite mapping and in situ mass spectrometry imaging of MUC2 glycopeptides via on-slide digestion with mucinase StcE, 2024. bioRxiv 2024.09.16.613285. https://doi.org/10.1101/2024.09.16.613285.
Lu T-Y, Smaruj PN, Fudenberg G, Mancuso N, Chaisson MJP. The motif composition of variable number tandem repeats impacts gene expression. Genome Res. 2023:33:511–524. https://doi.org/10.1101/gr.276768.122.
Luis Villanueva-Cañas J, et al. New genes and functional innovation in mammals. Genome Biol Evol. 2017:9:1886–1900. https://doi.org/10.1093/gbe/evx136.
Ma Y, Yuan W, Cui W, Li MD. Meta-analysis reveals significant association of 3′-UTR VNTR in SLC6A3 with smoking cessation in Caucasian populations. Pharmacogenomics J. 2016:16:10–17. https://doi.org/10.1038/tpj.2015.44.
Mac Donagh J et al. Structured tandem repeats in protein interactions. Int J Mol Sci. 2024:25:2994. https://doi.org/10.3390/ijms25052994.
MacKenzie A, Quinn J. A serotonin transporter gene intron 2 polymorphic region, correlated with affective disorders, has allele-dependent differential enhancer-like properties in the mouse embryo. Proc Natl Acad Sci U S A. 1999:96:15251–15255. https://doi.org/10.1073/pnas.96.26.15251.
Makova KD, et al. The complete sequence and comparative analysis of ape sex chromosomes. Nature. 2024:630:401–411. https://doi.org/10.1038/s41586-024-07473-2.
Malaker SA, et al. Revealing the human mucinome. Nat Commun. 2022:13:3542. https://doi.org/10.1038/s41467-022-31062-4.
Manca M, et al. The regulation of monoamine oxidase A gene expression by distinct variable number tandem repeats. J Mol Neurosci. 2018:64:459–470. https://doi.org/10.1007/s12031-018-1044-z.
Manigbas CA, et al. A phenome-wide association study of tandem repeat variation in 168,554 individuals from the UK Biobank. Nat Commun. 2024:15:10521. https://doi.org/10.1038/s41467-024-54678-0.
Marshall JN, et al. Variable number tandem repeats—their emerging role in sickness and health. Exp Biol Med (Maywood). 2021:246:1368–1376. https://doi.org/10.1177/15353702211003511.
Mill J, Asherson P, Craig I, D'Souza UM. Transient expression analysis of allelic variants of a VNTR in the dopamine transporter gene (DAT1). BMC Genet. 2005:6:3. https://doi.org/10.1186/1471-2156-6-3.
Mills RE, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011:470:59–65. https://doi.org/10.1038/nature09708.
Mukamel RE, et al. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science. 2021:373:1499–1505. https://doi.org/10.1126/science.abg8289.
Mukamel RE, et al. Repeat polymorphisms underlie top genetic risk loci for glaucoma and colorectal cancer. Cell. 2023:186:3659–-3673.e23. https://doi.org/10.1016/j.cell.2023.07.002.
Nakamura Y, Koyama K, Matsushima M. VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. J Hum Genet. 1998:43:149–152. https://doi.org/10.1007/s100380050059.
Nikkanen J, et al. An evolutionary trade-off between host immunity and metabolism drives fatty liver in male mice. Science. 2022:378:290–295. https://doi.org/10.1126/science.abn9886.
Nurk S, et al. The complete sequence of a human genome. Science. 2022:376:44–53. https://doi.org/10.1126/science.abj6987.
O'Donnell S, et al. Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in Saccharomyces cerevisiae. Nat Genet. 2023:55:1390–1399. https://doi.org/10.1038/s41588-023-01459-y.
Oliver PL, et al. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet. 2009:5:e1000753. https://doi.org/10.1371/journal.pgen.1000753.
Olson ND, et al. Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet. 2023:24:464–483. https://doi.org/10.1038/s41576-023-00590-0.
Pajic P, et al. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. Elife. 2019:8:e44628. https://doi.org/10.7554/eLife.44628.
Pajic P, et al. A mechanism of gene evolution generating mucin function. Sci Adv. 2022:8:eabm8757. https://doi.org/10.1126/sciadv.abm8757.
Pajic P, Landau L, Gokcumen O, Ruhl S. Saliva protein genes in humans were shaped during primate evolution. Genome Biol Evol. 2025:17:evaf165. https://doi.org/10.1093/gbe/evaf165.
Pajic P, Lin Y-L, Xu D, Gokcumen O. The psoriasis-associated deletion of late cornified envelope genes LCE3B and LCE3C has been maintained under balancing selection since Human Denisovan divergence. BMC Evol Biol. 2016:16:265. https://doi.org/10.1186/s12862-016-0842-6.
Pajic PM. Mucins provide insights into mechanisms of functional evolution. PhD dissertation. University at Buffalo; 2025.
Parvanov ED, Petkov PM, Paigen K. Prdm9 controls activation of mammalian recombination hotspots. Science. 2010:327:835. https://doi.org/10.1126/science.1181495.
Plender EG, et al. Structural and genetic diversity in the secreted mucins MUC5AC and MUC5B. Am J Hum Genet. 2024:111:1700–1716. https://doi.org/10.1016/j.ajhg.2024.06.007.
Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Hum Mol Genet. 2018:27:R234–R241. https://doi.org/10.1093/hmg/ddy177.
Ponting CP. Biological function in the twilight zone of sequence conservation. BMC Biol. 2017:15:71. https://doi.org/10.1186/s12915-017-0411-5.
Ponting CP, Hardison RC. What fraction of the human genome is functional? Genome Res. 2011:21:1769–1776. https://doi.org/10.1101/gr.116814.110.

Porubsky D, et al. Human de novo mutation rates from a four-generation pedigree reference. Nature. 2025:643:427–436. https://doi.org/10.1038/s41586-025-08922-2.
Prodanov T, et al. Locityper enables targeted genotyping of complex polymorphic genes. Nat Genet. 2025:57:2901–2908. https://doi.org/10.1038/s41588-025-02362-4.
Quilez J, et al. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res. 2016:44:3750–3762. https://doi.org/10.1093/nar/gkw219.
Readman C, Indhu-Shree R-B, Friedman JM, Birol I. A comprehensive tandem repeat catalog of the human genome, 2024. medRxiv. https://doi.org/10.1101/2024.06.19.24309173
Reily C, Stewart TJ, Renfrow MB, Novak J. Glycosylation in health and disease. Nat Rev Nephrol. 2019:15:346–366. https://doi.org/10.1038/s41581-019-0129-4.
Reith MEA, et al. The dopamine transporter gene SLC6A3: multidisease risks. Mol Psychiatry. 2022:27:1031–1046. https://doi.org/10.1038/s41380-021-01341-5.
Ren J, Gu B, Chaisson MJP. Vamos: variable-number tandem repeats annotation using efficient motif sets. Genome Biol. 2023:24:175. https://doi.org/10.1186/s13059-023-03010-y.
Rhie A, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021:592:737–746. https://doi.org/10.1038/s41586-021-03451-0.
Rhie A, et al. The complete sequence of a human Y chromosome. Nature. 2023:621:344–354. https://doi.org/10.1038/s41586-023-06457-y.
Ross CA, Tabrizi SJ. Huntington's disease: from molecular pathogenesis to clinical treatment. Lancet Neurol. 2011:10:83–98. https://doi.org/10.1016/S1474-4422(10)70245-3.
Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. Elife. 2014:3:e03523. https://doi.org/10.7554/eLife.03523.
Sabino FC, et al. Evolutionary history of the PER3 variable number of tandem repeats (VNTR): idiosyncratic aspect of primate molecular circadian clock. PLoS One. 2014:9:e107198. https://doi.org/10.1371/journal.pone.0107198.
Sabol SZ, Hu S, Hamer D. A functional polymorphism in the monoamine oxidase A gene promoter. Hum Genet. 1998:103:273–279. https://doi.org/10.1007/s004390050816.
Saitou M, et al. Sex-specific phenotypic effects and evolutionary history of an ancient polymorphic deletion of the human growth hormone receptor. Sci Adv. 2021:7:eabi4476. https://doi.org/10.1126/sciadv.abi4476.
Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012:13:745–753. https://doi.org/10.1038/nrg3295.
Schaper E, Gascuel O, Anisimova M. Deep conservation of human protein tandem repeats within the eukaryotes. Mol Biol Evol. 2014:31:1132–1148. https://doi.org/10.1093/molbev/msu062.
Scheer K, et al. Adaptive increase of amylase gene copy number in Peruvians driven by potato-rich diets, 2025. bioRxiv. https://doi.org/10.1101/2025.03.25.644684
Schloissnig S, et al. Structural variation in 1,019 diverse humans based on long-read sequencing. Nature. 2025:644:442–452. https://doi.org/10.1038/s41586-025-09290-7.
Ségurel L, Bon C. On the evolution of lactase persistence in humans. Annu Rev Genomics Hum Genet. 2017:18:297–319. https://doi.org/10.1146/annurev-genom-091416-035340.
Simon M, et al. Evidence that filaggrin is a component of cornified cell envelopes in human plantar epidermis. Biochem J. 1996:317:173–177. https://doi.org/10.1042/bj3170173.
Steigmeyer AD, Lowery SC, Rangel-Angarita V, Malaker SA. Decoding extracellular protein glycosylation in human health and disease. Annu Rev Anal Chem (Palo Alto Calif). 2025:18:241–264. https://doi.org/10.1146/annurev-anchem-071024-124203.
Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science. 2020:368:1449–1454. https://doi.org/10.1126/science.aaz1646.
Sulovari A, et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc Natl Acad Sci U S A. 2019:116:23243–23253. https://doi.org/10.1073/pnas.1912175116.
Tanudisastro HA, et al. Polymorphic tandem repeats influence cell type-specific gene expression across the human immune landscape, 2025. bioRxiv. https://doi.org/10.1101/2024.11.02.621562
Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011:12:692–702. https://doi.org/10.1038/nrg3053.
Tishkoff SA, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007:39:31–40. https://doi.org/10.1038/ng1946.
Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012:13:36–46. https://doi.org/10.1038/nrg3117.
van Dyck CH, et al. Increased dopamine transporter availability associated with the 9-repeat allele of the SLC6A3 gene. J Nucl Med. 2005:46:745–751.
van Leeuwen EM, et al. Population-specific genotype imputations using minimac or IMPUTE2. Nat Protoc. 2015:10:1285–1296. https://doi.org/10.1038/nprot.2015.077.
Van Oss SB, Carvunis A-R. De novo gene birth. PLoS Genet. 2019:15:e1008160. https://doi.org/10.1371/journal.pgen.1008160.
Vasiliou SA, et al. The SLC6A4 VNTR genotype determines transcription factor binding and epigenetic variation of this gene in response to cocaine in vitro: epigenetics of the SLC6A4 gene. Addict Biol. 2012:17:156–170. https://doi.org/10.1111/j.1369-1600.2010.00288.x.
Veilleux CC, et al. Human subsistence and signatures of selection on chemosensory genes. Commun Biol. 2023:6:683. https://doi.org/10.1038/s42003-023-05047-y.
Venter JC, et al. The sequence of the human genome. Science. 2001:291:1304–1351. https://doi.org/10.1126/science.1058040.
Villanea FA, et al. The MUC19 gene: an evolutionary history of recurrent introgression and natural selection. Science. 2025:389:eadl0882. https://doi.org/10.1126/science.adl0882.
Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. Unstable tandem repeats in promoters confer transcriptional evolvability. Science. 2009:324:1213–1216. https://doi.org/10.1126/science.1170097.
Vollger MR, et al. Segmental duplications and their variation in a complete human genome. Science. 2022:376:eabj6965. https://doi.org/10.1126/science.abj6965.
Wagner CE, Wheeler KM, Ribbeck K. Mucins and their role in shaping the functions of mucus barriers. Annu Rev Cell Dev Biol. 2018:34:189–215. https://doi.org/10.1146/annurev-cellbio-100617-062818.
Wang E, et al. The genetic architecture of selection at the human dopamine receptor D4 (DRD4) gene locus. Am J Hum Genet. 2004:74:931–944. https://doi.org/10.1086/420854.
Wang Y, et al. Contribution of both positive selection and relaxation of selective constraints to degeneration of flyability during geese domestication. PLoS One. 2017:12:e0185328. https://doi.org/10.1371/journal.pone.0185328.

Weisburd B, et al. Defining a tandem repeat catalog and variation clusters for genome-wide analyses, 2025. bioRxiv. 2024.10.04.615514. https://doi.org/10.1101/2024.10.04.615514

Weisman CM. The origins and functions of De Novo genes: against all odds? J Mol Evol. 2022:90:244–257. https://doi.org/10.1007/s00239-022-10055-3.

Wenzel A, et al. Single molecule real time sequencing in ADTKD-MUC1 allows complete assembly of the VNTR and exact positioning of causative mutations. Sci Rep. 2018:8:4170. https://doi.org/10.1038/s41598-018-22428-0.

Wu H, et al. Telomere-to-telomere genome assembly of a male goat reveals variants associated with cashmere traits. Nat Commun. 2024:15:10041. https://doi.org/10.1038/s41467-024-54188-z.

Xia S, Chen J, Arsala D, Emerson JJ, Long M. Functional innovation through new genes as a general evolutionary process. Nat Genet. 2025:57:295–309. https://doi.org/10.1038/s41588-024-02059-0.

Xu D, et al. Recent evolution of the salivary mucin MUC7. Sci Rep. 2016:6:31791. https://doi.org/10.1038/srep31791.

Xu D, Gokcumen O, Khurana E. Loss-of-function tolerance of enhancers in the human genome. PLoS Genet. 2020:16:e1008663. https://doi.org/10.1371/journal.pgen.1008663.

Xu G, Mei Q, Zhou D, Wu J, Han L. Vitamin D receptor gene and aggrecan gene polymorphisms and the risk of intervertebral disc degeneration—a meta-analysis. PLoS One. 2012:7:e50243. https://doi.org/10.1371/journal.pone.0050243.

Yang Q, et al. Tuning the tropism and infectivity of SARS-CoV-2 virus-like particles for mRNA delivery. Nucleic Acids Res. 2025:53:gkaf133. https://doi.org/10.1093/nar/gkaf133.

Yilmaz F, et al. High level of complexity and global diversity of the 3q29 locus revealed by optical mapping and long-read sequencing. Genome Med. 2023:15:35. https://doi.org/10.1186/s13073-023-01184-5.

Yılmaz F, et al. Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. Science. 2024:386:eadn0609. https://doi.org/10.1126/science.adn0609.

Yoo D, et al. Complete sequencing of ape genomes. Nature. 2025:641:401–418. https://doi.org/10.1038/s41586-025-08816-3.

Zhang S, et al. Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: implications for biological regulation and human traits. Cell Genom. 2024:4:100699. https://doi.org/10.1016/j.xgen.2024.100699.

Zhang S, et al. Integrated analysis of the complete sequence of a macaque genome. Nature. 2025:640:714–721. https://doi.org/10.1038/s41586-025-08596-w.

Zhao L, Svetec N, Begun DJ. De Novo genes. Annu Rev Genet.2024:58:211–232. https://doi.org/10.1146/annurev-genet-111523-102413.

Zhao L, Zhou W, He J, Li D-Z, Li H-T. Positive selection and relaxed purifying selection contribute to rapid evolution of male-biased genes in a dioecious flowering plant. Elife. 2023:12:RP89941. https://doi.org/10.7554/elife.89941.2.

Ziaei Jam H, et al. LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. Genome Biol. 2024:25:176. https://doi.org/10.1186/s13059-024-03319-2.

Associate editor: George Zhang