The expression of satellite DNA-encoded proteins in the human genome

Yongxian Chen (陈勇先)[1#], Jie Li (李杰)[1#], Xuan Zhou (周璇)[1#], Qian Zhao (赵倩)[2], Jian Li (李健)[1], Yong Wang (王勇)[3], Huolian Liu (刘伙莲)[4], Mingqian Feng (丰明乾)[3], Wen Ni (倪雯)[4], Yin Zhang (张寅)[1]*, Yabin Guo (郭雅彬)[1]*

[1]Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangdong-Hong Kong Joint Laboratory for RNA Medicine, Medical Research Center; [4]Department of Pathology, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China 510120.

[2] Department of Gynecology, Guangdong Women and Children Hospital, Guangzhou, China.

[3]College of Biomedicine and Health, Huazhong Agricultural University, Wuhan, Hubei, China

[#]Equal contribution

* zhangy525@mail.sysu.edu.cn; guoyb9@mail.sysu.edu.cn

Abstract

The discovery of dark matter proteins has primarily relied on predicting open reading frames (ORFs) from known RNA molecules, followed by experimental validation. Although this approach has yielded many excellent studies in recent years, its positive output has become increasingly low while costs continue to rise. Meanwhile, whether large portions of the genomic dark matter—such as satellite DNA—encode proteins remains unknown. In this study, we employed 6-frame *in silico* translation to directly predict ORFs (istORFs) from genomic DNA sequences and their corresponding amino acid sequences (isteins), and then searched for signature peptides and matching RNA sequences in proteomics and ribo-seq databases. Unexpectedly, the positive rate was remarkably high. We identified a large number of satellite DNA-derived ORFs; among the 1,000 highest-copy-number isteins, 415 had detectable signature peptides in mass spectrometry datasets via Pepquery, and 401 showed matching translation signals in ribo-seq data. The MESSS gene family, derived from a novel form of Sat2/3, comprises over 10,000 predicted ORFs with a combined coding region exceeding 6 Mb. Using a specific antibody, we detected MESSS expression in multiple cell lines. These findings provide initial evidence that at least some satellite DNA sequences possess coding capacity and are capable of expressing proteins. We also analyzed conserved isteins across multiple primate and mouse T2T genomes and found that even among highly conserved ORFs, many had not been previously discovered or annotated. Based on these results, we propose a series of hypotheses that await future validation.

Introduction

The recent completion of telomere-to-telomere (T2T) human reference genomes has closed hundreds of millions of base pairs of gaps that existed in traditional reference genomes such as GRCh38/hg38, particularly in centromeres, the short arms of acrocentric chromosomes, and highly repetitive regions, thus ushering in an era of sequence-based exploration of these "dark matter" regions of the genome [1, 2]. For a long time, traditional comparative genomics held that the human and great ape (e.g., chimpanzee) genomes are highly similar, with sequence identity exceeding 98% [3]. However, cross-species comparisons using T2T assemblies have revised this view: the Chinese population reference genome T2T-YAO [2], when compared with the European hydatidiform mole T2T-CHM13 [1], differs by thousands of structural variants and approximately 10% non-syntenic sequences, revealing population-specific genomic diversity; similarly, systematic analysis of six non-human ape T2T genomes [4] has shown that approximately 10–15% of the genome—including subtelomeric heterochromatin, lineage-specific segmental duplications, and recently evolved gene families—had been completely masked or severely underestimated in previous studies due to their high repetitiveness and structural dynamics. Of course, these differences were not entirely unrecognizable before; the so-called >98% similarity was actually based on comparisons of genes shared between humans and apes. By the same metric, human and mouse genes are also ~95% similar. Yet even now, it remains difficult to interpret the species-level and individual-level differences revealed by T2T genomes, because these genomic dark matter regions appear to have little obvious meaning. Seeing sequence differences on the surface without understanding what they mean is, in effect, not really seeing differences at all.

Proteins are the primary drivers of phenotypes, whereas most ncRNAs play regulatory roles. Thus, protein-coding genes occupy a more central position in the genome than ncRNA (non-coding) genes, a view supported by decades of research. In recent years,

studies of non-canonical ORFs (ncORFs) and small ORFs (smORFs) have become highly active, yielding many high-quality publications [5–8]. Yet the field has not seen the same explosive momentum that accompanied the rise of ncRNA research. Current efforts remain largely at the stage of high-throughput screening, with limited progress toward deeper functional characterization. Screening technologies are becoming more sophisticated and costly, but the rate of positive discoveries continues to decline. This is because: (i) the ORFs being identified are mostly located in the untranslated regions of known mRNAs, annotated lncRNAs, or circular RNAs, and the encoded peptides are generally very short; (ii) these small peptides are present at very low abundance in the proteome; (iii) most ncORFs are poorly conserved, and short, poorly conserved ORFs are more likely to have arisen by chance, with their products resembling background noise; and (iv) detection of these small peptides relies heavily on highly sensitive, advanced technologies and can rarely be confirmed by conventional methods such as Western blot. At the same time, a vast amount of genomic dark matters—including satellite DNA, tandem repeats, and low-complexity sequences—has not been examined for coding potential.

More than 70% of the human genome is actively transcribed, and even centromeres and telomeres can be transcribed under certain conditions such as stress [9], implying that, in principle, the entire human genome is capable of being transcribed. In recent years, various forms of non-canonical translation have been discovered, including translation of RNAs that lack 5' caps and 3' poly(A) tails [10]. When such dark matter RNAs reach the cytoplasm, they have a finite probability of being translated into peptides. Thus, logically, any sequence in the genome that can encode a peptide may, under certain conditions, be transcribed and translated. We also noticed that GenBank contains proteins from various species with highly periodic repeats and high-copy-number tandem repeat ORFs, whose encoding DNA sequences are also highly repetitive, suggesting a satellite or tandem repeat origin (file: sat_pro). However, the human genome annotation contains almost no such proteins. The few periodically

repetitive proteins in humans—such as certain keratins, mucins, and zinc finger proteins—have genes that do not resemble typical satellite DNA but rather conventional genes. Human and mouse gene annotation is very stringent, requiring strong evidence—a practice that certainly avoids the pitfalls of misannotation, but may also exclude certain repetitive sequences that are genuinely capable of being expressed. Identifying them would greatly enhance our understanding of the current genome.

We therefore reasoned that it might be possible to predict dark proteins directly from genomic DNA sequences and then search for evidence of their expression, thereby including large regions of genomic dark matter that had not been previously explored. In this study, we performed 6-frame *in silico* translation on the chromosomes of multiple T2T genomes—three forward and three reverse frames—and extracted the predicted peptide sequences for further analysis. We are fully aware that such DNA-based machine translation does not reflect the actual situation in the cell—not least because it does not account for RNA splicing—but our goal was not to identify intact coding genes and full-length proteins in one step. Rather, we aimed to find evidence for the expression of previously unknown peptides, which would already be a meaningful advance. In this study, we predicted a large number of previously unreported proteins/peptides, some with extremely high copy numbers, some unique to humans, and some showing conservation across primates or even between primates and mouse. We validated the translation of a subset of these proteins using mass spectrometry-based proteomics and ribo-seq. We also carried out a more in-depth investigation of the most abundant satellite ORF family, MESSS, including confirmation of its expression in multiple cell lines using a specific antibody.

Results

6-frame *in silico* translation of T2T genomes

Although 6-frame translation is a mature method for predicting coding regions, it has

previously been used mostly for compact genomes such as those of prokaryotes. In the human genome, the discovery of smORFs and ncORFs has largely relied on known RNA transcripts (mRNAs, lncRNAs, etc.). This transcript-centric approach is robust, but it may miss certain unknown proteins due to extremely low transcription levels or unusual spatial distribution of the transcripts. We reasoned that directly mining potential protein-coding information from DNA—even for a genome as complex as the human genome—offers unique advantages that complement transcript-based strategies.

Machine translation is not merely a conversion of nucleotide sequences into amino acid sequences; it also provides a different way of presenting information. Although we primarily rely on software tools, computer programs, and increasingly on AI tools for genomic data processing, the role of human inspection has not been replaced. The ability to process ambiguous patterns, to integrate diverse prior knowledge, and to engage in divergent thinking—all of which are strengths of the human mind—are not yet within the reach of computer programs or artificial intelligence. Some patterns that are obvious to a human observer may be completely invisible to a computer. Therefore, manual examination of sequences remains an important approach for discovering novel patterns. However, DNA sequences composed of only four nucleotides are not intuitively readable to the human brain, whereas sequences composed of 20 amino acids are far more interpretable. Although a DNA sequence contains all the information necessary for translation, the reading frame and strand orientation are not immediately obvious to a human reader. Translating a DNA sequence into six amino acid sequences flattens the information of reading frames and orientations into one dimension—a process of dimensionality reduction. Furthermore, M (representing the start codon) and Z (representing the stop codon) delimit the boundaries of potential proteins, breaking what appears to be an infinitely long sequence into discrete, finite-length units. These three aspects—(i) expanding the alphabet from four letters to twenty, (ii) dimensionality reduction, and (iii) segmentation into units—greatly enhance the readability of the genome to the human

brain.

We are fully aware that the proteins predicted in this way may not all be expressed—indeed, most of them may not be. However, we do not need to find many unknown proteins at once. We can follow up with validation of the most promising candidates among these predictions. Even finding just a few—or even one—previously unknown protein would be a meaningful discovery. We read the machine-translated sequences from left to right and extracted the sequences between the first M and the first Z (Supplementary Fig. 1A). Because these sequences are neither annotated protein genes nor even exons, we named these units ***isteins*** (in-silico-translated-ein), and the corresponding genomic DNA sequences ***istORFs*** (in-silico-translated open reading frames). For each T2T genome, we extracted isteins of ≥100 amino acids, as longer polypeptides are more likely to have functional potential. This does not mean that isteins shorter than 100 aa are devoid of function. We also recognize that non-canonical translation may not necessarily initiate at the first ATG or terminate at the classical stop codons (TAA, TAG, TGA), and that our predicted isteins may not represent the full-length proteins that would be produced after RNA splicing. However, for an initial investigation, it is more important to focus on a representative set of candidates rather than to aim for exhaustive coverage. Should this study succeed, additional candidates can be incorporated in future work.

The CHM13 genome is derived from a hydatidiform mole, in which some repetitive sequences tend to be amplified relative to somatic cells. For istORFs in repetitive regions, we therefore used T2T-YAO as the primary reference, as it is derived from a somatic cell and is currently one of the highest-quality human assemblies. For low-copy-number istORFs in conventional coding regions, we relied on CHM13 because of its more comprehensive annotation. YAO contains 982,057 isteins of ≥100 aa, while CHM13 contains 1,039,418. We also performed 6-frame translation on additional human T2T genomes, including Han1 [11], CN1 CN1[12], and hg002 [13] (Files: isteins). For annotations not available in YAO, we used CHM13. The same

procedure was applied to various primate and mouse T2T genomes (Supplementary Table 1). For each genome, we further extracted tandem-repeat isteins (File: Tandem-repeat isteins). The proportion of tandem-repeat isteins varies across genomes, and even among human genomes, reflecting either genuine individual differences or assembly quality. Mouse and gorilla genomes showed particularly high numbers of tandem-repeat isteins (Supplementary Table 2).

As the istein length threshold decreases, the number of predicted isteins increases sharply, but this relationship is not smooth. For example, the YAO genome shows two prominent spikes at 99 aa and 127 aa (Supplementary Fig. 1B). The 99-aa isteins are not included in our ≥100 aa dataset, but we examined them specifically and found that they are highly abundant on chromosome Y, with one particular sequence present at 1,492 identical copies (which we named MEW99, due to its enrichment in the MEW 3-mer motif), also encoded by Sat2/3. These spikes correspond to HOR (high-order repeats) events in the genome and merit further attention.

Distribution of istORFs across different genomic regions

We first examined the genomic distribution of istORFs based on genome annotations and RepeatMasker (Fig. 1A). Interestingly, the distribution of istORFs across different annotation categories was roughly proportional to the overall genomic composition of these categories, suggesting that, at a purely mathematical level, the potential to form ORFs does not differ dramatically across regions. However, satellite DNA regions clearly deviated from this general pattern. Satellite DNA is primarily located at centromeres, the short arms of acrocentric chromosomes (chr13, chr14, chr15, chr21 and chr22), and the long arm of chrY. Some satellite DNA regions, such as the centromeres of chr7, chr9 and chr20, contained sparse istORFs, whereas others—notably the chr1q12 region near the centromere and the short arm of chr15—were among the most istORF-dense regions in the entire genome (Fig. 1B–C, Supplementary Fig. 2). Under the traditional view, satellite DNA is not considered to

possess coding capacity; its sequences are monotonously repetitive and appear uninteresting to the human eye. Yet when we forcibly translated them, we observed dense ORF clusters in certain regions. Satellite DNA frequently undergoes replication slippage and unequal crossing-over, leading to frameshifts, which theoretically should make it difficult to maintain open reading frames. This unexpected observation raises the possibility that satellite DNA may possess some coding capacity and may even be under positive selective pressure.

We manually inspected the istein files of human T2T genomes, including YAO and CHM13. The first feature that caught our attention was the overwhelming abundance of repetitive peptides. This repetitiveness manifests at two levels: first, the internal periodic repeats or simple repeats within individual amino acid sequences; and second, the copy number of isteins themselves, with many highly similar isteins forming large families and even identical isteins occurring in multiple copies. Such a striking pattern would have been extremely difficult to detect without translation into amino acid sequences. We believe that most molecular biologists, upon seeing so many identical or highly similar amino acid sequences, would immediately infer protein-level selection, because it is mathematically implausible to maintain such amino acid identity solely through constraints at the DNA or RNA level without incurring nonsense mutations. In scientific discovery, noticing anomalies is a critical first step—anomalies hint at hidden phenomena that merit investigation, even if the eventual findings differ from our initial expectations.

The complexity distribution of isteins and satellite-derived ORFs

Motivated by the abundance of repetitive proteins observed in the previous section—which exhibit considerably simpler primary structures than conventional proteins—we sought to quantify protein sequence complexity. We developed a formula (see Methods) to compute sequence complexity on a scale from 0 to 1, and applied it to all isteins ≥100 aa from each T2T genome. Fig. 2 and Supplementary Fig. 3 show the

complexity distributions across species. Great ape genomes displayed a characteristic two-peak pattern with a valley in the middle, whereas non-great-ape genomes lacked the low-complexity peak entirely. Previous studies have shown that the mouse genome is more repetitive than the human genome [14], and our analyses below also indicate that the genomes of gibbons and macaques are no less repetitive than the human genome. Thus, these species do not lack simple repeat sequences *per se*; rather, such sequences rarely form isteins ≥100 aa. If at least some of the low-complexity ORFs (complexity 0.2–0.35) in great apes are translated, this would suggest that the great-ape proteome contains a larger number of longer low-complexity proteins—a possibility that remains to be tested. Notably, the low-complexity peak in the human istein distribution is the least prominent among the great apes. Instead, humans have more mid-complexity isteins, and their distribution is the least smooth, with several pronounced "spikes" (red columns in Fig. 2A), suggesting that specific istein families have undergone extensive amplification—consistent with previous reports of HORs in the human genome.

To identify these high-copy-number isteins, we ranked isteins by copy number within each genome (Supplementary Table 3). The copy-number distribution of human istORFs is more aggregated than those of chimpanzee and gorilla (Supplementary Fig. 4), consistent with a higher prevalence of HORs in the human lineage [15]. The highest-copy-number ORFs in the genome are almost exclusively encoded by satellite DNA. In the human genome, the most abundant istein families fall into three groups. One group is encoded by alpha satellite DNA, with multiple alpha-satellite variants and both strands capable of forming ORFs. The other two groups are derived from distinct types of Sat2/3: one is the DYZ1 satellite on chrY (a specialized form of Sat2/3), and the third—and by far the most abundant—is a specialized Sat2/3 variant located primarily on chr1q12. Its canonical isoform is 127 aa in length (corresponding to the spike in Supplementary Fig. 1B) and contains imperfect periodic repeats. Because its sequences frequently begin with MESSS, we named this family **MESSS** (also meaning "*more than a mess*"). To distinguish among the numerous MESSS

isoforms, we adopted the nomenclature "MESSSm_n", where m is the chromosome number and n is the copy number in YAO.

Transposable elements constitute the most abundant repeat class in the human genome, and both LINEs and SINEs are present in very high copy numbers. However, transposable element regions do not contain high-copy-number istORFs; the highest copy number observed is only a few dozen. Transposable elements are ancient sequences: LTR and non-LTR retrotransposons are distributed across all major eukaryotic lineages (animals, fungi, plants, and SAR) and possess conserved reverse transcriptase domains, tracing their origins to the last eukaryotic common ancestor. DNA transposons show a similar pattern. The istORFs in transposable element regions are diverse and divergent.

In contrast to transposons, satellite DNA is remarkably young. Most satellite sequences in the human genome are specific to primates (or even narrower clades, such as Old World monkeys). This fundamental difference in origin dictates that their coding modes, if any, must be very different. Satellite DNA can generate large arrays of tandem repeats through concerted evolution, enabling identical or nearly identical ORFs to reach extremely high copy numbers. Thus, although dark matter proteins may originate from various sources, the high-copy-number istORFs are predominantly satellite-derived, suggesting that satellite-encoded proteins are more likely to be expressed and functional, and should therefore be prioritized for investigation.

We also examined other species. Although the highest-copy-number ORFs in every genome are satellite-derived, the most abundant ORFs differ markedly across species. Alpha satellite DNA has high copy numbers in Old World monkey genomes, where it is primarily located in centromeric regions and plays important roles in centromere structure and sister chromatid segregation during mitosis. Our data show that alpha satellite forms ORFs in multiple primate species, with both strands exhibiting coding potential, and that these ORFs display some degree of conservation (Supplementary Table 3). This raises the possibility that alpha satellite DNA may encode peptides or

proteins in addition to its well-established roles at the DNA level.

In the bonobo genome, the most abundant istORFs derive from a complex two-unit satellite DNA, whose two strands together form four distinct ORFs in both orientations, predominantly located on acrocentric chromosomes. This satellite family is ancient, already present in Old World monkeys, and shows an increasing trend within the hominine lineage (Supplementary Table 3), reaching its extreme in bonobo, where its copy number far exceeds that in humans and chimpanzees.

Interestingly, despite the close relatedness of chimpanzees and bonobos, their high-copy-number ORF repertoires are very different. Chimpanzees possess three high-copy-number ORFs that are either absent or extremely rare in bonobos and humans (Supplementary Table 3).

When comparing primate—and particularly hominid—genomes based on conventional protein-coding genes, the differences appear minimal, because these genes originated very early in evolution; even human and mouse share ~95% similarity by this metric. Thus, comparisons based on conventional genes substantially underestimate the differences among hominid genomes. To identify the key differences among them—and to address questions such as "what makes us human"—satellite proteins, and perhaps other dark matter proteins, may deserve far greater attention. A recent study that sequenced near-T2T genomes of eight lemur species also highlighted rapid centromere evolution and its role in speciation [16].

Comparison between genomes of different species based on isteins

To quantitatively compare istORF repertoires across genomes, we calculated Jaccard distances between species. Given that DNA analysis commonly uses 21-mers, we chose 7-mer peptides for amino acid sequences, as 7-mers provide sufficient complexity while remaining computationally tractable. We first applied this approach to annotated proteins (Fig. 3A). The result showed 98.32% similarity between human

and chimpanzee, consistent with previous reports [3], confirming that our method performs as expected. We then applied the same analysis to isteins predicted from T2T genomes (Fig. 3B). The differences between species increased substantially. Importantly, the phylogenetic hierarchy remained intact—for example, ANI (CHM13, GorGor) ≈ ANI (PanTro, GorGor), and ANI (YAO, SymSyn) ≈ ANI (PonAbe, SymSyn)—and the topology of the phylogenetic tree was unchanged (Supplementary Fig. 5).

Clearly, istein-based comparisons better reflect genomic differences between species. We do not claim that this approach is necessarily more "biologically accurate" than traditional comparisons; rather, it provides a different lens through which to examine genomes and uncover previously overlooked information.

Next, we searched for istORFs that may have been subject to positive selection during human evolution. We extracted 7-mers that show increasing abundance along the lineage from orangutan to gorilla to chimpanzee to human (Fig. 3C). Remarkably, the top-ranking 7-mer corresponded to MESSS (shown in red), suggesting that MESSS represents a major genomic difference between humans and other great apes, and may have played a role in human evolution. The green 7-mers correspond to a nearly human-specific satellite DNA, encoding isteins of even lower complexity than MESSS, rich in tyrosine, and predicted by AlphaFold to form a beta-barrel structure (File: SNIDYIL_54). The black 7-mers correspond to another Sat2/3-derived istein family that is more similar to classical Sat2/3-encoded sequences, with increased serine content—showing a similar evolutionary trend to MESSS, but with less dramatic divergence between humans and other hominines.

The major differences among closely related species lie in satellite DNA and short tandem repeats, as conventional protein-coding genes are highly similar. Even within a species, differences between populations and individuals are even more dependent on these regions. The recent maturation of T2T technologies has allowed us to probe the information hidden in satellite DNA more deeply [16, 17]. If satellite DNA is

indeed capable of encoding proteins, this may provide new explanations for physiological differences and disease susceptibilities among geographically distinct human populations.

MESSS is the most abundant ORF family in the human genome

We extracted all MESSS ORFs from the YAO genome based on sequence features and positional information, identifying a total of 16,775 ORFs (7,329 unique isteins; 8,487 unique istORFs; Supplementary Table 4; Files: MESSS_istein, MESSS_istORF). The vast majority—12,379—are located on chr1q12. An additional 2,094 are found on chr16, with smaller numbers scattered across other chromosomes. Chimpanzees, bonobos, and gorillas also possess MESSS ORFs, but only a few hundred copies each (Files: MESSS). We constructed a phylogenetic tree of MESSS sequences from these four species (Fig. 4A). On most branches, human and great ape MESSS sequences are intermingled; however, one cluster (highlighted in red) represents a rapidly evolving lineage that has undergone massive amplification specifically in humans. This cluster is located on chr1q12, contains the majority of MESSS copies, and is human-specific. We refer to this group as MESSS1, which constitutes one of the largest HORs in the human genome. chr1q12 is one of the most evolutionarily active regions of the human genome and also harbors the DUF1220 gene cluster—a primate-specific locus implicated in brain development and cognitive expansion [18]. Taken together, these observations suggest that MESSS may have contributed to human evolution, potentially playing a role in the development of the human brain.

AlphaFold predictions show that the classical Sat2/3 heavy-chain-encoded protein adopts a beta-barrel structure (Supplementary Fig. 6A–B), whereas the chr16-encoded MESSS is more loosely structured (Fig. 4B). The highly amplified chr1-encoded MESSS variants are even more disordered (Fig. 4C), suggesting that this structural loosening may have enabled MESSS to acquire new functions. The MESSS family is

diverse and abundant, forming a continuum of related structures (Supplementary Fig. 6C).

In the hg38 genome, only a few hundred MESSS copies can be found, underscoring the necessity of T2T assemblies for studying dark matter proteins—especially high-copy-number satellite-derived ORFs. This also helps explain why MESSS and similar ORFs had gone unnoticed for so long.

Evolution of Sat2/3 in Hominidae

Given the importance of satellite ORFs in this study, we evaluated the overall repetitiveness of T2T genomes using a previously described method [14]. As expected, the human T2T genome is more repetitive than hg38 (Fig. 5A), reflecting the inclusion of previously unassembled repeat-rich sequences. Among great apes, however, the human genome is the least repetitive. Gibbons and macaques show higher repetitiveness than great apes, and the mouse genome is more repetitive than all primates examined. The Y chromosome is consistently the most repetitive chromosome in each genome. Since the mouse embryonic stem cell line used lacks a Y chromosome, mouse repetitiveness is not inflated by Y-linked sequences; even excluding chrY, the mouse genome remains substantially more repetitive than primates.

Sat2/3 constitutes the most abundant source of high-copy-number ORFs, prompting a closer examination. Sat2/3 is a primate-specific satellite DNA, with the heavy (purine-rich) strand having the canonical repeat unit GGAAT, and the light (pyrimidine-rich) strand ATTCC. In mouse, only scattered Sat2/3-like sequences exist and do not form tandem arrays. The heavy strand encodes characteristic peptides such as MEW and GMEWN, whereas the light strand encodes PFH, IPFHS, and related motifs. We quantified these diagnostic peptides across genomes and found that Sat2/3-derived ORFs increase sharply from monkeys to gibbons to great apes (Fig. 5B–C). A striking asymmetry emerged: in orangutan, gibbon, macaque, and mouse,

light-strand peptides outnumber heavy-strand peptides; in hominines (gorilla, chimpanzee, and human), this ratio is reversed, with heavy-strand peptides vastly outnumbering light-strand peptides. Since the total amounts of light and heavy strands are equal, this reversal reflects a lineage-specific shift in which the heavy strand acquired more long ORFs in hominines, while the light strand lost them.

Although the number of classical Sat2/3 istORFs in humans is lower than in chimpanzees and gorillas, the human genome retains a substantial number of such ORFs, particularly near the chr9 centromere, where they are densely concentrated (Fig. 1C)—in fact, outnumbering MESSS copies, though lacking the higher-order repeats characteristic of MESSS. Complexity analysis reveals that the human chr1q12 region has undergone substantial HOR expansion relative to bonobo and gorilla (Supplementary Fig. 7), and this region contains the ~6 Mb MESSS-encoding locus. The human chrYq, in contrast, is composed almost entirely of Sat2/3. Even closely related species show profoundly different Y chromosomes, each with a distinct HOR profile (Supplementary Fig. 7). We compared the amino acid compositions of classical chr9 Sat2/3-encoded isteins, the highly amplified chrY-encoded MEW99, and chr16- and chr1-encoded MESSS variants (Supplementary Fig. 8A). Classical Sat2/3 isteins closely resemble the theoretical amino acid composition of the Sat2/3 heavy chain. MEW99 is characterized by increased lysine (K) content. MESSS variants show increased serine (S) content; however, the human-specific chr1 MESSS does not have appreciably more serine than the chr16 ancestral form, but is instead distinguished by a marked increase in arginine (R).

We also examined the nucleotide composition of these istORFs (Supplementary Fig. 8B). Relative to classical Sat2/3 istORFs, MEW99 has increased A content, whereas MESSS has acquired substantial C and T at the expense of G. Consequently, the light strand of MESSS regions has gained many G residues, increasing stop codons on the reverse strand and abolishing the long istORFs that are characteristic of bidirectional coding in classical Sat2/3. The purine-to-pyrimidine ratio in MESSS regions has shifted from ~3:1 (heavily strand-asymmetric) to ~3:2 (more balanced), which may

affect DNA properties and transcriptional activity.

Previous work reported that Sat3 exists in two types, with hominines possessing both type I and type II, whereas other primates have only type II [15]. The emergence of type III Sat3 may have contributed to the light-to-heavy strand switch and to the origin of MESSS in hominines. The human-specific MESSS expansion involved an additional step: the introduction of more C residues into the heavy strand, increasing serine content and potentially improving solubility.

Alpha satellite DNA is specific to Old World monkeys and is the most abundant satellite DNA in the human genome. We observed high-copy-number istORFs in alpha satellite regions across all primate T2T genomes examined, suggesting that alpha satellite also warrants further investigation, although we have not yet pursued this direction in the present study.

Conserved isteins/istORFs across species

In contrast to the highly divergent satellite-derived isteins, we also asked whether our predictions included conserved, previously unannotated proteins. To obtain a rapid but stringent assessment, we devised a conservative screening strategy based on shared 7-mer peptides across species. For each taxonomic group, we first selected 7-mers that are present exactly once in each species' istein set and are completely absent from the CHM13 and hg38 annotated proteomes. We then extracted the corresponding isteins, and retained only those groups in which the sequences from different species share the same first four and last four amino acids—ensuring that the N- and C-termini are aligned—and finally collapsed identical sequence groups. This yielded a set of conserved isteins for each taxonomic level (Fig. 6A).

Using this approach, we identified 180,427 conserved isteins within African apes (CHM13, PanPan, GorGor). Expanding the scope to great apes (CHM13, PanPan, GorGor, PonAbe) reduced the number to 54,950; within apes (adding SymSyn) to

20,837; within Old World monkeys (adding MacFas) to 5,437; and when including mouse (Euarchontoglires: CHM13, PanPan, GorGor, PonAbe, SymSyn, MacFas, MusMus), the number fell to 179 (Supplementary Table 5, File: Conserved_istein_ORF).

For each conserved group, we extracted the CHM13 representative sequence and annotated its genomic location using the CHM13 GTF file (GCF_009914755.1_T2T-CHM13v2.0_genomic.gtf). As the taxonomic scope expanded, the number of conserved isteins dropped sharply, while the proportions located in gene regions, protein-coding genes, and exons increased correspondingly (Fig. 6B–C, Supplementary Table 6, File: Conserved_Hominidae_annotation).

We focused on the 179 isteins conserved across all seven genomes, including mouse. BLASTP searches confirmed that these isteins have no annotated counterparts in the human proteome; a small number matched non-primate proteins, suggesting that they represent more broadly conserved genes that have been missed in human annotation. Over 98% of these istORFs are located within gene regions, with >80% in exons. This is not surprising, given that the common ancestor of primates and rodents lived approximately 90 million years ago; sequences that remain highly conserved to this day are almost exclusively the coding regions of essential proteins. These istORFs arise primarily through overlapping coding (in both sense and antisense orientations) and intronic read-through. The host genes include SOX1, FOXD2, NFIX, AHDC1, JUN, and multiple HOX genes—almost all are well-known “star” genes (Supplementary Table 6). Only three of the 179 istORFs are located in intergenic regions; notably, all three overlap with regions annotated in the Fetal Gene Atlas binned by cell type (Cao et al. 2020 [19]), suggesting that these previously uncharacterized intergenic regions may contain important coding information.

We further calculated dN/dS ratios for these 179 istORFs. Fig. 6D–E shows that the majority have experienced significant purifying selection, with roughly half exhibiting dN/dS < 0.3 (Supplementary Table 7). Such deep conservation over an evolutionary

timescale of nearly 100 million years strongly suggests that these istORFs are genuine protein-coding regions, even though their mRNA structures remain unknown. It also implies that, even within Euarchontoglires, there may still be hundreds of conserved protein-coding genes awaiting discovery. A small subset of these has been reported in OpenProt; the convergence of independent methods on the same targets further supports the reliability of our approach. As we expand the evolutionary scale, the number of conserved istORFs increases substantially. Even within Homininae (~10 million years of divergence), maintaining intact and conserved ORFs suggests protein-level selective pressure; at this level, we identified >100,000 conserved istORFs, of which 32.8% are intergenic and 46.2% are intronic. Notably, these numbers were obtained using very stringent parameters. Between the extremes of ultra-conserved ORFs and young, rapidly evolving ORFs (such as satellite ORFs), a vast intermediate landscape remains largely unexplored. Although we cannot yet estimate the true total amount of coding sequence with confidence, it seems increasingly likely that the unannotated coding regions in the human genome may exceed the currently annotated ones (~2%).

Beyond these single-copy, non-repetitive conserved genes, some tandem-repeat isteins (Supplementary Table 2) are also highly conserved. However, because their analysis is considerably more complex than that of non-repetitive isteins, we have not yet performed a systematic study of them.

Proteomics mass spectrometry evidence for istein expression

To detect istein-derived peptides at the protein level, we searched seven proteomics datasets—including oocyte and embryonic stem cell samples—against the 1,000 highest-copy-number isteins using Pepquery. At least 415 isteins had one or more confident peptide matches in at least one experiment (Fig. 7A, File: protein_explorer). Among the top 1,000, the most abundant family was MESSS (172 unique sequences, 3,536 total copies). We therefore examined MESSS specifically

and found that it is highly expressed in oocytes, consistent with the large stores of RNA and ribosomes in these cells, which support active translation. MESSS peptides were also detected in several cancer tissues (Fig. 7B).

We extracted 5-mer peptides from these confident matches and removed those present in the annotated proteomes (hg38 and CHM13), yielding a set of dark-matter-specific 5-mers. We then examined the relationship between these 5-mers and the genome-wide distribution of dark-matter-specific 5-mers in isteins. Supplementary Fig. 9A shows no obvious correlation, which is not unexpected, as we only searched the top 1,000 isteins, and expression levels vary widely across different istORFs. We also examined the correlation between these 5-mers and those found specifically in MESSS, and observed a clear relationship (Supplementary Fig. 9B). This is also not surprising, given that MESSS is the most abundant family in the top 1,000. Nevertheless, because these 5-mers are absent from all annotated proteins, the simplest interpretation is that satellite ORFs are transcribed and translated into polypeptides or proteins. Conversely, if we maintain the traditional view that satellite DNA does not encode proteins, we would need alternative explanations for the origin of these 5-mers.

We also searched public proteomics datasets for conserved isteins. Among the 179 isteins conserved from primates to mouse, 90 had at least one peptide match in 10 mass spectrometry libraries. Among the 5,437 isteins conserved across six primate species, 1,764 had at least one peptide match in 9 libraries. These are both high validation rates, warranting further in-depth investigation (Fig. 7C–D, File: protein_explorer).

Ribo-seq evidence for istORF translation

Because we did not account for RNA splicing, many of the istORFs we identified may be fragmented or incomplete in the actual RNA molecules. We therefore anticipated that alignment of istORFs to ribo-seq data might be suboptimal. Nevertheless, we

performed the analysis. We downloaded four ribo-seq datasets (oocyte, 293T, MDA-MB-231, and heat-shocked cells) and aligned them against the istORFs corresponding to the top 1,000 most abundant isteins. A total of 401 istORFs (40.1%) showed evidence of translation in at least one experiment (Fig. 8A). As in the mass spectrometry results, oocytes showed the highest number of positive istORFs.

For conserved genes, many are located in conventional coding regions (exons and introns), making it difficult to distinguish their translation signals from those of known genes. We therefore focused on the 727 istORFs that are conserved in Old World monkeys and located in intergenic regions. When aligned against the same four ribo-seq datasets, 270 (37.1%) showed translation evidence in at least one experiment (Fig. 8B).

We then compared the mass spectrometry and ribo-seq results. Among the top 1,000 istORFs, 161 (16.1%) had evidence from both types of experiments (Fig. 8C). Of these, MESSS accounted for 49 (4.9%) and was the predominant family with dual evidence (Fig. 8D). Among the intergenic Old World monkey-conserved istORFs, 27.8% had evidence from both approaches (Fig. 8E), consistent with our expectations for conserved proteins.

We also isolated ribosomal fractions from MDA-MB-231 and HeLa cells and performed RT-PCR using primers specific to MESSS1_446. In both cell lines, the polysome fractions showed strong signals, suggesting that MESSS is being translated in these cells (Fig. 8F–G).

Biochemical detection of MESSS expression

To further validate MESSS expression at the biochemical level, we designed and synthesized peptides based on the MESSS amino acid sequence (Supplementary Fig. 10A) and immunized mice to generate polyclonal antisera. Among the resulting antisera, HNM3-2 showed the best performance (Supplementary Fig. 10B). We first

expressed recombinant MESSS in *E. coli* BL21(DE3) using a codon-optimized coding sequence cloned in pET15-b that preserves the amino acid sequence of MESSS1_446. Western blot with HNM3-2 confirmed that the antiserum specifically recognized the bacterially expressed MESSS (Supplementary Fig. 10C). We then used this antiserum to detect MESSS expression in multiple cell lines. Fig. 9A shows that MESSS was detected in several cell lines, with variable band sizes. This is consistent with expectations, as MESSS exists in many isoforms, has high copy numbers, and contains tandem repeat structures that may give rise to heterogeneous protein products. The cell lines tested included those from pig, dog, monkey, and mouse. Species outside Homininae lack MESSS ORFs, yet we detected bands in pig kidney PK15, canine kidney MDCK, and mouse fibroblast L929 cells. This may reflect cross-reactivity of the antiserum with similar epitopes in these species, or the presence of peptides with sequence similarity to our target peptide.

We also overexpressed MESSS1_446 in 293T cells. Anti-Flag Western blot confirmed successful expression (Supplementary Fig. 11A), but when probed with HNM3-2, the band position differed from that detected by anti-Flag (Supplementary Fig. 11B), even though the predicted amino acid sequence was identical to that of the bacterially expressed construct. This discrepancy may be due to antibody performance issues, or to post-translational modifications or conformational masking of the epitope in the eukaryotic expression system. The ~100 kDa band may represent a multimer or a spliced variant of MESSS in which the epitope is exposed. CCK-8 assays showed no difference in cell viability or proliferation between MESSS-overexpressing cells and wild-type controls (Supplementary Fig. 11C). These results at least indicate that MESSS can be stably expressed in human cells without apparent toxicity.

Since the polysome fractionation experiments suggested active MESSS translation in MDA-MB-231 cells, we performed RNA interference experiments in this cell line. We first tested siRNA efficiency in 293T cells; siMESSS1-1 and siMESSS1-2 showed good knockdown, while siMESSS1-3 was ineffective (Supplementary Fig. 11D). We therefore used siMESSS1-1 and siMESSS1-2 for knockdown in MDA-MB-231 cells.

At 48 hours post-transfection, the knockdown groups exhibited obvious cell death under light microscopy, which was confirmed by flow cytometry and colony formation assays (Supplementary Fig. 12). This suggests that MESSS may promote cell survival in MDA-MB-231 cells. Western blot confirmed reduced MESSS bands in the knockdown groups (Fig. 9C–D). Immunofluorescence with the same antiserum showed that MESSS is predominantly cytoplasmic, with slightly reduced fluorescence in knockdown cells and a marked increase in bright puncta (Fig. 9E).

To further investigate the tissue expression profile of MESSS, immunohistochemistry (IHC) was performed on normal human liver, lung, and small intestine tissues, as well as ovarian cancer specimens. In healthy tissues, MESSS exhibited robust, cell-type-specific immunoreactivity within barrier and mucosal epithelia; strong cytoplasmic and membrane staining was observed in mature enterocytes of the small intestine, and prominent apical and cytoplasmic signals were concentrated in the bronchial pseudostratified ciliated epithelium. In contrast, normal hepatocytes and infiltrating inflammatory cells in the liver showed negligible basal expression, whereas intense, granular cytoplasmic staining was observed exclusively within focal clusters of parenchymal hepatocytes under apparent stress. Additionally, ovarian cancer tissues displayed a heterogeneous and relatively weak-to-moderate expression pattern with distinct punctate cytoplasmic signals within specific tumor cell subpopulations. Taken together, these initial IHC observations demonstrate the endogenous translation of MESSS and reveal its highly specific spatial distribution across various physiological and pathological tissue contexts (Supplementary Fig. 13).

Collectively, these results provide preliminary biochemical evidence for MESSS expression, although they are not yet definitive. The main limitations are the suboptimal specificity and titer of the antiserum, likely due to the low complexity of the MESSS sequence, which may result in weak immunogenicity. This may be a common challenge in future studies of satellite-encoded proteins.

Misannotation of human satellite sequences as bacterial hypothetical proteins

One additional observation merits attention. When searching the NCBI databases, we found that many of the high-copy-number genomic sequences we identified—which are not annotated in the human genome—are instead annotated as hypothetical proteins in various bacteria (normal flora or pathogens) and parasitic organisms. Supplementary Table 8 lists a few examples, but the actual number is far larger, and we have not yet performed a systematic survey. These sequences, such as Sat2/3, are well-established primate-specific satellite DNAs that exist as contiguous arrays in primate genomes. In bacteria, by contrast, their distribution is patchy: some strains of a given species contain the sequence while others do not. Given the habitats of these organisms and the fact that they are handled by humans, their samples are highly susceptible to contamination with human DNA. The fact that the "bacterial proteins" identified in this way are almost exclusively high-copy-number satellite isteins—rather than single-copy isteins—further supports a contamination origin, as high-copy sequences are more likely to be detected in contaminated samples. Moreover, many bacterial genome assemblies are generated in a largely automated manner and are far less accurate than human T2T genomes. Thus, the most parsimonious explanation is that these sequences represent human genomic sequences misannotated as bacterial genes, rather than the reverse. Because bacteria are generally thought to lack junk DNA, automated annotation pipelines confidently assign open reading frames as hypothetical proteins—whereas in human or other mammalian genomes, the same sequences remain unannotated due to more stringent annotation criteria. This creates an ironic situation: potential coding sequences in the human genome are absent from human annotation but are instead annotated as hypothetical proteins in bacteria.

An alternative explanation is that some of these sequences may have been acquired by bacteria or parasites from their hosts through horizontal gene transfer (HGT). However, the observed distribution pattern does not show the typical features of HGT. In the absence of strong evidence for HGT, we consider contamination the more likely

explanation. Although we have not systematically demonstrated that these sequences are indeed coding regions in the human genome, we suggest that at minimum, bacterial genomes should be cleaned of such misannotations, and that future bacterial genome assemblies should be screened against human T2T genomes to remove human-derived sequences unless clear evidence of horizontal transfer exists. These misannotations caused considerable confusion and misled us during the early stages of this study. We hope that sharing this observation will help colleagues avoid similar pitfalls.

Discussion

Proteomic dark matter is fundamentally different from genomic or transcriptomic dark matter. With today's powerful sequencing technologies, there is almost no truly invisible dark matter at the nucleic acid level—sequences may have unknown functions, but they are detectable. Protein-level dark matter, however, is truly dark. The discovery and identification of dark matter proteins differ from those of non-coding RNAs. Nucleic acid sequencing directly yields sequence reads; as long as sequencing technology is sufficiently advanced, researchers will naturally see large amounts of RNA beyond the conventional categories (rRNA, tRNA, mRNA). Protein mass spectrometry does not output a sequence file; it requires a library of predicted peptide sequences for retrieval. Prediction thus becomes a bottleneck in the identification of unknown proteins. The conventional transcript-centric approach is limited, on the one hand, by incomplete annotation of RNA transcripts, and on the other hand, by the fact that RNA-seq analysis has largely been performed using hg38 as the reference genome, which lacks large amounts of dark matter sequences—reads derived from these regions are simply discarded because they fail to align. Recent large-scale studies have shown that the yield from transcript-based screening and validation of ncORFs has become very low [8], highlighting the urgent need for new approaches to discover dark matter proteins. Our *in silico* translation approach opens

a door, allowing us to see ORFs that we had previously not even been aware of but that may actually be expressed.

Translating genomic DNA directly into amino acid sequences, without considering transcription and splicing, is a quick-and-dirty method. It certainly does not reflect the actual expression process in the cell, but it enables rapid identification of DNA sequences with coding potential that have never been reported before. Combined with mass spectrometry and ribo-seq data, we can filter for sequences with high likelihood of expression for further investigation. For proteins of interest, one can then go back and use long-read RNA-seq and RACE to identify their RNAs. Our goal was not to precisely report one or a few unknown proteins in one step, but rather to reveal long-overlooked potential coding regions and make it possible for others to discover them. When we opened this door, however, we were flooded with so much information that this paper covers too many aspects, most of which remain at the hypothesis level. We hope to present as much as possible to readers so that interested colleagues can begin their own explorations as early as possible.

The positive rates in our searches were unexpectedly high (Fig. 7–8). In fact, many conventional proteins are difficult to detect in proteomic datasets [20], yet we downloaded only 10 proteomic datasets. Indeed, some isteins that were not identified in our high-throughput search could still be found via the Pepquery web interface, suggesting that expanding the search scope would likely yield even higher positive rates. Theoretically, unbiased genome-wide translation should have a lower probability of finding coding regions than transcript-based mining. However, several factors give our approach an advantage over conventional methods: (i) known transcripts have already been mined repeatedly, leaving little room for new ORF discovery; (ii) many transcript annotations are built upon the hg38 assembly, which is incomplete for repetitive regions; and (iii) we prioritized high-copy-number or highly conserved ORFs for validation, greatly reducing computational cost and improving the positive rate.

Among the dark-matter ORFs we screened, satellite DNA-derived ORFs were the most striking. Some have extremely high copy numbers, and to our knowledge, their coding potential has never been reported. We provide multiple lines of evidence—ordered from weaker to stronger—that at least some satellite ORFs are expressed: (i) extremely high copy numbers of certain satellite ORFs; (ii) conservation across species (e.g., MESSS is present in all hominine species); (iii) ribo-seq alignments; (iv) peptide matches in proteomic datasets; and (v) detection by Western blot, immunofluorescence, and immunohistochemistry. For MESSS, we have obtained all of the above lines of evidence, providing considerable confidence that it is indeed expressed. To our knowledge, this is the first study to show that satellite DNA can encode proteins or peptides, and that these ORFs can reach exceptionally high copy numbers. Although we cannot yet present a complete and systematic picture of dark matter protein expression, we have revealed a direction that is well worth pursuing.

The exceptionally high copy numbers of satellite ORFs such as MESSS deserve serious attention. The combined coding region of the MESSS family (~6 Mb) exceeds 10% of the total conventional coding sequence in the human genome. Even if only a small fraction is expressed, its impact cannot be ignored. The ancestral form of MESSS emerged in Homininae, and the novel MESSS underwent a burst in the human lineage, suggesting that MESSS may have contributed to human evolution, particularly to the expansion of brain size and cognitive capacity that distinguishes humans from other great apes. The regions where MESSS is concentrated—chr1q12, chr10q11.21, and chr16q11.2—have been linked to brain disorders such as schizophrenia and autism, further supporting this hypothesis [21–23]. The higher incidence of Alzheimer's disease and autism in humans compared to great apes may also be related. Of course, these hypotheses await further investigation.

Beyond the young satellite-derived istORFs, we also examined the opposite end of the spectrum—the ultra-conserved istORFs—and found that even between human and mouse, a considerable number remain to be discovered, with evidence of translation, and some are conserved even across vertebrates. In between these two extremes—

primate-conserved, ape-conserved, hominid-conserved, hominine-conserved—there lies a continuum of istORFs waiting to be explored. Thus, 6-frame in silico translation is a highly efficient approach. Rather than being confined to the traditional RNA-centered discovery pipeline, we can directly predict amino acid sequences from DNA, rapidly identify candidate proteins of interest, and then proceed to validation. Our data also suggest that dark matter proteins may outnumber conventional proteins (~20,000) in terms of distinct species, although their functional importance may be lower on average.

It was once thought that overlapping coding was mainly a feature of viruses, prokaryotes, and organelles such as mitochondria. However, our results on conserved isteins suggest that the frequency of overlapping coding in the human genome may have been underestimated, and that many such overlapping ORFs occur within highly important protein-coding genes. It is possible that some important genes form locally compact regulatory modules within the vast genome—a compact structure that not only saves space and facilitates precise regulation but also prevents the disruption of coding regions by chromosomal rearrangements that could scatter them across different loci.

In summary, our study provides an efficient approach for mining previously unannotated coding regions in the genome, and offers initial evidence that some sequences long assumed to be non-coding—such as satellite DNA—not only can encode proteins but may constitute a significant component of the human proteome. These findings also suggest that the difference between genomic regions in terms of translation is not a binary yes/no, but rather a continuum of expression levels and regulatory precision; the entire genome has the capacity for transcription and translation, with excess and non-functional RNAs and proteins being degraded—a more parsimonious model than one that assumes perfect regulation of transcription and translation.

Hypotheses

A hypothesis on the burst-like evolution of satellite DNA and the birth of new genes

Current gene annotation practices suffer from a systematic bias. We tend to equate "genes" with mature coding units that are broadly conserved across major taxonomic groups, equipped with well-defined regulatory elements, and readily detectable at the protein level by Western blot. Under this standard, the vast majority of sequences considered "true genes" have homologs in the mouse genome. Given that the common ancestor of primates and rodents lived approximately 90 million years ago, our current gene annotation essentially reflects a collection of genes that had already largely taken shape at least 90 million years ago. Yet evolution has never ceased. The birth of new genes should be an ongoing process, not something that stopped at some ancient node. A gene requires tens of millions of years of evolution to transition from a newly emerged orphan to a gene shared at the family or order level. If we acknowledge only the end products of evolution while ignoring the intermediate stages, our definition of a gene is inherently incomplete.

From an evolutionary perspective, orphan genes and "star" genes are not fundamentally different—they merely occupy different positions along a continuous spectrum. An orphan gene, if it is "fortunate" enough to accumulate beneficial mutations and regulatory elements over a sufficiently long period, may eventually become a widely conserved star gene. The ancient genes that we now see distributed across all major eukaryotic lineages were once orphan genes in some distant lineage. Just as every elderly person was once young, every star gene was once an orphan. This implies that a species must survive over an extended evolutionary timescale without being eliminated in order for its orphan genes to gradually spread to the level of genus, family, order, or beyond—and even horizontal gene transfer requires considerable time for dissemination. Over this long process, the vast majority of orphan genes are lost; only a tiny fraction survive and eventually "succeed."

Therefore, we should not exclude an ORF from being a gene simply because it lacks an ideal promoter, poly(A) site, or high expression level. What we may be observing is instead a snapshot of a young gene in the process of evolving toward maturity.

Satellite DNA can rapidly generate ORFs of various lengths and reading frames through replication slippage and unequal crossing-over, acting as a “master key” that can quickly assemble different coding sequences, making them ideal “incubators” for the continuous birth of new genes. A newly born gene is unlikely to immediately acquire perfect regulatory elements such as promoters, enhancers, poly(A) sites, or splicing signals. The tandem amplification characteristic of satellite DNA compensates for this “quality” deficit: a newly emerged ORF may have very low expression efficiency, but if it is amplified to hundreds or thousands of copies in the genome, its total macroscopic expression can reach a considerable level, providing raw material for natural selection.

But where can these primitive polypeptides and proteins—which almost certainly lack specific functional domains and precise catalytic activities—go to be tested by evolution? The answer may be the intracellular matrix. Newly born proteins lack various localization signals; the default path is to be released directly into the cytoplasm upon translation. The intracellular matrix, while important, is arguably the least demanding environment in the cell. Its functions are physical—maintaining macromolecular crowding, providing weak interaction interfaces, and mediating phase separation—rather than chemical (such as enzymatic catalysis or specific binding by transcription factors). This means that matrix proteins have low sequence stringency; they only need to maintain certain macroscopic properties, such as amino acid composition, chain flexibility, or net charge distribution. This “low-threshold” characteristic makes the matrix an ideal “apprenticeship workshop” for new genes—nascent, functionally crude, sequence-imperfect proteins can gain an evolutionary foothold here without being immediately eliminated. This may also explain why we have long lacked a deep biochemical understanding of the “cytosol”—what exactly is dissolved in it? It is unlikely that the total protein concentration of 200–300 mg/mL in

cells is maintained solely by precisely regulated enzymes, signaling molecules, and structural proteins. What is this crowded “background”? The contrast with our precise understanding of albumin in extracellular fluids is striking. Perhaps it is precisely because a large fraction of the intracellular matrix is composed of dark matter proteins that they have remained invisible to us.

Transposable elements and satellite DNA are often mentioned together, but they are fundamentally different. Transposable elements are ancient sequences that have existed throughout eukaryotic history; satellite DNA, by contrast, is mostly very young. If satellite DNA is indeed very young—for example, most major satellite families in the human genome are specific to Old World monkeys, Homininae, or even humans—then where are the older satellites? The answer is that they have already disappeared, leaving at most some remnants. Lemurs and other strepsirrhines lack alpha satellite DNA, and even closely related species can have vastly different centromeres. We know what the ancestral ribosome looked like, but we can hardly infer what the ancestral centromere looked like—all because satellite DNA evolves too fast.

Satellite DNA derives its name from its electrophoretic banding pattern, but if we use an astronomical analogy, satellite DNA is not like a satellite—it is more like a supernova: it expands explosively, amplifies rapidly, but has a short lifespan. Thus, the life cycle of a typical satellite DNA family follows a spindle-shaped trajectory on a timescale of tens of millions of years: (i) birth of the ancestral sequence; (ii) massive tandem amplification, accompanied by the expansion and diversification of various ORFs; (iii) peak expansion, with continued ORF divergence; (iv) collapse, with copy number reduction and some ORFs becoming increasingly complex, occasionally acquiring regulatory elements such as promoters; and (v) extinction, replaced by emerging satellite DNAs—but some ORFs with clear beneficial functions are retained, becoming young genes at the family-to-order level.

From this perspective, Sat2/3 may currently be at its peak, with abundant ORF diversity and copy numbers, yet even sequences like MESSS remain relatively simple

and repetitive, without complex ORFs having yet emerged. Moreover, Sat2/3 has generated vast numbers of 5-mers such as GMEWN, yet none of these are found in annotated genes—no 15-mer DNA sequence from Sat2/3 has yet been translocated in-frame into a mature gene's ORF. The simplest explanation is that there has not been enough time: Sat2/3 is still young, and there remains a clear barrier between it and conventional coding regions; its distinctive k-mers have not yet spread. Genes such as DUF1220 and NOTCH2NL, although they also display repetitive features and are located near chr1q12, are clearly not products of Sat2/3—they are older and share no sequence homology. They may be remnants of an earlier "supernova" explosion. It is possible that descendants of MESSS could one day give rise to genes resembling them.

Overall, genome evolution is highly complex. On one hand, important genes are extremely conserved, some with histories as long as or longer than that of eukaryotes themselves, and they possess modular regulatory precision. On the other hand, "supernova" bursts (satellite DNA) create local hotspots on chromosomes, generating vast numbers of ORFs in a short time, a few of which will eventually disperse across the genome to become the new generation of genes.

A hypothesis on the unknown role of the Y chromosome

Given the abundance of istORFs on the Y chromosome, it warrants separate discussion. The human Y chromosome is highly repetitive; except for gorilla, the human Y chromosome is more repetitive than that of any other primate examined, even though the overall repetitiveness of the human genome is lower than that of other primates. The long arm of the human Y chromosome is composed almost entirely of a specialized Sat2/3 repeat, DYZ1, a feature not found in other primates (Supplementary Fig. 7). Given the important role of Sat2/3 in hominine evolution, the Yq region should be taken seriously rather than dismissed as junk DNA. Traditional studies have suggested that the Y chromosome contains few genes and is

degenerating. However, if these ORFs are taken into account, the Y chromosome may not be as gene-poor as previously thought, and its evolutionary rate is much faster. The Y chromosome is not only the most divergent chromosome between hominid species, but also the most variable chromosome among human males—for example, the Y chromosome of CN1 is 27% longer than that of YAO, even though both are from Chinese individuals. The conventional view regards Y-linked sequences largely as junk DNA, but from an evolutionary perspective, maintaining such a large amount of sequence without using it would be a disadvantage, as it still consumes energy for replication. If some individuals use these sequences to encode useful proteins, this could confer a competitive advantage—particularly in males, where competition is intense—making it more rational to make full use of the Y chromosome.

Beyond the high-copy-number Y-linked ORFs in the top 1,000 list (Supplementary Table 2), we specifically examined MEW99. Using the PepQuery web interface, we searched for its signature peptides across various normal and cancer tissues and found that detectable peptides covered almost its entire sequence (Supplementary Fig. 14). This suggests that a substantial number of Y-linked satORFs may indeed be capable of expressing proteins. The YAO Y chromosome contains 24,571 isteins of ≥100 aa; even if only a small fraction undergoes translational leakage, the cumulative effect could be non-negligible. Compared to other great apes, the human Y chromosome shows the most pronounced HORs, further suggesting that the human Y chromosome may be evolving more rapidly and playing a greater role than previously appreciated. We also used AlphaFold to predict the structures of two Y-linked isteins with the highest copy numbers among those ≥100 aa—one named NAMESKEL (based on its signature peptide) and MEW99. The former resembles the beta-barrel structure of the classical Sat2/3 heavy strand, while the latter is more loosely structured than MESSS (Supplementary Fig. 15).

The Y-linked ORFs may also partially explain why males exhibit greater phenotypic diversity than females. Previously, this was attributed to hormonal effects and the fact that males have only one X chromosome. Our findings suggest that Y-encoded dark

matter proteins may also contribute. Furthermore, for cancers originating from the same organ, male tumors tend to be more aggressive; Y-linked dark matter proteins may also play a role in this disparity.

Questions to be answered by future studies

- What fraction of the human genome is actually translated?
- What is the contribution of dark matter proteins to the proteome—both in terms of the number of distinct species and their total abundance?
- What are the RNA structures of key dark matter proteins? Do they differ from classical mRNAs in features such as 5′ caps, poly(A) tails, or Kozak sequences?
- What proportion of dark matter proteins are translated through non-canonical mechanisms?
- What roles do important dark matter proteins play in development and disease? Can any of them serve as therapeutic targets?

Materials and Methods

Genome data sources

The human T2T genomes, primate genomes, and mouse genome used in this study are listed in Supplementary Table 1 with their file names and download URLs.

In silico translation and extraction of isteins and istORFs

To systematically survey the coding potential of previously unannotated genomic regions, we performed six-frame translation on telomere-to-telomere (T2T) genome assemblies. For each

genome, both strands were translated *in silico* from the first to the last nucleotide. Translation was conducted in all three forward reading frames and three reverse complement reading frames, generating six continuous amino acid sequences per chromosome.

We defined an **istORF (in-silico-translated open reading frame)** as a sequence between the first in-frame start codon (ATG, encoding M) and the first subsequent in-frame stop codon (TAA, TAG, or TGA, encoding Z) in any of the six reading frames. The corresponding conceptual translation product, bounded by M and Z, was termed an **istein** (in-silico-translated peptide/protein) (Supplementary Fig. 1). For each istein identified, the corresponding genomic DNA sequence was defined as an **istORF** (in-silico-translated open reading frame). Only isteins with a minimum length of 100 amino acids were retained for further analysis, as longer polypeptides are more likely to have functional potential, although we acknowledge that shorter isteins may also be biologically relevant.

The T2T genome assemblies analyzed in this study include seven human individuals, five great ape species, as well as gibbon, crab-eating macaque, and mouse (see Supplementary Table 1 for details).

Detection of tandem repeat-encoded proteins

To identify proteins derived from tandemly repeated genomic regions, we developed a pattern-based detection algorithm that operates directly on the amino acid sequence. The method identifies proteins whose sequences are dominated by repeated motifs, without relying on alignment to known repeat databases or sequence homology.

Algorithm Overview

For each protein sequence, the algorithm performs two independent assessments: (1) detection of repetitive patterns through sliding-window k-mer analysis, and (2) identification of the dominant repeat unit and its regularity. A sequence is classified as "repeat-encoded" if it meets a coverage-based or entropy-based criterion (see below).

Pattern Detection

For a given sequence S of length L, all substrings (k-mers) of lengths k ranging from

min_len (default: 5) to max_len (default: 20) are enumerated. For each distinct k-mer, its frequency and all occurrence positions are recorded. A k-mer is considered a candidate repeat if its frequency exceeds:

$$freq \geq \max\left(2, \frac{L \times freq\_threshold}{k}\right)$$

where freq_threshold is set to 0.1 by default. This threshold penalizes longer motifs, requiring higher absolute occurrence counts.

Coverage-Based Classification

For all candidate k-mers meeting the frequency threshold, their occurrence positions are collected and the **coverage** C is calculated as the proportion of sequence positions overlapped by at least one candidate repeat:

$$C = \frac{|\bigcup_{positions}[pos, pos + k)|}{L}$$

A sequence is classified as repetitive if $C \geq coverage\_threshold$ (default: 0.25). This threshold was empirically determined using a training set of known satellite-derived proteins and non-repetitive controls.

Entropy-Based Classification

As a complementary measure, the **pattern entropy** is calculated from the frequency distribution of 2-mers:

$$H_2 = -\sum_i p_i \log p_i$$

where p_i is the frequency of the i-th unique 2-mer in the sequence. Sequences with $H_2 <$ 4.0 exhibit low informational diversity and are classified as repetitive, capturing cases where high-coverage repeats are absent but the sequence remains compositionally simple.

Additionally, if at least **three distinct patterns** are identified as candidates (regardless of coverage), the sequence is also classified as repetitive.

Dominant Repeat Unit

For each sequence, the **dominant repeat unit** is identified as the k-mer that maximizes a scoring function balancing frequency, pattern length, and positional regularity:

$$Score = count \times k \times \frac{1}{\sigma_{interval} + 1} \times \frac{n_{occurrences}}{L/k}$$

where $\sigma_{interval}$ is the standard deviation of the intervals between consecutive occurrences of the k-mer. This scoring function favors motifs that appear frequently, are longer, appear at regular intervals, and are dense relative to the sequence length.

Implementation

The algorithm was implemented in Python (v3.10) using standard libraries (numpy, collections, csv, re). For each input FASTA file, the script generates three output files:

1. A TSV table (*_results.csv) — listing for each protein: ID, length, repetitive status (Yes/No), dominant pattern sequence, pattern length, occurrence count, pattern score, interval standard deviation, 2-mer entropy, and 3-mer entropy.
2. A FASTA file (*_repeats.fasta) — containing only sequences classified as repetitive, with their original headers preserved.
3. A detailed log file (*_analysis.log) — recording per-sequence metrics and a summary of the analysis.

Parameters

All parameters can be adjusted by the user:

Parameter	**Default**	**Description**
min_len	5	Minimum repeat unit length to consider
max_len	20	Maximum repeat unit length to consider
coverage_threshold	0.25	Coverage threshold for repetitive classification
freq_threshold	0.1	Frequency threshold for candidate k-mers

Justification

This approach was chosen over homology-based methods (e.g., RepeatMasker or BLAST against repeat databases) because it does not require reference databases and can detect previously uncharacterized repeat structures, including nested or degenerate repeats typical of satellite-derived open reading frames. The coverage and entropy thresholds were empirically optimized using a manually curated set of known tandem repeat-encoded proteins from human and model organisms.

Quantification of istein sequence complexity

To systematically distinguish satellite DNA-encoded proteins with different structural complexities, we developed a multi-dimensional complexity profiling approach that integrates seven complementary metrics. The overall complexity score ranges from 0 (extremely simple, highly repetitive) to 1 (highly complex, diverse sequence composition).

Metrics

For each protein sequence, the following metrics were calculated:

1. **Periodic pattern coverage** — the proportion of sequence positions covered by short motifs (k = 2–5) that appear at least three times with regular intervals (standard deviation of intervals ≤ 15% of the mean interval). This metric captures locally repeated structural units.

2. **Lempel-Ziv complexity** — a measure of sequence compressibility. A normalized version was used, where the maximal expected complexity is estimated as 0.6 × sequence length. Lower values indicate higher repetitiveness.

3. **Periodicity** — the maximal self-similarity score computed over all possible periods (1–15 aa). This reflects the overall periodic structure of the sequence.

4. **Pattern simplicity index** — derived from k-mers (k = 2–5) that occur at least three times, combining inverse motif length, inverse alphabet diversity, coverage, and occurrence frequency. Higher values correspond to simpler repeating units.

5. **Normalized Shannon entropy** — calculated as the ratio of the observed entropy to the maximum possible entropy given the observed alphabet size, thus controlling for alphabet composition.
6. **Amino acid bias** — the summed frequency of the three most abundant amino acids in the sequence, reflecting compositional skew.
7. **Repetition intensity** — the fraction of the total k-mer count contributed by the top 5% of most frequent k-mers (k = 3). Higher values indicate that a small set of short motifs dominates the sequence.

Composite Complexity Score

The final composite complexity score was obtained as a weighted sum of these metrics, with weights optimized through iterative human-guided feedback using a small set of representative protein sequences (including ultra-simple repeats, medium-complexity repeats, and complex sequences). The optimized weights are:

Metric	**Weight**
Periodic pattern coverage	0.20
Lempel-Ziv complexity	0.15
1 − Periodicity	0.25
1 − Pattern simplicity index	0.25
Normalized Shannon entropy	0.05
1 − Amino acid bias	0.05
1 − Repetition intensity	0.05

All metrics were first converted to a "complexity contribution" direction (higher value = more complex) before weighting. The final complexity score is:

Complexity = 0.20 × periodic_cov + 0.15 × lz_comp + 0.25 × (1 − periodicity) + 0.25 × (1 −

pattern_simplicity) + 0.05 × shannon_norm + 0.05 × (1 − aa_bias) + 0.05 × (1 − rep_intensity)

The resulting score is capped at 1.0. Based on this score, sequences were classified into six categories:

Score range	Classification
< 0.20	Very simple
0.20–0.35	Simple
0.35–0.50	Medium
0.50–0.65	Complex
0.65–0.80	Very complex
≥ 0.80	Extremely complex

Implementation

The method was implemented in Python (v3.10) using standard libraries (numpy, collections, math, csv). For each input FASTA file, the script outputs a CSV table containing all per-sequence metrics and the composite complexity score, along with a summary report of the overall distribution.

Amino acid k-mer profiling

To enable systematic cross-species comparison of short sequence motifs, we generated k-mer frequency spectra from each non-redundant predicted proteome. Unlike DNA-based k-mer analysis, which is limited to four nucleotides, amino acid k-mer analysis leverages the 20-letter amino acid alphabet, providing vastly richer sequence space and greater biological interpretability, as k-mers more directly correspond to potential functional motifs, interaction interfaces, or post-translational modification sites.

For each species, we extracted all contiguous amino acid k-mers of lengths k = 5, 7, 9, and 11

from every protein sequence in the non-redundant predicted proteome file (*.rep.fasta). The k = 7 dataset was prioritized for downstream analyses as it offers an optimal balance between specificity (sufficient length for species discrimination) and sensitivity (frequent occurrence within functional motifs). Shorter k-mers (k = 5) were used for broader phylogenetic comparisons, while longer k-mers (k = 9 and 11) facilitated the identification of nearly unique sequence signatures.

For k-mer counting, we employed Jellyfish (version 2.3.0), a memory-efficient and high-performance k-mer counting tool. Since Jellyfish is designed for nucleotide alphabets (typically DNA or RNA), we adapted it for amino acid sequences by first converting each protein sequence into a continuous string of single-letter amino acid codes using a custom Python preprocessing script. Ambiguous or non-standard amino acids, including the stop codon represented as 'Z', were retained in the sequence to preserve termination signals and avoid artificial fragmentation of ORFs. Preprocessed sequences were written in FASTA format and used as input for Jellyfish.

The Jellyfish counting command was as follows:

bash

jellyfish count -m 7 -s 100M -t 16 -o ${SPECIES}_k7.jf ${SPECIES}_proteins.fasta

Where -m 7 specifies the k-mer length, -s 100M sets the initial hash size (100 million entries, adjusted based on genome size), -t 16 utilizes 16 threads for parallel processing, and -o specifies the output Jellyfish binary database. Following counting, the k-mer frequency histograms were dumped to human-readable text format using:

bash

jellyfish dump -c -t ${SPECIES}_k7.jf > ${SPECIES}_k7_counts.tsv

The -c flag outputs canonical k-mers (though for amino acids, this is not strictly necessary), and -t tab-separates the output with k-mer sequence and count.

To facilitate downstream comparative analysis in Python, we parsed the Jellyfish output files and converted them into serialized Python dictionaries using the pickle module. Each dictionary mapped a unique k-mer (string of length k) to its total occurrence count (integer) across the entire

species-specific proteome. The resulting pickle files (${SPECIES}_k7.pkl) enabled rapid loading and intersection operations across multiple species without repeated disk I/O.

For quality control, we verified that the number of unique k-mers scaled as expected with k-mer length (e.g., longer k-mers producing more unique entries) and that the distribution of k-mer frequencies approximated a Poisson-like distribution, consistent with random sampling of the amino acid sequence space. Species with unusually low unique k-mer counts were flagged for further inspection, potentially indicating high sequence redundancy or assembly artifacts.

This k-mer profiling approach offered several advantages over alignment-based methods: (i) it required no *a priori* reference genome or annotation, (ii) it was computationally efficient for whole-proteome comparisons, (iii) it naturally captured both perfect matches and implicit structural variations, and (iv) it enabled unbiased detection of lineage-specific sequence innovations that may be invisible to homology-based methods.

k-mer extraction and Jaccard similarity calculation

For each species, all possible 7-mer peptides were extracted from six-frame translations of whole-genome sequences using a sliding window approach. For annotated proteins, canonical protein sequences were used directly. k-mer counting was performed using Jellyfish (v2.3.0), and unique k-mer sets were stored as Python pickle files for subsequent analysis.

Pairwise Jaccard similarity was calculated as:

$$J(A,B) = \frac{|\,A \cap B\,|}{|\,A \cup B\,|}$$

where A and B represent the sets of unique 7-mers from two species. Jaccard distance was defined as 1 − J(A,B).

Average Nucleotide Identity (ANI) estimation

To enable comparison with traditional sequence identity metrics, we estimated ANI from k-mer sharing using the formula:

$$ANI = \left(\frac{|\,A \cap B\,|}{(|\,A\,| + |\,B\,|)/2}\right)^{1/k}$$

where |A| and |B| denote the total number of unique k-mers in each species. This approach approximates the average nucleotide identity by accounting for the probability that a randomly selected k-mer from one genome is present in the other, raised to the power of 1/k to normalize for k-mer length.

Data processing and visualization

All pairwise comparisons were computed for seven species/two strains: CHM13 (human T2T reference), YAO (human Asian individual), *Pan troglodytes* (chimpanzee), *Pan paniscus* (bonobo), *Gorilla gorilla* (gorilla), *Pongo abelii* (orangutan), and *Nomascus leucogenys* (gibbon). For annotated proteins, only CHM13 was included as human representative due to data availability.

Heatmaps were generated using Python (v3.13) with matplotlib (v3.10.0). Color scales were standardized across both datasets to enable direct comparison of ANI distributions. The custom colormap transitions from blue (85–90% ANI) through white (95% ANI) to red (98–100% ANI) for intuitive interpretation.

dN/dS (Ka/Ks) ratio calculation

Genome and Protein Data

We analyzed 179 conserved protein-coding genes across seven primate species and mouse: human (T2T-CHM13v2.0), bonobo (*Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo abelii*), siamang (*Symphalangus syndactylus*), crab-eating macaque (*Macaca fascicularis*), and house mouse (*Mus musculus*). Conserved proteins were identified based on k-mer conservation analysis across all seven species, requiring identical k-mer counts in all species and conservation of terminal 4 amino acids.

CDS Sequence Extraction

Coding sequences (CDS) were extracted from telomere-to-telomere (T2T) genome assemblies using custom Python scripts. Genomic coordinates were parsed from protein identifiers in the format chromosome~reading_frame~serial_index~protein_length, where reading frames included forward strands (for0, for1, for2) and reverse strands (rev0, rev1, rev2). CDS

positions were calculated as follows: for forward strand frame 0, start = index × 3 − length × 3 + 3; for reverse strands, positions were calculated from the chromosome end and reverse-complemented. Terminal stop codons (TAA, TAG, or TGA) were removed from all CDS sequences to ensure accurate dN/dS estimation, retaining only codons encoding amino acids.

dN/dS Calculation

Pairwise dN/dS ratios between human (CHM13) and mouse were calculated using the Nei-Gojobori method (Nei & Gojobori, 1986) with Jukes-Cantor correction for multiple substitutions. The method distinguishes synonymous (dS) and non-synonymous (dN) substitution rates by:

1. **Site classification**: Each codon position was classified as synonymous or non-synonymous by enumerating all possible single-nucleotide substitutions and determining whether they change the encoded amino acid using the standard genetic code.

2. **Substitution counting**: Observed synonymous and non-synonymous differences between sequence pairs were counted. Codons containing gaps or ambiguous nucleotides (N) were excluded from analysis.

3. **Jukes-Cantor correction**: Substitution rates were corrected for multiple hits using the formula: $d = -3/4 \times \ln(1 - 4p/3)$, where p is the proportion of observed differences.

4. **Selection pressure interpretation**: dN/dS ratios were interpreted as: < 1 indicating purifying (negative) selection, ≈ 1 indicating neutral evolution, and > 1 indicating positive (Darwinian) selection.

All calculations were implemented in Python using Biopython (v1.87) for sequence handling. Custom scripts are available upon request.

Phylogenetic tree construction and visualization

Sequence Sampling and Preparation

Open reading frame (ORF) sequences were extracted from four primate species: human (YAO), gorilla (*Gorilla gorilla*), bonobo (*Pan paniscus*), and chimpanzee (*Pan troglodytes*). To balance computational efficiency with representative diversity, we employed a stratified sampling strategy. Human sequences were sampled at 55% (n = 247–248) with proportional representation across chromosomes (chr1: ~70%, chr10: ~10%, chr16: ~11%, others: ~9%), while the three non-human primate species were each sampled at 15% (n = 67–68). This resulted in a final dataset of 450 sequences for both nucleotide and amino acid analyses.

Multiple Sequence Alignment and Distance Calculation

For nucleotide sequences, ORFs were truncated to 500 bp (central region retained) to optimize computational performance. For protein sequences, ORFs were truncated to 300 amino acids. Multiple sequence alignment was performed using a gap-padding approach to equalize sequence lengths. Pairwise evolutionary distances were calculated using the identity matrix method, which computes the proportion of identical sites between sequence pairs.

Phylogenetic Tree Construction

Phylogenetic trees were constructed using two distance-based methods implemented in Biopython (v1.87):

1. **Neighbor-Joining (NJ)**: A minimum evolution method that iteratively joins the closest taxa based on corrected distance matrices, producing unrooted trees.
2. **UPGMA (Unweighted Pair Group Method with Arithmetic Mean)**: A hierarchical clustering method assuming a molecular clock, producing rooted ultrametric trees.

Both methods utilized the same distance matrices derived from the identity model.

Tree Visualization

Phylogenetic trees were visualized using custom Python scripts with Matplotlib. Branches were color-coded by species and chromosome: human chr1 (crimson, #DC143C), human other chromosomes (dark orange, #FF8C00), gorilla (dark turquoise, #00CED1), bonobo (light green, #90EE90), and chimpanzee (hot pink, #FF69B4). All visualizations used Arial

font and were exported as both PDF (vector format for editing) and PNG (raster format for preview). Additionally, trees generated by iTOL were post-processed using custom SVG manipulation scripts to apply the same color-coding scheme based on sequence identifiers.

Mining peptide sequences from MS datasets using PepQuery

This study integrated tandem mass spectrometry (MS/MS) datasets derived from various biological samples, including cancer tissues, human embryonic stem cells (hESCs), heat-stressed 293T cells, and oocytes (Supplementary Table 9). To meet the input requirements of the downstream targeted search software, the raw mass spectrometry data (.raw format) generated by the instruments were first subjected to peak picking and converted into the standard mass spectra file format (.mgf) using ThermoRawFileParser. Subsequently, spectra from the same sample types were merged for downstream analysis.

To validate target peptides across the MS datasets, hypothesis-driven identification was performed using PepQuery (v2.0.2). A custom peptide library was used as the target input, while the UniProt human reference proteome (January 2026 release) served as a background database to exclude interference from known proteins. Searches were executed in protein matching mode with strict high-confidence filtering to rigorously control the false positive rate. All analyses were performed on a high-performance computing (HPC) cluster using high-concurrency resources to ensure efficient large-scale validation.

Ribo-seq data preprocessing

Raw sequencing reads (Supplementary Table 10) were downloaded and processed through a four-step pipeline. First, adapter sequences (AGATCGGAAGAGCACACGTCT) were trimmed using cutadapt (v5.2) with a minimum length cutoff of 15 nt. Read quality was assessed using FastQC (v0.11.9). Reads derived from ribosomal RNA (rRNA) were removed by aligning to a custom human rRNA index (18S, 28S, 5.8S, and 5S rRNA sequences; NCBI accessions NR_046235.1, NR_046236.1, NR_046237.1, NR_046238.1) using bowtie2 (v2.5.0) in --very-sensitive mode, retaining only unaligned reads (--un-gz). Mitochondrial RNA (mtRNA) contamination was subsequently removed by aligning the rRNA-depleted reads to the human mitochondrial genome (NC_012920.1) using the same bowtie2 parameters. The final clean reads were subjected to a

second round of FastQC quality assessment. All preprocessing steps were parallelized across 8 threads and processed at the individual sample level using a pipelined workflow to maximize throughput.

Ribo-seq alignment and quantification

Two sets of predicted ORF sequences were used as reference. Set 1 ("YAO_top1000") contained 1,000 human istORFs derived from ribosome profiling data. Set 2 ("Conserved_ORF") contained 5,437 ORFs conserved across six primate species (*Homo sapiens*, *Pan paniscus*, *Gorilla gorilla*, *Pongo abelii*, *Symphalangus syndactylus*, *Macaca fascicularis*), identified by k-mer conservation analysis. For each set, duplicate sequences were removed based on nucleotide identity.

A seed-and-extend algorithm was implemented for read mapping. Briefly, a k-mer position index (k = 15) was constructed from all ORF sequences, storing each k-mer as key → [(gene_id, position), ...]. For each read, all constituent 15-mers were queried against the index; candidate alignment positions were generated from k-mer matches (read position i matches gene position j ⇒ alignment start = j − i). Full-length Hamming distance was then computed at each candidate position, and matches with ≤1 mismatch were retained. Both forward and reverse-complement orientations were considered.

Read counts per ORF were normalized to RPK (Reads Per Kilobase) and average coverage depth (assuming mean ribosome footprint length of 30 nt).

RT-PCR for MESSS

Forward primer: CAGCATCTAGATATGACGTAGATGATTCCATTCGGGTC

Reverse primer: CAGCATCTAGATATGACG

siRNA knockdown for MESSS

siMESSS1-1:

- sense: CGAAUGAAUUGAAUGCAAUCA
- antisense: AUUGCAUUCAAUUCAUUCGAU

siMESSS1-2:

- sense: CGAAUGGUCUCGAAUGGAAUC
- antisense: UUCCAUUCGAGACCAUUCGAU

siMESSS1-3:

- sense: GAAUGGAAUCAUCUUCAAAUG
- antisense: UUUGAAGAUGAUUCCAUUCGA

Data availability

All supplementary files can be found at: https://osf.io/v2hqy/files/. If files cannot be downloaded by clicking the links in the main text, please browse them manually under this URL.

Acknowledgments

We thank Xu Zhan (Henan Vocational College of Tuina) for her assistance in the annotation for the animal and plant species where we first saw those satellite-derived proteins. We thank Yafeng Zhu (Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University) for technical support in proteomics analysis. The 6-frame in silico translation approach revealed so many unexpected phenomena that we wished to share them with readers as early as possible, so that interested colleagues may begin similar explorations sooner. Due to the breadth of content and time constraints, the current version of this manuscript remains preliminary, and may contain errors or omissions. We welcome comments and criticism from readers.

References

1. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53. https://doi.org/10.1126/science.abj6987.

2. He Y, Chu Y, Guo S, Hu J, Li R, Zheng Y, et al. T2T-YAO: A Telomere-to-telomere Assembled Diploid Reference Genome for Han Chinese. Genomics Proteomics Bioinformatics. 2023;21:1085–100. https://doi.org/10.1016/j.gpb.2023.08.001.

3. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437:69–87. https://doi.org/10.1038/nature04072.

4. Yoo D, Rhie A, Hebbar P, Antonacci F, Logsdon GA, Solar SJ, et al. Complete sequencing of ape genomes. Nature. 2025;641:401–18. https://doi.org/10.1038/s41586-025-08816-3.

5. Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, et al. Pervasive functional translation of noncanonical human open reading frames. Science. 2020;367:1140–6. https://doi.org/10.1126/science.aay0262.

6. Kesner JS, Chen Z, Shi P, Aparicio AO, Murphy MR, Guo Y, et al. Noncoding translation mitigation. Nature. 2023;617:395–402. https://doi.org/10.1038/s41586-023-05946-4.

7. Yang M, Xie Y, Wang L, Jungreis I, Ou T, Kellis M, et al. Proteogenomics-enabled discovery of novel small open reading frame (sORF)-encoded polypeptides in human and mouse tissues. Nucleic Acids Res. 2025;53:gkaf687. https://doi.org/10.1093/nar/gkaf687.

8. Deutsch EW, Kok LW, Mudge JM, Valls CF, Jungreis I, Ruiz-Orera J, et al. Expanding the human proteome with microproteins and peptideins. Nature. 2026;:1–13. https://doi.org/10.1038/s41586-026-10459-x.

9. Fonseca-Carvalho M, Veríssimo G, Lopes M, Ferreira D, Louzada S, Chaves R. Answering the Cell Stress Call: Satellite Non-Coding Transcription as a Response Mechanism. Biomolecules. 2024;14:124. https://doi.org/10.3390/biom14010124.

10. Kwan T, Thompson SR. Noncanonical Translation Initiation in Eukaryotes. Cold Spring Harb Perspect Biol. 2019;11:a032672. https://doi.org/10.1101/cshperspect.a032672.

11. Chao K-H, Zimin AV, Pertea M, Salzberg SL. The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual. G3 Genes|Genomes|Genetics. 2023;13:jkac321. https://doi.org/10.1093/g3journal/jkac321.

12. Yang C, Zhou Y, Song Y, Wu D, Zeng Y, Nie L, et al. The complete and fully-phased diploid genome of a male Han Chinese. Cell Res. 2023;33:745–61. https://doi.org/10.1038/s41422-023-00849-5.

13. Hansen NF, Dwarshuis N, Ji HJ, Rhie A, Loucks H, Logsdon GA, et al. A complete diploid human genome benchmark for personalized genomics. 2025;:2025.09.21.677443. https://doi.org/10.1101/2025.09.21.677443.

14. Haubold B, Wiehe T. How repetitive are genomes? BMC Bioinformatics. 2006;7:541. https://doi.org/10.1186/1471-2105-7-541.

15. Jarmuż M, Glotzbach CD, Bailey KA, Bandyopadhyay R, Shaffer LG. The Evolution of Satellite III DNA Subfamilies among Primates. The American Journal of Human Genetics. 2007;80:495–501. https://doi.org/10.1086/512132.

16. Trivedi M, Gianfrate F, Gennaro L de, Ayllon M, Munson KM, Hoekzema K, et al. Rapid centromere turnover and the adaptive radiation of lemurs. 2026;:2026.05.16.725662. https://doi.org/10.64898/2026.05.16.725662.

17. Hsieh P, Soisangwan N, Gordon DS, Javidh A, Harvey WT, Porubsky D, et al. A global map for introgressed structural variation and selection in humans. Science. 2026;392:eadz7518. https://doi.org/10.1126/science.adz7518.

18. Popesco MC, Maclaren EJ, Hopkins J, Dumas L, Cox M, Meltesen L, et al. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. Science. 2006;313:1304–7. https://doi.org/10.1126/science.1127980.

19. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet. 2014;46:1063–71. https://doi.org/10.1038/ng.3092.

20. Yue L, Jiang W, Li S, Luo M, Fan N, Zhan X, et al. Spatial distribution of the proteome in the human body and in cancers. Nature. 2026;:1–10. https://doi.org/10.1038/s41586-026-10660-y.

21. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, et al. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009;41:1223–7. https://doi.org/10.1038/ng.474.

22. Maillard AM, Ruef A, Pizzagalli F, Migliavacca E, Hippolyte L, Adaszewski S, et al. The 16p11.2 locus modulates brain structures common to autism, schizophrenia and obesity. Mol Psychiatry. 2015;20:140–7. https://doi.org/10.1038/mp.2014.145.

23. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021;372:eabf7117. https://doi.org/10.1126/science.abf7117.

Figure Legends

Fig. 1. Distribution of istORFs across different genomic regions. (A) Proportions of istORFs in different genomic annotation categories. (B) Circos plot showing the distribution of istORFs in comparison with annotated protein-coding genes, LINEs, SINEs, and satellite DNA. (C) Density plot of istORF distribution along chromosomes; red bars indicate satellite DNA density.

Fig. 2. Complexity distribution of isteins. Complexity scores were calculated using the Wootton & Federhen method. Density values are relative densities with a fixed AUC of 500 for each genome.

Fig. 3. Pairwise similarity estimates based on 7-mer analysis. (A) Annotated proteins show high similarity between human and chimpanzee (98.3%) and progressively decreasing similarity with increasing evolutionary distance. (B) 6-frame translation-predicted proteins reveal substantially lower inter-species similarity (human-chimpanzee: 94.5%; human-gibbon: 88.3%), while maintaining high intra-species identity (human individuals: 99.3%). Color scale ranges from blue (85) to red (100). Values represent ANI percentages.

Fig. 4. Phylogeny and predicted structures of MESSS. (A) Phylogenetic tree of MESSS proteins. Colors indicate species and chromosome: human chr1 (crimson, 173 sequences), human other chromosomes (dark orange, 74 sequences), bonobo (light green, 68 sequences), chimpanzee (hot pink, 67 sequences), and gorilla (cyan, 68 sequences). (B–C) AlphaFold-predicted structures of MESSS16_186 (B) and MESSS1_446 (C).

Fig. 5. Genome repetitiveness and Sat2/3-derived k-mer distributions. (A) Repetitiveness of individual chromosomes across different genomes. (B–C) Distribution of characteristic 3-mer (B) and 5-mer (C) peptides encoded by Sat2/3. Orange tones represent heavy-strand k-mers; green tones represent light-strand k-mers.

Fig. 6. Highly conserved isteins. (A) Numbers of conserved isteins identified at each taxonomic level. (B) Distribution of conserved isteins across intergenic regions, introns, and exons. (C) Proportions of protein-coding genes, lncRNAs, and pseudogenes among the host genes of conserved isteins. (D–F) dN/dS analysis of the 179 istORFs conserved across primates and mouse: scatter plot of dN vs. dS (D); distribution of dN/dS ratios (E); proportions of genes under positive selection, purifying selection, and neutral evolution (F).

Fig. 7. Proteomics validation of istein expression. (A) Coverage heatmap of the top 1,000 isteins from YAO across 7 datasets. (B) Coverage heatmap of 136 MESSS isteins from the top 1,000. (C) Coverage of the 179 isteins conserved across six primates and mouse across 10 datasets. (D) Coverage of 5,437 isteins conserved in Old World monkeys across 9 datasets. Color scales in (C) and (D) are identical.

Fig. 8. Ribo-seq evidence for translation of top 1,000 and conserved istORFs. (A) Translation signals of the top 1,000 istORFs in four experiments. (B) Translation signals of istORFs conserved in Old World monkeys in four experiments. (C) Overlap between mass spectrometry and ribo-seq evidence for the top 1,000 istORFs. (D) Overlap between mass spectrometry and ribo-seq evidence for MESSS within the top 1,000. (E) Overlap between mass spectrometry and ribo-seq evidence for conserved istORFs in Old World monkeys. (F–G) RT-PCR detection of MESSS RNA in single- and polysome fractions from MDA-MB-231 (F) and HeLa (G) cells using primers specific to MESSS1_446.

Fig. 9. Western blot and immunofluorescence detection of MESSS expression. (A) Western blot detection of MESSS in multiple human and animal cell lines. Lanes: 1, 293T; 2, HeLa; 3, Vero (African green monkey kidney); 4, NT2 (teratocarcinoma); 5, PK15 (pig kidney); 6, CHO (Chinese hamster ovary); 7, MDCK (dog kidney); 8, L929 (mouse fibroblast); 9, HepG2 (hepatoblastoma); 10, Huh7; 11, LM3; 12, PLC8024 (hepatocellular

carcinoma); 13, NCM460 (immortalized colonic epithelial); 14, SW480; 15, LoVo; 16, HCT116 (colon cancer); 17, CAL120; 18, MDA-MB-468; 19, SUM159PT; 20, BT549; 21, MDA-MB-231 (triple negative breast cancer); 22, 293T; 23, HeLa. (C–D) Western blot detection of MESSS knockdown by siRNA in MDA-MB-231 cells, two independent experiments. (E) Immunofluorescence detection of MESSS knockdown by siRNA in MDA-MB-231 cells: siRNA vs. control siRNA; DAPI/MESSS/merge.

Fig. 1

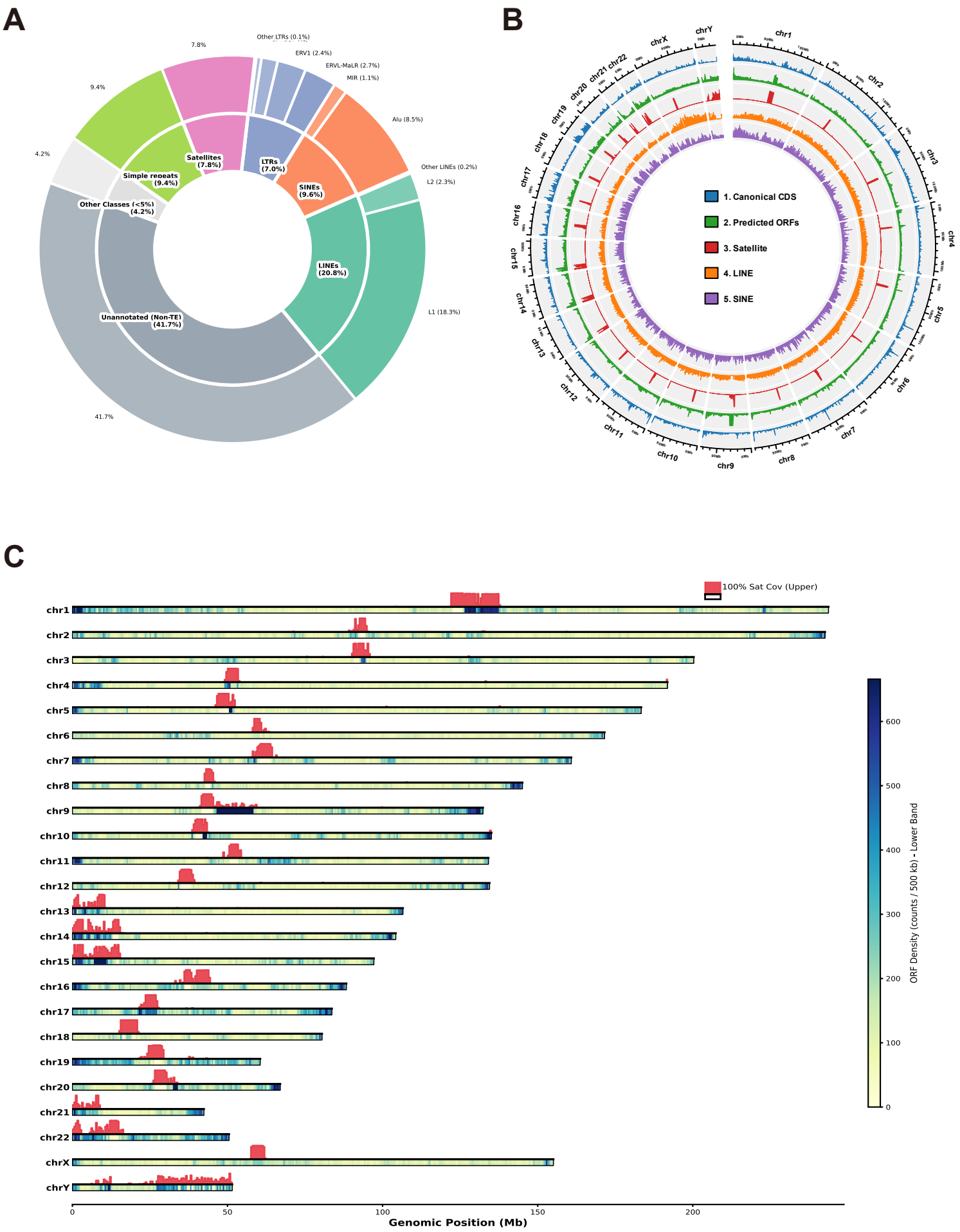


Fig. 2

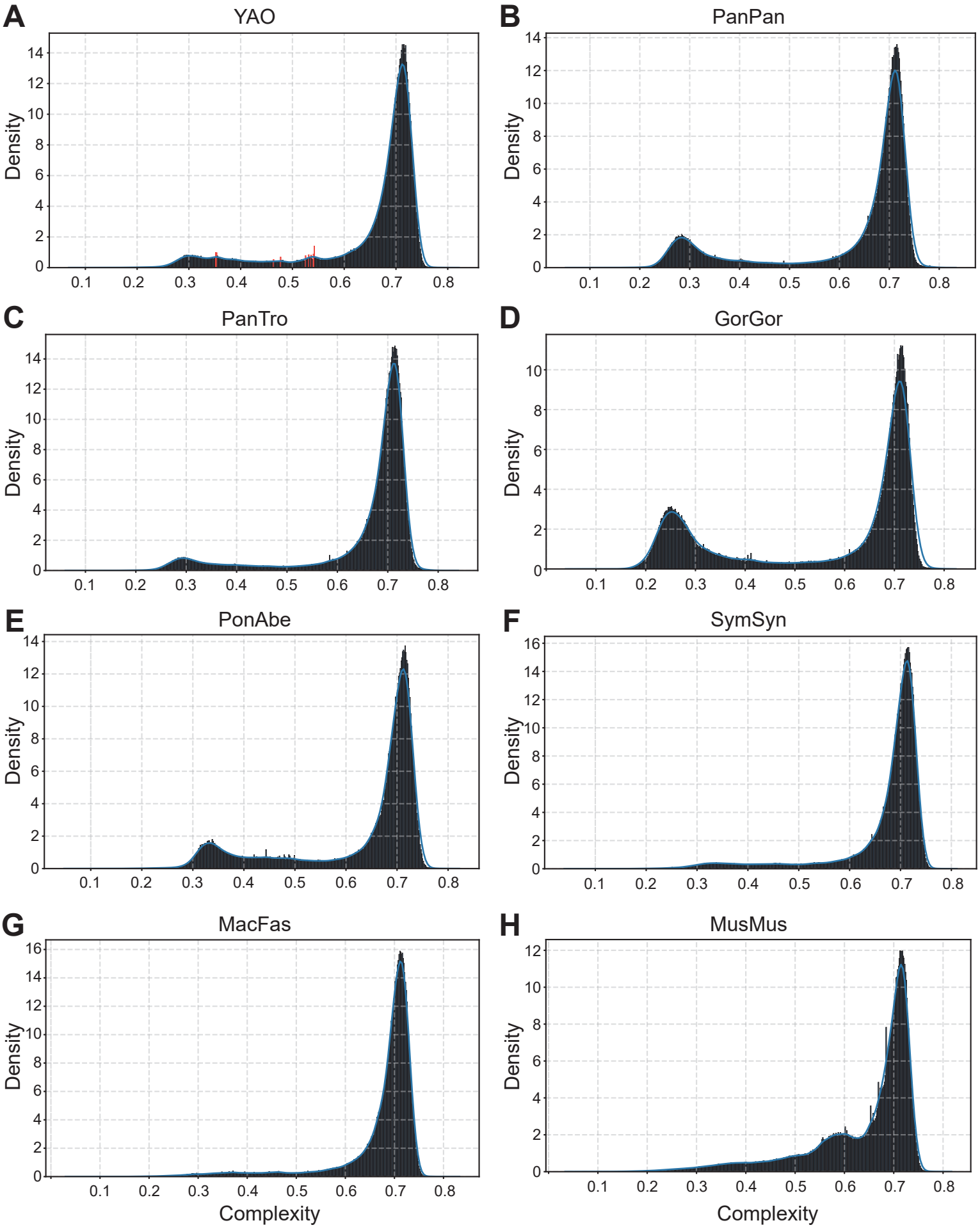
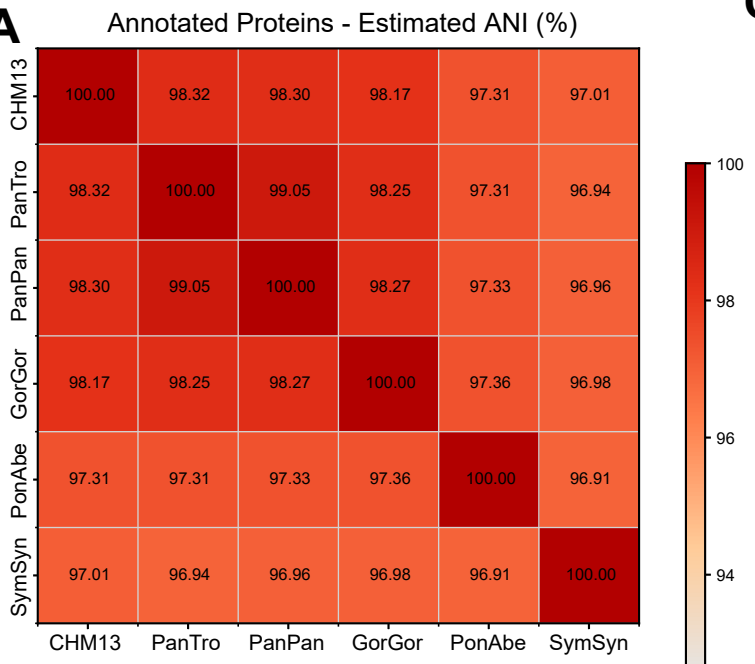
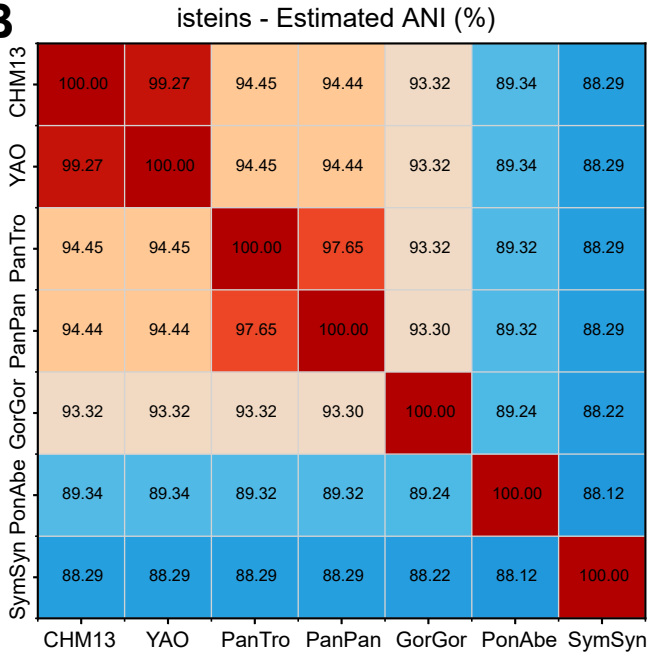


Fig. 3

A



B

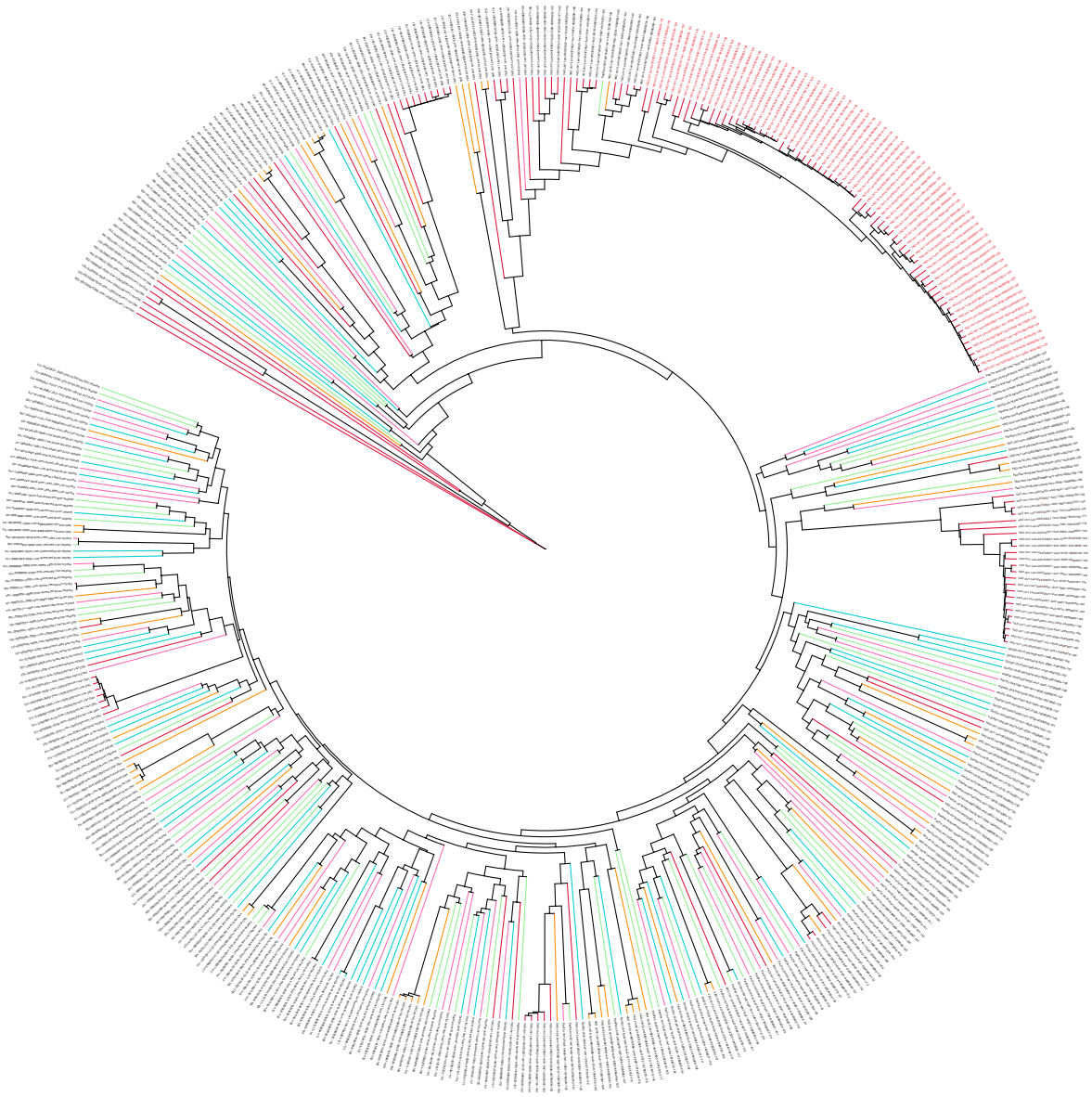


C

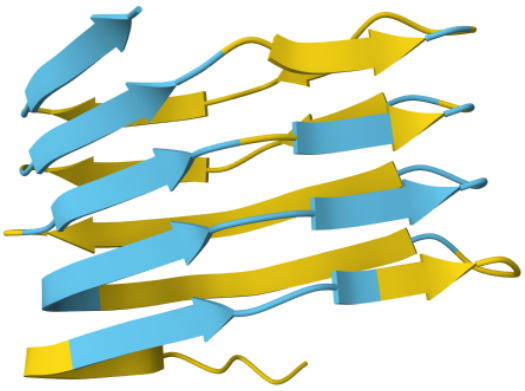
7-mer	YAO	PanTro	GorGor	PonAbe
SSNGMEW	14087	47	20	5
MDSNGII	13050	161	141	0
MESSSNG	11068	326	300	0
HRMDSNG	11050	86	81	0
SNIDYIL	11039	2	1	0
GIEWNHR	10806	77	72	0
QGIYSIL	10555	2	1	0
KKYQSTQ	10535	18	16	0
NHRMDSN	10430	62	42	0
EWTRMES	9540	211	157	0
NGMELTR	8925	619	501	0
RMESSSN	8763	222	190	0
GMEWNNP	8738	178	8	0
IIEWTRM	8407	116	87	0
NGINPSA	8328	40	31	0
WTRMESS	8092	157	109	4
NGIIIER	7811	11	9	0
EWNGINS	7803	3645	1141	0
GIIIERN	7785	8	4	0
WNHRMDS	7781	57	38	0
PWTRMQS	7760	2	1	0
NPWTRMQ	7731	2	1	0
MESTSNG	7592	14	11	0
EWNNPWT	7431	2	1	0
IIEWSRM	7275	13	3	0
MESSSNE	7223	20	12	0
INPNRME	7126	678	511	0
MEWNQLD	7124	33	28	0
NNPWTRM	7096	2	1	0
WNNPWTR	7090	2	1	0
SRMESSS	7031	10	5	1
MEWNNPW	6995	2	1	0
GINPNRM	6904	691	472	0
NPNRMEW	6670	581	351	0
TRIEWNG	6128	1384	867	31
MELTRIE	6122	43	17	0
LTRIEWN	6106	51	7	0
ELTRIEW	5993	49	20	1
IIIKWNR	5772	62	61	0
MESKRII	5522	7	1	0
GINPSGM	5366	891	840	0
RMESTSN	5366	7	5	0
SKRIIEW	5120	6	2	0
EWNQPEC	5022	2222	68	0
GMEWNHR	4818	19	15	0
NAMEWNH	4733	98	10	0
AMEWNHP	4717	80	6	0
IIKWNRM	4710	80	76	0

Fig. 4

A



B



C

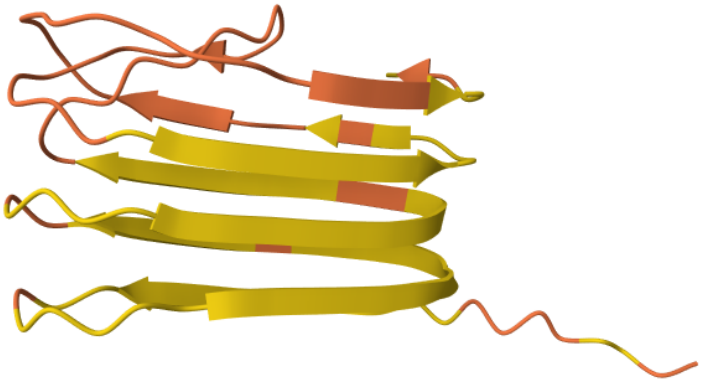


Fig. 5

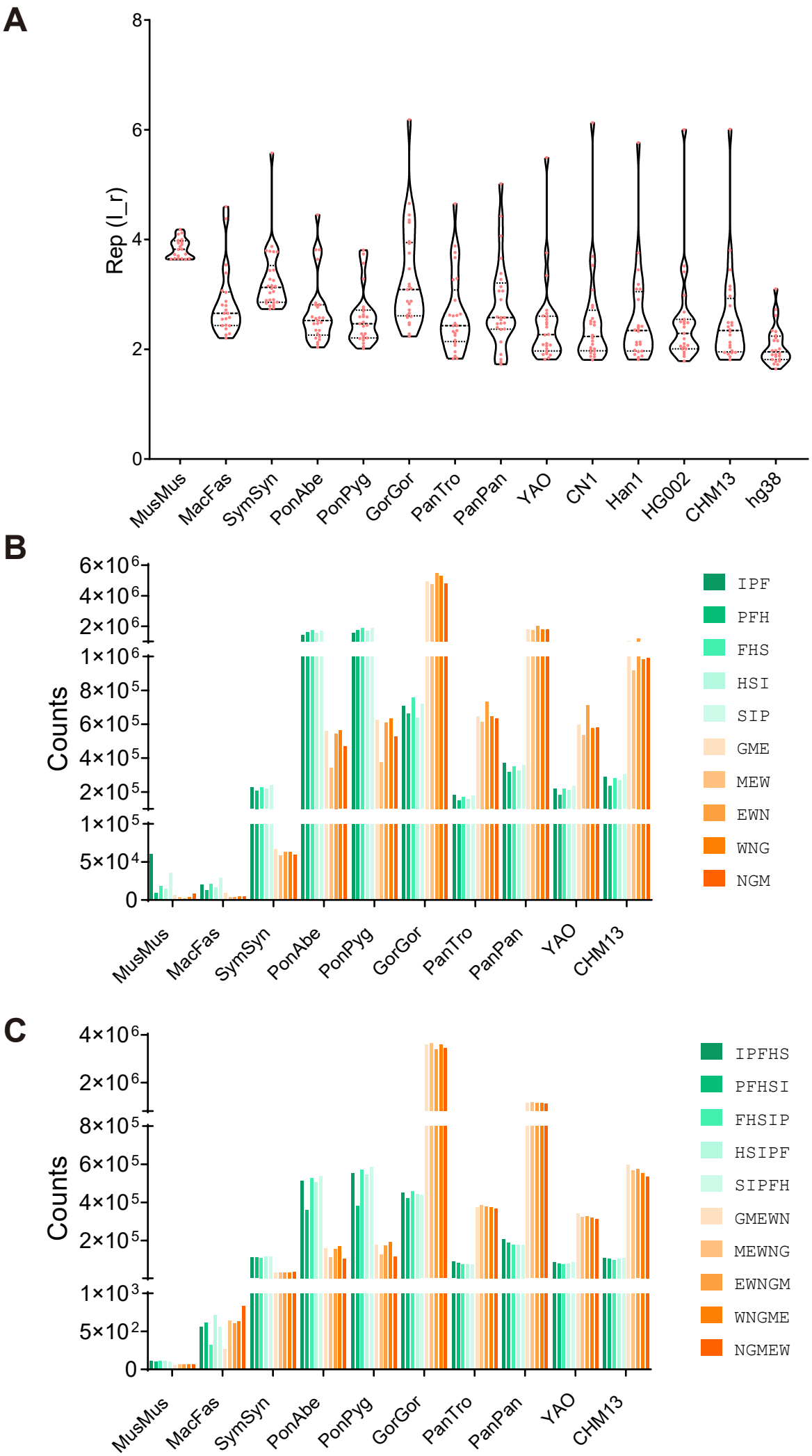


Fig. 6

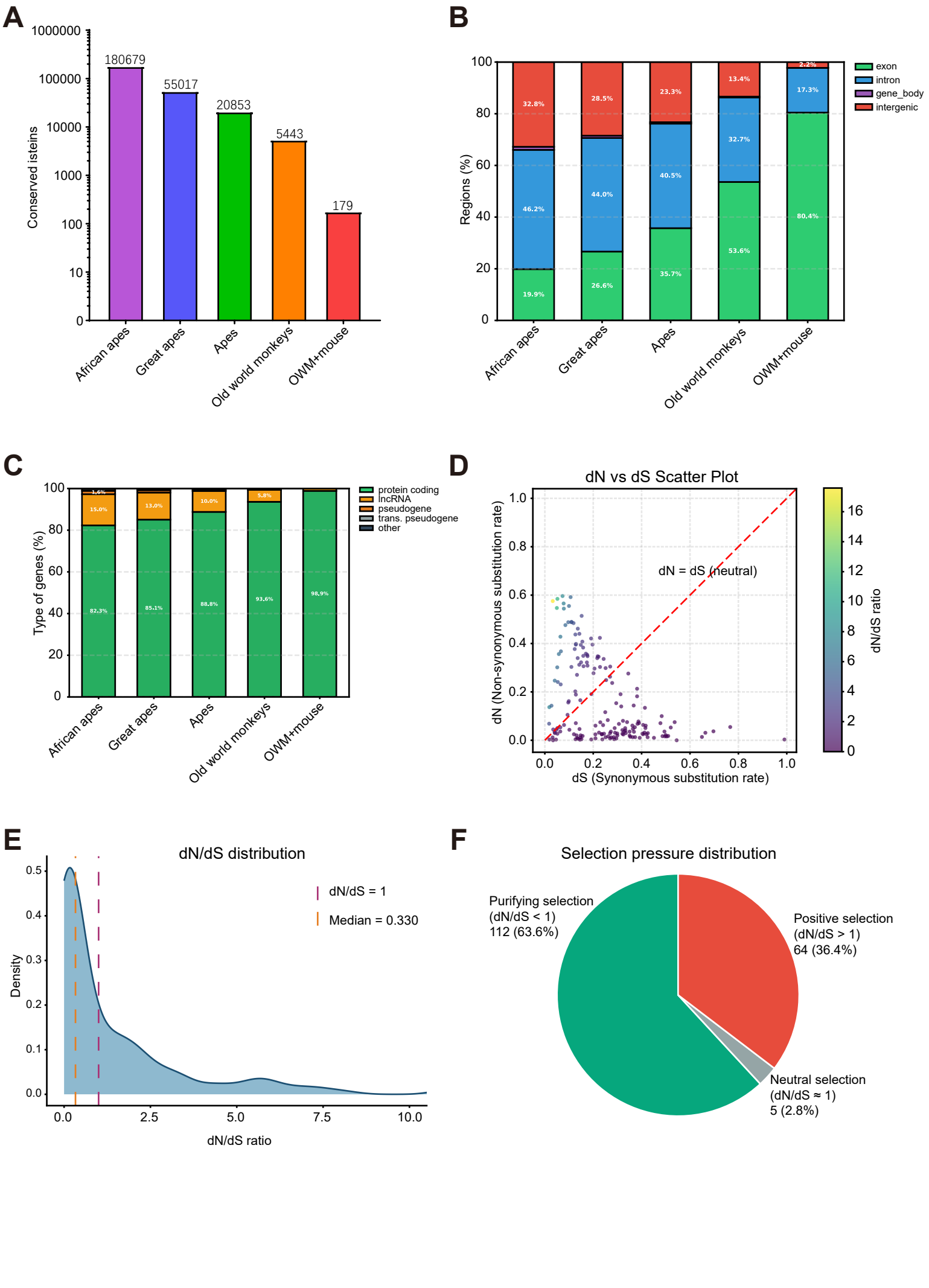


Fig. 7

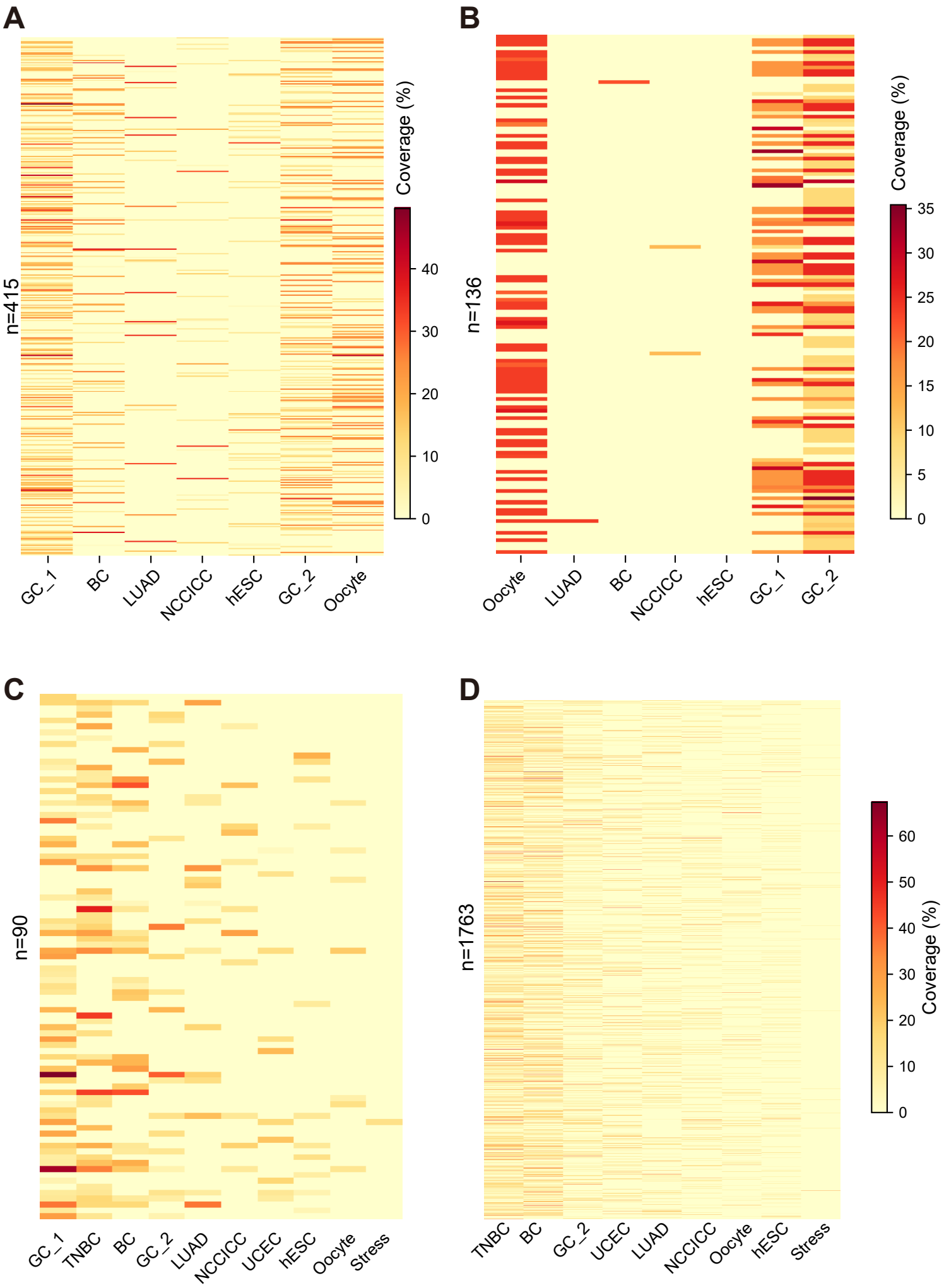


Fig. 6

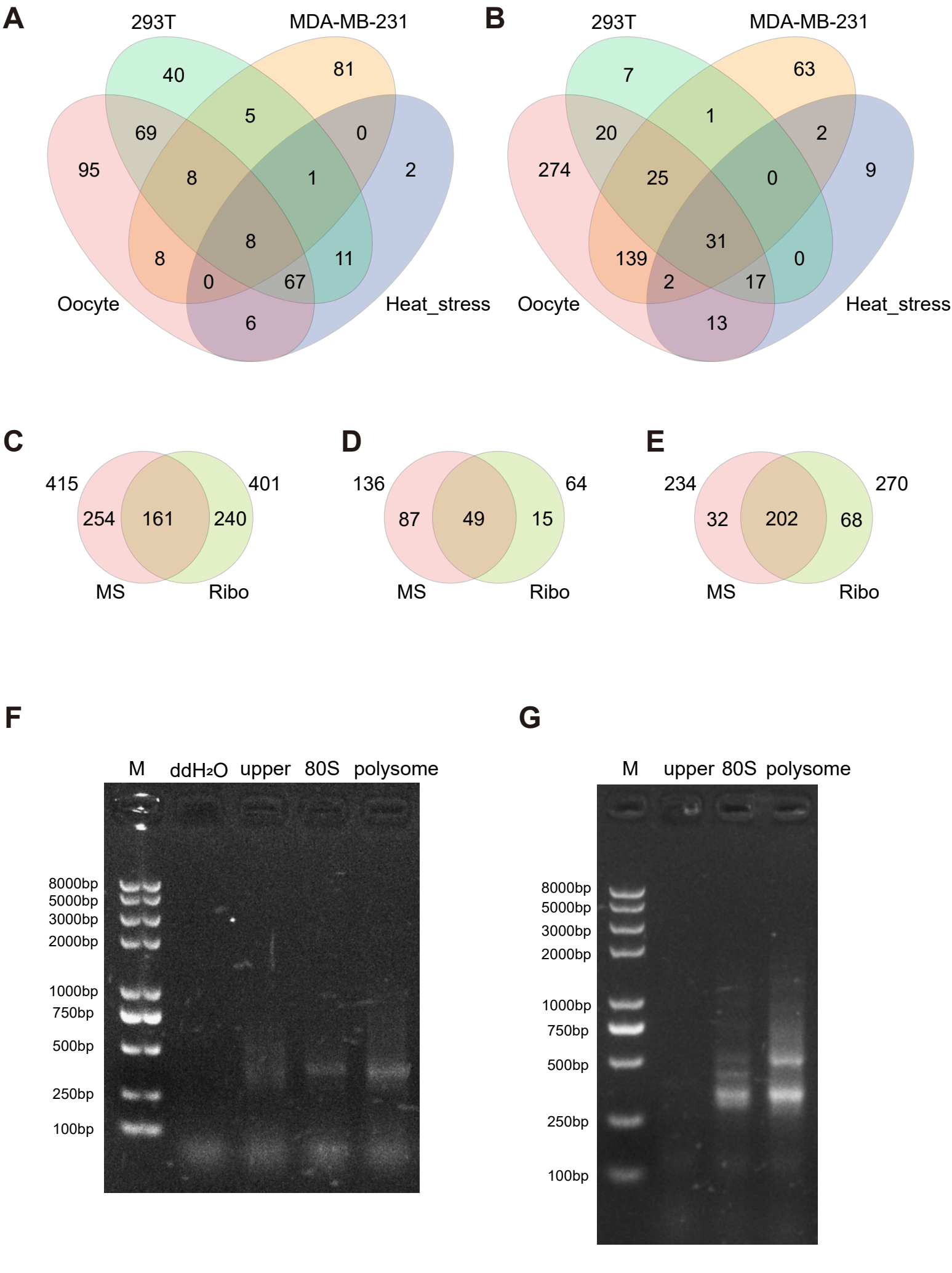


Fig. 9