

Structural Genomics & Information Complexity | Signal Rating: 1/5 Stars | Oct 28, 2026 | 8 min read

Human GRCh38 Chromosome 1p36.33 Unbiased Sliding-Window Scan Reveals a Z = +4.24 Non-B DNA & Z-DNA Repeat Anomaly Inside the AGRN Neuromuscular Locus

Annotation-agnostic sliding-window profiling across 60 kb of subtelomeric Human Chromosome 1 Section A (NC_000001.11:1,000,001-1,060,000) isolates extreme Z-DNA alternating purine-pyrimidine and G-quadruplex structural anomalies embedded within the 5' and 3' coding architecture of Agrin (AGRN).

1. Abstract & Emergent Discovery Overview

Standard functional genomics pipelines partition reference chromosomes strictly along annotated start and stop codon boundaries, discarding intronic and perigenic repeat architectures as background noise. Using the CRV Science Labs multi-threaded Rust engine (bio-suite v0.1.0), we executed two coupled worker-agent workflows across a 60,000 bp subtelomeric segment of Human Chromosome 1p36.33 (GRCh38.p14 primary assembly NC_000001.11:1,000,001-1,060,000 , designated Section A) [1] : (i) targeted protein-coding locus extraction across HES4 , ISG15 , and AGRN , and (ii) an annotation-agnostic sliding-window structural complexity scan (2,000 bp windows, 1,000 bp step size, 59 windows total) combining 3-mer Shannon block entropy (H_3), Jensen-Shannon compositional divergence (D_{JS}), Lempel-Ziv 1976 (LZ76) algorithmic compressibility [2] , Purine- Pyrimidine DNA walk Detrended Fluctuation Analysis (alpha DFA) [3] , and Non-B DNA secondary structure scoring (G4Hunter G-quadruplex [4] and ZHunt -class alternating purine-pyrimidine Z-DNA propensity [5]).

Cross-intersecting the unbiased complexity scan against RefSeq gene models isolates a local sequence repeat window: Region-X01 (relative coordinates 29,000-31,000 bp, NC_000001.11:1,029,001-1,031,000), located inside the 5' interval of the neuromuscular synapse organizer gene Agrin (AGRN), registers a structural complexity Z-score of $Z = +4.2392$. While normalized Lempel-Ziv algorithmic complexity drops to 0.3893 due to a tandem dinucleotide/tetranucleotide repeat tract, its Z-DNA alternating purine-pyrimidine score reaches 19.95 (8.2x the locus mean of 2.41) alongside a G4Hunter G-quadruplex score of 2.12 .

2. Chr1p36.33 Locus Dataset & Coding Complexity Scores

The 60,000 bp subtelomeric 1p36.33 segment exhibits a GC-rich isochore signature (GC = 63.16%, GC3 = 63.90%) characteristic of human telomere-proximal gene clusters [1] . Click Copy for Excel (TSV) to copy the locus dataset and complexity scores directly into Excel:

Locus / Feature ID	Strand	Start (bp)	End (bp)	Length (bp)	GC (%)	GC3 (%)	CpG (O/E)	Shannon H3 (bits)	Norm LZ76	DFA Alpha	Hurst H	G4Hunter Max	Z-DNA Score
Full Chr1 Section A (NC_000001.11)	+	1	60000	60000	63.16	63.90	0.4672	5.7054	0.8532	0.6026	0.6234	3.12	2.4133

| HES4 (bHLH Notch target) | | - | | 1100 | | 2350 | | 1250 | | 69.44 | | 65.87 | |
0.6529 | | 5.5066 | | 0.9053 | | 0.6324 | | 0.6344 | | 2.36 | | 0.7200 |

| ISG15 (Ubiquitin-like modifier) | | + | | 13400 | | 14550 | | 1150 | | 64.43 | |
55.35 | | 0.4817 | | 5.6160 | | 0.9106 | | 0.5059 | | 0.5150 | | 1.64 | | 0.7826 |

| AGRN 5 Sector (Exons 1-8) | | + | | 20100 | | 36500 | | 16400 | | 65.71 | | 65.26
| | 0.3868 | | 5.5635 | | 0.8059 | | 0.5939 | | 0.5961 | | 3.12 | | 3.7195 |

| AGRN 3 Sector (Exons 9-36) | | + | | 36501 | | 56100 | | 19599 | | 65.75 | | 66.26
| | 0.4820 | | 5.6289 | | 0.8861 | | 0.6053 | | 0.6490 | | 3.04 | | 2.5665 |

3. Sliding-Window Complexity Anomalies (Regions X01-X05)

When the worker-agent ignores start/stop codon boundaries and scans the raw nucleotide stream in 2,000 bp overlapping windows, six windows exceed the $Z \geq +1.15$ structural anomaly threshold within the AGRN gene body:

| Anomaly ID | | Window ID | | Start (bp) | | End (bp) | | Complexity Z-Score | | GC
(%) | | CpG (O/E) | | G4Hunter Score | | Z-DNA Score | | Norm LZ76 | | DFA Alpha | |
Intersected RefSeq Annotation |

| Region-X01 | | W030 | | 29000 | | 31000 | | 4.2392 | | 56.70 | | 0.2425 | | 2.12 |
| 19.95 | | 0.3893 | | 0.3874 | | AGRN (5 exons 1-8 interval) |

| Region-X02 | | W031 | | 30000 | | 32000 | | 3.6828 | | 56.75 | | 0.1877 | | 1.80 |
| 17.80 | | 0.4606 | | 0.4067 | | AGRN (5 exons 1-8 interval) |

| Region-X03 | | W053 | | 52000 | | 54000 | | 2.8810 | | 58.65 | | 0.4272 | | 2.00 |
| 14.60 | | 0.7868 | | 0.5899 | | AGRN (3 exons 9-36 interval) |

| Region-X04 | | W052 | | 51000 | | 53000 | | 2.2648 | | 61.00 | | 0.5274 | | 2.16 |
| 11.95 | | 0.8115 | | 0.6109 | | AGRN (3 exons 9-36 interval) |

| Region-X05 | | W056 | | 55000 | | 57000 | | 1.5306 | | 62.10 | | 0.4607 | | 2.92 |
| 8.65 | | 0.8279 | | 0.6240 | | AGRN (3 laminin-G/EGF domain boundary) |

4. Biophysical Synthesis & 1* Signal Limitations

This baseline pipeline run illustrates why algorithmic complexity (LZ76) and biophysical motif propensity (G4Hunter / Z-DNA) diverge in simple repeat windows. In Kolmogorov information theory, a tandemly repeated purine-pyrimidine tract -- such as (TG)_n / (CA)_n or (CG)_n -- exhibits low algorithmic entropy (normalized LZ76 = 0.3893 in Region-X01) because the Lempel-Ziv parser compresses repeated words into a small dictionary [2]. Biophysically, alternating syn/anti glycosidic torsion patterns can adopt left-handed Z-DNA (Z-DNA score = 19.95) and guanine-rich tracts can score positively on G4Hunter (2.12-2.92) under negative supercoiling stress behind RNA Polymerase II [4, 5], potentially intersecting regions involved in alternative splicing and spliceosome kinetics.

Signal Rating Assessment (1* Baseline / Review Note): Although Region-X01 is a useful sanity check demonstrating that the bio-suite sliding-window scanner detects low-complexity tandem repeats inside annotated gene bodies, this run is rated 1* because the dataset is restricted to a single 60 kb reference slice (n = 59 windows) without whole-chromosome background normalization, population structural-variant

controls, or experimental ChIP-seq / kethoxal / permanganate footprinting linking the sequence motif to a verified biological factor in vivo.

References & Primary Sources

- * [1] NCBI RefSeq Homo sapiens Chromosome 1 GRCh38.p14 Primary Assembly (NC_000001.11:1,000,001-1,060,000, 1p36.33 subtelomeric locus).
https://www.ncbi.nlm.nih.gov/nucore/NC_000001.11
- * [2] Lempel, A., & Ziv, J. (1976). On the Complexity of Finite Sequences. IEEE Transactions on Information Theory, 22(1), 75-81.
<https://doi.org/10.1109/TIT.1976.1055501>
- * [3] Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., & Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. Physical Review E, 49(2), 1685-1689. <https://doi.org/10.1103/PhysRevE.49.1685>
- * [4] Bedrat, A., Lacroix, L., & Mergny, J.-L. (2016). Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic Acids Research, 44(4), 1746-1759.
<https://doi.org/10.1093/nar/gkw006>
- * [5] Ho, P. S., Ellison, M. J., Quigley, G. J., & Rich, A. (1986). A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. The EMBO Journal, 5(10), 2737-2744.
<https://doi.org/10.1002/j.1460-2075.1986.tb04558.x>