Supplementary Information

A common human anticodon polymorphism in an alanine-like tRNA gene predicts alanine incorporation at valine codons

Davis W. Cheng

Supplementary tables and data files

Table S1. Acceptor-stem third base pair for all 431 human tRNA genes in the analysis set (429 GtRNAdb high-confidence genes plus the 2 isotype-mismatch genes). Columns give locus, gene name, anticodon isotype, anticodon, tRNAscan-SE body isotype model, the third acceptor-stem pair as read from the folded structure, a binary G3:U70 call, and the GtRNAdb note field. Supports Figure 3a. 431 rows.

Table S2. All single-nucleotide variants at anticodon positions 34–36 of the selected genes, from gnomAD v4. Queries returned results for 429 of the 431 selected loci; the two failures, both on an unplaced contig, were excluded from analysis rather than recorded as variant-negative, and all gnomAD-based denominators in the manuscript are 429. Columns give locus, gene, isotype, anticodon, body model, variant identifier, position, reference and alternate alleles, allele frequency, allele count, allele number and the GtRNAdb note. 858 rows; 2 of 431 loci on an unplaced contig failed to return and are absent.

Table S3. The same variants translated into the amino acid each new anticodon would decode, accounting for strand and wobble. Adds the original and new anticodon, the original and new amino acid, and flags for identity change and for stop codons. Sorted by allele frequency; rs1233572 is the first row. Supports Figure 3b. 856 rows.

Data S4. Primate synteny analysis. Orthologous coordinates in chimpanzee, gorilla and macaque, the extracted and reverse-complemented gene sequence for each, the number of differences from the human gene, and the anticodon and acceptor-stem pair read from each. Supports Figure 1a.

Data S5. Expression scan. Per-sample read counts for the eight human mim-tRNAseq samples of GSE152621, with locus assignments, observed anticodons, decoy counts and the two informative reads in full. Also records the two series that were attempted and excluded, with the failed positive control that justified excluding them.

Data S6. Population differentiation and association analysis. Per-population allele frequencies, the Fst value for rs1233572, the two matched background distributions with medians and percentiles, and the GWAS Catalog and linkage results for the locus. Supports the Results section on population differentiation and linkage disequilibrium.

Data S7. Power calculation for a genotype-stratified proteomic test, including the assumptions behind the predicted misincorporation rate and the valine observations required at three effect sizes. Supports the Discussion.

Table S8. Anticodon-shift loci of Rogers and Griffiths-Jones (2014) tested for standing variation in humans. One row per human gene in a shift set on a primary chromosome, giving their set identifier, gene identifier, isotype and anticodon, their GRCh37 coordinates exactly as published in their Supplementary File 3 and unmodified by any step here, the lifted GRCh38 coordinates, the GtRNAdb gene matched, and the most frequent anticodon variant that gene carries in gnomAD v4 with its allele frequency and the amino-acid change it implies. Rows are ordered by that frequency. 59 rows, of which 51 matched an annotated GRCh38 gene with concordant isotype and anticodon. Supports the Results section on cataloged shifts.

Table S9. Variants at the alanine identity element. Every gnomAD v4 variant with called genotypes at either member of the G3:U70 base pair in the 38 high-confidence human alanine tRNA genes, giving the gene, the genomic position and which member of the pair it is, the strand, the variant, and its allele count, allele number and frequency. 72 rows across 24 genes; the highest frequency is 0.28%. Supports the Results section on identity-element variants.

Supplementary methods notes

Analyses that were attempted and rejected. Three analyses were run and are not reported as results. Each is recorded here because the reason for rejection constrains how the same questions should be approached.

Variable-arm measurement on genomic sequences. An initial test for serine and leucine identity elements measured variable-arm length on genomic sequences and failed its positive control: serine and leucine genes, which must have the longest arms, did not. tRNA introns inflate apparent arm length in intron-containing isotypes. Repeating the measurement on mature sequences passed the control and is the version reported.

Expression in GSE141436 and GSE272400. Both series returned no reads at the locus, but the result is uninformative. Their reads are untrimmed and 111 nucleotides long, and the apparent matches were Illumina adapter sequence. After adapter trimming, a positive control was run: none of 3,000 sampled reads matched any of the 431 human tRNA genes in the analysis set within six edits. These libraries require their protocol-specific preprocessing, which was not reproduced.

Code availability

All analysis scripts accompany this submission in the analysis_scripts folder. They are plain Python and depend only on standard scientific libraries together with edlib for sequence alignment and statsmodels for the regression. Each script regenerates its output from the public sources named in Methods; no private data are required.

Script index: aim1_parse.py and aim1_stats.py (gene-set parsing and summary statistics); gu70.py (acceptor-stem identity-element scan); acceptor_and_discriminator.py (acceptor-stem pairs, anticodon bases and discriminator across the alanine and valine families, for Table 2); isoacceptor_dosage.py (isoacceptor gene copy number); identity.py (pairwise sequence identity, masked and unmasked); identity_element_scan.py (gnomAD scan at both members of G3:U70, for Table S9); gnomad.py (anticodon variant retrieval); rogers_liftover.py and rogers_match_and_enrichment.py (GRCh37 to GRCh38 mapping and matching of the cataloged shift loci); enrichment_denominator_fix.py (shift-locus enrichment recomputed against the 429 successfully queried genes, for Table 4); fst.py and fst_maf_matched.py (population differentiation with MAF-matched backgrounds); fetch_all.py (cross-species scan, not reported); aim2_scan2.py, aim2_control.py and aim2_bulk.py (expression scan with decoy controls); fig1.py and fig23.py (figures).

Data availability

All primary data are public. tRNA gene sets, structures and mature sequences are from GtRNAdb (https://gtrnadb.ucsc.edu/). Population variation is from gnomAD v4 (https://gnomad.broadinstitute.org/) and dbSNP. Reference sequence and whole-genome alignments are from Ensembl (https://rest.ensembl.org/). Associations are from the GWAS Catalog full release (https://www.ebi.ac.uk/gwas/). Sequencing reads are from NCBI GEO series GSE152621, with GSE141436 and GSE272400 attempted and excluded as described above. The variant described here is rs1233572 at chr6:28,735,467 (GRCh38); the gene is tRNA-Val-AAC-6-1 at chr6:28,735,429–28,735,500, minus strand.