



Biases in arginine codon usage correlate with genetic disease risk

Katharina V. Schulze, PhD^{1,2}, Neil A. Hanchard, MD, PhD^{1,2} and Michael F. Wangler, MD^{1,3}

Purpose: The persistence of hypermutable CGN (CGG, CGA, CGC, CGU) arginine codons at high frequency suggests the possibility of negative selective pressure at these sites and that arginine codon usage could be a predictive indicator of human disease genes.

Methods: We analyzed arginine codons (CGN, AGG, AGA) from all canonical Ensembl protein coding gene transcripts before comparing the frequency of CGN codons between genes with and without human disease associations and with gnomAD constraint metrics.

Results: The frequency of CGN codons among a gene's total arginine codon count was higher in genes linked to syndromic autism spectrum disorder (ASD) compared with genes not associated with ASD. A comparison of genes annotated as dominant or recessive with control genes not matching either classification revealed a progressive increase in CGN codon

frequency. Moreover, CGN frequency was positively correlated with a gene's probability of loss-of-function intolerance (pLI) score and negatively correlated with observed-over-expected ratios for both loss-of-function and missense variants.

Conclusion: Our findings indicate that genes utilizing CGN arginine codons rather than AGG or AGA are more likely to underlie single-gene disorders, particularly for dominant phenotypes, and thus constitute candidate genes for the study of human genetic disease.

Genetics in Medicine (2020) 22:1407–1412; <https://doi.org/10.1038/s41436-020-0813-6>

Keywords: arginine substitution; de novo; codon usage; autism spectrum disorders

INTRODUCTION

Cytosines in a cytosine–guanine (CpG) dinucleotide context are known for their propensity to mutate at a rate that can be several hundred fold greater than transversions at other bases,¹ and these are disproportionately represented in human genetic variation.² This increased rate of variants can be explained by the deamination of methylated cytosine (5mC) that results in thymine, which is not readily detected by DNA repair mechanisms.³ Interestingly, arginine is the only amino acid to contain CpG dinucleotides at the first and second codon positions, which are less redundant than the third position and therefore more likely to lead to an amino acid change. Moreover, codon usage for arginine can have an impact on GC content for a given gene⁴ as arginine is encoded by six codons in vertebrate genomes—four CGN codons (CGA, CGC, CGG, CGU) and two AGR codons (AGG, AGA).

Based on data archived in the database of clinically relevant variants (ClinVar),⁵ arginine substitutions underlie 20.0% of all pathogenic single-nucleotide variants (SNVs), making arginine the most commonly substituted amino acid (Fig. 1a; Supplementary Table 1). Arginine is also the most commonly substituted amino acid among benign variants, but at a

significantly lower frequency (13.9%; Chi-squared test, $p = 5.53 \times 10^{-29}$). Even after accounting for amino acid usage, the number of nonsynonymous SNVs affecting arginine are exceptional relative to other amino acids (Fig. 1b; Supplementary Fig. 1). For pathogenic as well as benign arginine substitutions, the most frequent SNVs are cytosine to thymine (C>T) and guanine to adenine (G>A) transitions—both products of 5mC deamination on the plus and minus DNA strand, respectively (Fig. 1c).

Usage of the four CGN codons in human genes, despite the hypermutability of CpG dinucleotides and existence of AGR alternative codons, suggests the possibility of negative selective pressure at these sites. We therefore hypothesized that the skewing of genes toward CpG codon usage for arginine could be predictive of disease risk, particularly for dominant phenotypes where one variant can produce a phenotype.

MATERIALS AND METHODS

ClinVar SNV analysis

Unique pathogenic and benign nonsynonymous SNV entries were extracted from ClinVar variant summary data (accessed

¹Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA; ²USDA/ARS/Children's Nutrition Research Center, Baylor College of Medicine, Houston, TX, USA; ³Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, USA. Correspondence: Michael F. Wangler (michael.wangler@bcm.edu)

Submitted 30 October 2019; revised 14 April 2020; accepted: 15 April 2020
Published online: 6 May 2020

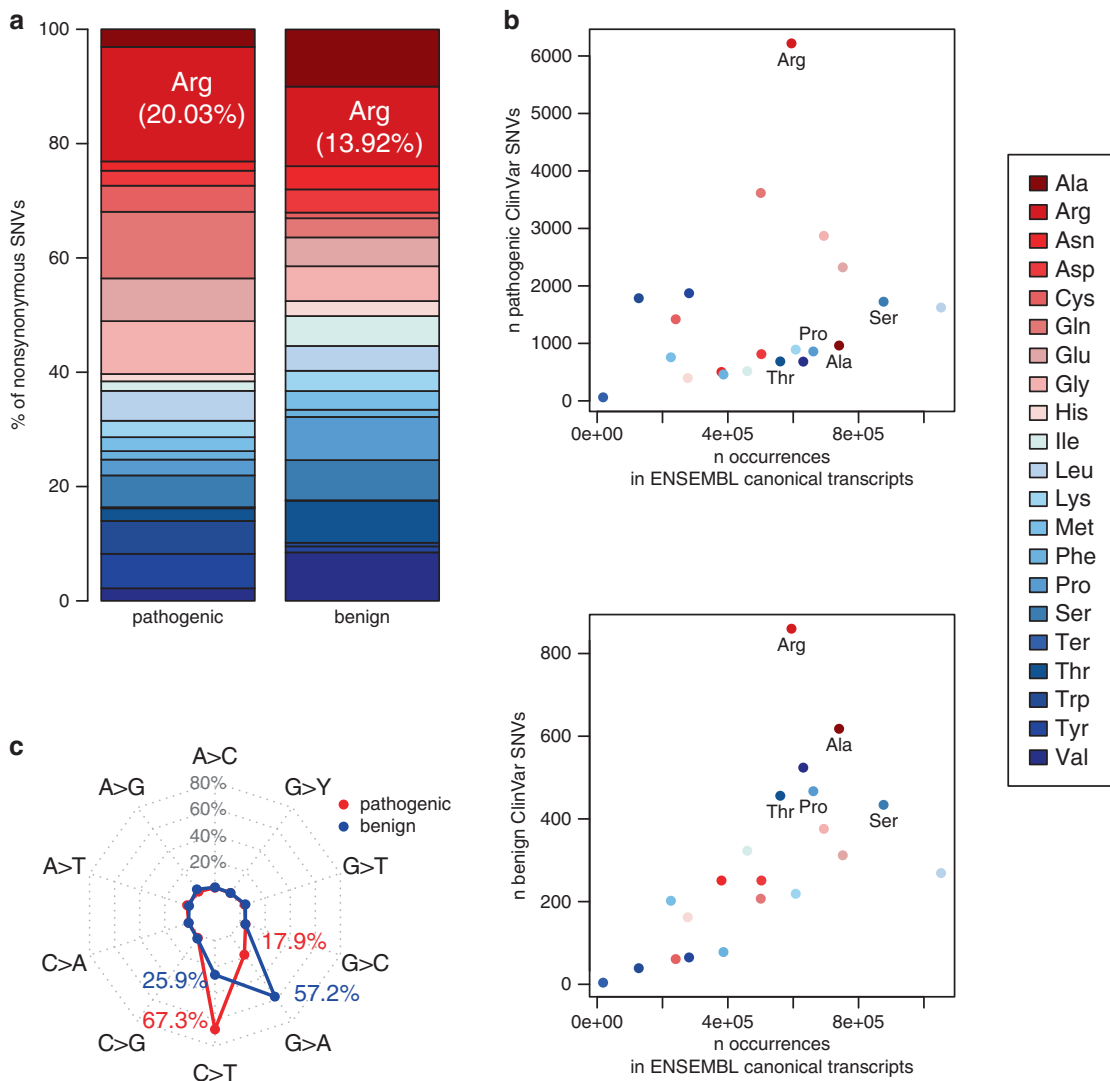


Fig. 1 ClinVar amino acid substitutions. (a) Amino acids with pathogenic and benign nonsynonymous single-nucleotide variants (SNVs) in ClinVar. “Ter” indicates a stop codon. Raw data underlying bar charts are listed in Supplementary Table 1. (b) The number of nonsynonymous, pathogenic ClinVar SNVs relative to the number of occurrences for each amino acid throughout canonical gene transcripts. Labeled amino acids are encoded by codons containing CpG dinucleotides. (c) Arginine substituting single-nucleotide changes for pathogenic and benign SNVs.

18 April 2019). Entries with three letter amino acid code format (e.g., p.Arg330Met; 82.8% of all unique SNV entries) were tallied. One entry of selenocysteine was omitted.

Pathogenic arginine codon usage analysis

All amino acid codons were tallied for each Ensembl coding sequence transcript (GRCh38, release 96)⁶ labeled canonical in the Genome Aggregation Database (gnomAD, v2.1.1).⁷ Trinucleotide counts were obtained from the same coding sequences by removing the first and second base to create reading frame offsets. Disease-associated genes were selected from the Simons Foundation Autism Research Initiative (SFARI; <https://www.sfari.org/resource/sfari-gene>, accessed 31 January 2018), human homologs of essential genes found in a *Drosophila melanogaster* X chromosome screen,⁸ and OMIM (<https://www.omim.org/downloads>, accessed 9 April 2018). To avoid ambiguity, only OMIM genes exclusively associated

with dominant or recessive diseases were used, while those associated with neither were used as controls. Human paralogs were identified with the Ensembl Comparative Genomics—Paralogues tab.

Expression analysis

Using the GRCh38 reference genome (release 96), guanine and cytosine (GC) content was calculated 5, 10, 25, and 50 kb up- and downstream of the canonical transcript boundaries’ genomic coordinates, which were downloaded from Ensembl BioMart. Gene expression data was downloaded from the Genotype-Tissue Expression (GTEx) database (accessed 16 April 2019).⁹

Statistics

CGN codon frequencies were compared between disease-associated genes and those lacking disease association using a

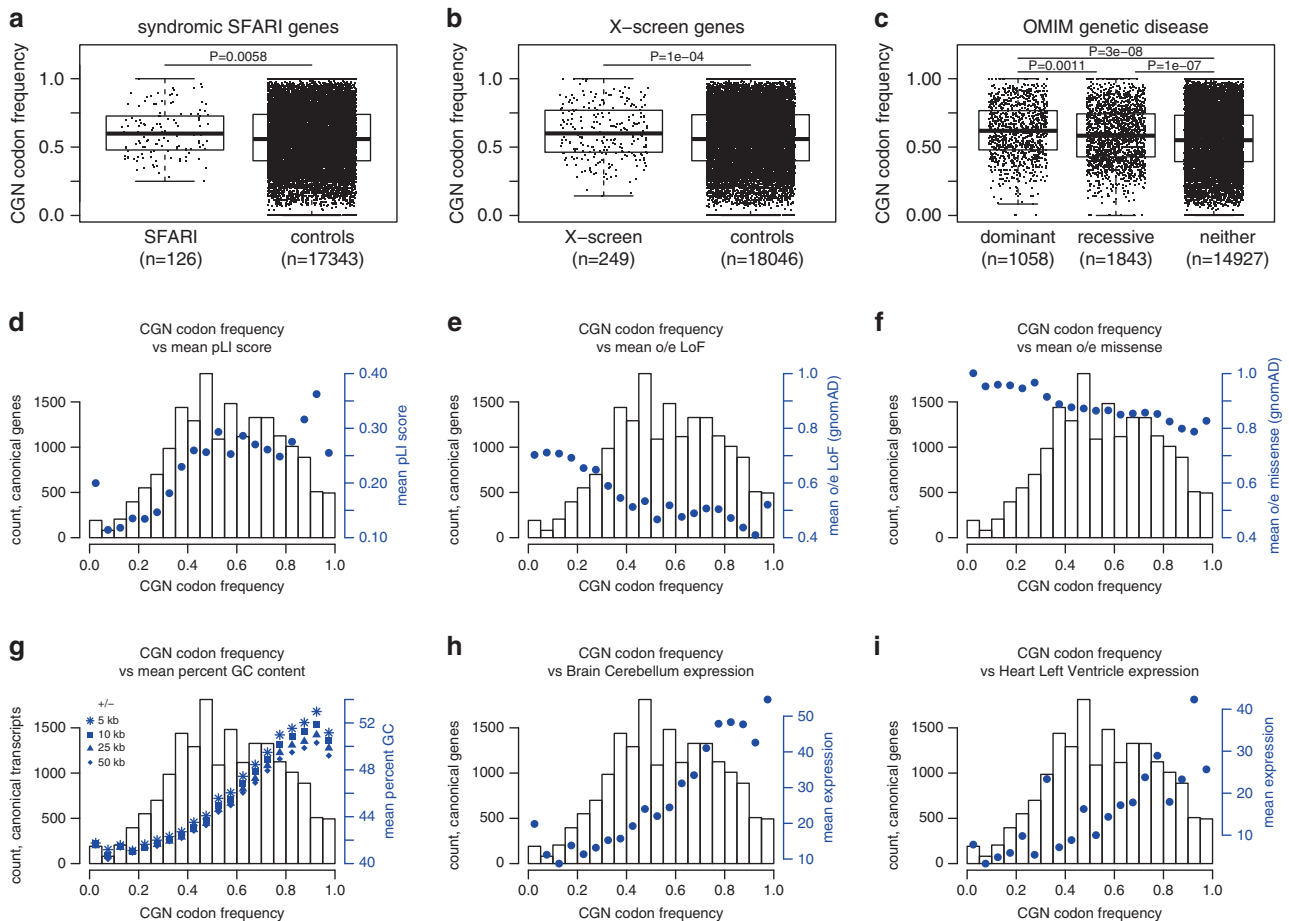


Fig. 2 CGN arginine codon frequency and gene pathogenicity. (a) CGN arginine codon frequency comparison between syndromic SFARI genes and control genes not linked to autism spectrum disease (ASD) in SFARI. (b) CGN arginine codon frequency comparison between human homologs to essential genes detected in *Drosophila* X-screen and all other genes (controls). (c) CGN arginine codon frequency comparison between OMIM disease genes and nondisease gene controls. (d–f) CGN codon frequency comparison with gnomAD constraint metrics: probability of loss-of-function intolerance (pLI) score, observed-over-expected (o/e) ratio of loss-of-function (LoF) and o/e ratio of missense variants. (g) CGN codon frequency comparison with GC content 5, 10, 25, and 50 kb up- and downstream of transcript boundaries. Pearson correlation coefficients were $r_{5\text{ kb}} = 0.537$, $r_{10\text{ kb}} = 0.532$, $r_{25\text{ kb}} = 0.519$, and $r_{50\text{ kb}} = 0.503$. (h–i) Examples of median gene transcripts per million (TPM) for cerebellum and left ventricle.

two-sided, two-sample Welch's *t* test or analysis of variance (ANOVA) in combination with Tukey's honestly significant difference (HSD) tests. Pearson correlation was used to test relationships between CGN codon frequency by gene and gene constraint metrics listed in gnomAD (v2.1.1), GC content, and gene expression levels.

RESULTS

Of all arginine codons found within 18,295 canonical gene transcripts in the human genome, the two AGR arginine codons were identified most frequently (AGA: 21.50%, AGG: 20.81%). However, the total frequency of these two AGR codons (42.31%) was below the total frequency of the four CGN codons (57.69%). Among CGN codons, CGG was the most frequent (20.40%), followed by CGC (18.41%), CGA (10.94%), and CGU (7.93%) (Supplementary Table 2). Therefore, there is not a dramatic depletion of CGN arginine codons, despite their relatively high risk of mutation.

Considering the persistence of CGN codons, we analyzed CGN codon frequency in the context of putative variant pathogenicity. The frequency of CGN codon usage was higher in genes associated with syndromic autism spectrum disorder (ASD) as listed by SFARI ($n = 126$) relative to the CGN arginine codon frequency in genes not in any way associated with ASD ($n = 17,343$; $p = 0.0058$, two-sided Welch's *t* test, Fig. 2a). This was interesting given the high rate of de novo mutation in the ASD gene set.^{10–12} Likewise, human genes that are homologous to genes found in a *Drosophila melanogaster* X chromosome screen for essential genes ($n = 249$)⁸ also showed higher CGN codon frequency compared with all other genes ($n = 18,046$; $p = 0.0001$, two-sided Welch's *t* test, Fig. 2b) suggesting that essential genes in model organisms are homologous to a gene set enriched for “high CGN” genes in the human genome.

To expand our analysis to monogenic Mendelian disorders, we next extracted a list of genes from OMIM that were

exclusively annotated as “dominant” ($n = 1058$) or “recessive” ($n = 1843$). A comparison of these two gene groups alongside control genes that matched neither classification ($n = 14,927$) revealed a significant difference in CGN arginine codon frequency ($p = 3.06E-22$, ANOVA; Fig. 2c) with a trend toward decreasing CGN frequency moving from dominant to recessive to control genes (dominant versus recessive: $p = 0.0011$, Tukey HSD; dominant versus controls: $p = 3.39E-08$, Tukey HSD; recessive versus controls: $p = 1.14E-07$, Tukey HSD). The differences remained significant across 1000 repeated analyses of variance, for which we used a randomly selected subset of non-OMIM control genes that was more comparable in size to the number of recessive and dominant genes ($n = 1500$; $p < 5.0E-5$ in 1000/1000 instances). In addition, we reduced control genes to those unlikely to underlie autosomal dominant disease; genes without OMIM disease association for which more than 20 individuals with homozygous loss-of-function (LoF) variants have been identified in gnomAD ($n = 302$) showed significantly lower CGN arginine codon frequencies compared with genes associated in OMIM with dominant or recessive disease ($p < 2.2E-16$, ANOVA; Supplementary Fig. 2A). Paralogous of these same 302 genes with OMIM disease associations had significantly higher CGN codon frequencies ($p = 1.4E-16$, ANOVA; Supplementary Fig. 2B).

When we expanded our analysis of OMIM disease genes to the amino acids alanine, proline, serine, and threonine that are encoded by NCG codons, we found that the frequency of NCG codons relative to non-NCG codons of the same amino acids was higher for dominant OMIM disease genes, but not for recessive ones when compared between gene groups with exclusively dominant, recessive, or neither OMIM disease association ($p < 0.05$, Tukey HSD; Supplementary Fig. 3). Similarly, we examined CGN versus AGR trinucleotides directly; by looking at these in “reading frame offsets” by one or two bases, we found that dominant genes had higher CGN trinucleotide frequencies than recessive genes (offset of two bases, $p = 2.3E-5$, Tukey HSD) and those without disease association (both offsets, $p < 0.05$, Tukey HSD; Supplementary Fig. 4). Taken together, we see that dominant disease genes in particular have higher frequencies of CpG dinucleotides, but that genes with high CGN arginine codon frequencies in particular are more frequently associated with both dominant and recessive disease.

A closer look at genes with ClinVar SNV entries revealed that the proportion of pathogenic SNVs resulting in arginine substitutions was higher for genes with CGN codon frequencies in the top 25th percentile ($n = 897$) compared with genes with CGN frequencies in the bottom 25th percentile ($n = 855$; $p = 0.0043$, two-sided Welch two-sample t test). Yet, there was no statistically significant difference in the total number of pathogenic SNVs (including all amino acid changes) ($p = 0.2709$) or the total number of arginine codons per gene ($p = 0.2702$). Similarly, the proportion of benign SNVs affecting arginine was higher in genes from the top 25th percentile of CGN frequency ($n = 732$)

relative to genes from the bottom 25th percentile ($n = 714$; $p = 0.0035$), while the total number of SNVs and total arginine codons did not differ ($p = 0.2391$ and $p = 0.4101$, respectively). However, the top benign genes showed a significantly lower proportion of total arginine SNVs than the top pathogenic SNVs ($p = 5.3E-44$).

These results indicated that genes associated with genetic disorders tend to have higher CGN arginine codon frequencies. Further, our analysis of OMIM classifications suggested that CGN codon frequency might have a greater likelihood of impacting the phenotype of heterozygous individuals. To this end, we compared CGN codon frequency with different gnomAD constraint metrics. There was a modest, yet significant, positive correlation between CGN codon frequency and the probability of loss-of-function intolerance (pLI) score ($r = 0.099$, $p = 5.3E-40$, Pearson correlation; Fig. 2d). This trend was stronger in a comparison of CGN frequency with observed-over-expected (o/e) ratios of LoF and missense variants for each gene, showing a significant negative correlation in both instances (LoF: $r = -0.128$, $p = 4.9E-66$, Pearson correlation, Fig. 2e; missense: $r = -0.163$, $p = 4.4E-109$, Pearson correlation, Fig. 2f).

We explored whether these effects could be due to local higher GC content, which can impact gene expression.⁴ We speculated that higher expressed genes in regions of high GC content might contribute to associations between CGN arginine codon frequency and human disease. We found a positive correlation between CGN codon frequency and the proportion of bases 5, 10, 25, and 50 kb up- and downstream of the transcript that were guanine or cytosine ($p < 2.2E-16$ for all distances, Pearson correlation; Fig. 2g). We next compared CGN codon frequency with the median transcripts per million (TPM) for each tissue in GTEx. With the exceptions of liver, minor salivary gland, stomach, testis, and whole blood, gene expression levels were positively correlated (Pearson) with CGN codon frequency in the 48 remaining tissues at a Bonferroni corrected significance threshold ($p < 9.4E-4$; Fig. 2h, i).

Finally, we examined arginine codon usage in the context of a family of human disease genes, namely actin loci. We note that there are six actin paralogs in the human genome (*ACTG2*, *ACTA1*, *ACTA2*, *ACTC1*, *ACTB*, and *ACTG1*). Interestingly, all six human actin genes are Mendelian disease genes, and arginine substitutions at CGN codons have been identified as pathogenic alleles for all six genes. Every actin paralog encodes 18 arginine residues that are highly conserved (Fig. 3a) and encoded either by a CGN (Fig. 3, red) or AGR trinucleotide (Fig. 3, blue). Given the severity of the disease phenotypes, we used the presence of missense substitution in the gnomAD database as a proxy for benign variation in each of the actin genes (Fig. 3a). Interestingly, these benign variants all occurred in CGN codons rather than AGR codons (Fig. 3a). We next compared these arginine residues across the actin paralogs to known pathogenic variants from ClinVar and discovered that all pathogenic alleles also occurred at CGN codons (Fig. 3b). Interestingly,

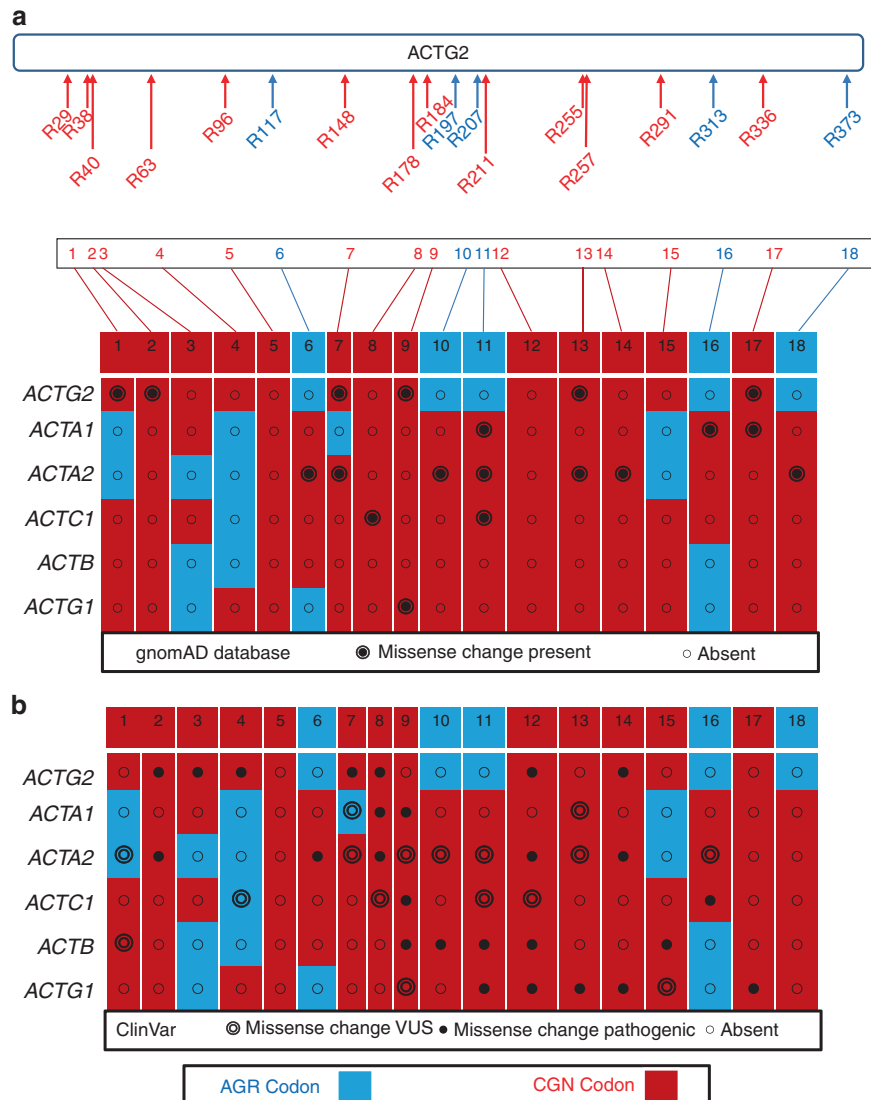


Fig. 3 CGN arginine codon usage and variation across actin gene family. CGN codons are shown in red font or shading, AGR codons in blue. (a) Eighteen conserved actin residues in the *ACTG2* gene are shown at top with numbering corresponding to the *ACTG2* protein. Each numbered arginine is then linked to a table of the actin paralogs. Solid black circle-dot symbols indicate the presence of a missense change in the gnomAD database, a proxy for benign alleles. (b) ClinVar variants corresponding to the same arginine positions in (a) are shown for all homologs. Empty circle-dot symbols indicate variants of uncertain significance (VUS), while solid dots indicate pathogenic or likely pathogenic missense changes. Arginines 8, 9, 12, 13, and 14 are notable for being encoded by CGN codons in all paralogs and having pathogenic alleles across multiple human disease phenotypes.

distinct positions exhibited conserved CGN codon usage despite this hypermutability and were consistently mutated in multiple human diseases (e.g., arginine 12 and 14). In summary, the analysis of arginine codon usage could identify key hotspots for pathogenic alleles across a family of related human disease genes.

DISCUSSION

We analyzed coding regions of the human genome and found that genes with high proportions of CGN arginine codons compared with AGR arginine codons were enriched among known disease genes and correlated with measures of variant intolerance, indicating that genes with CGN codon preferences are more likely to underlie single-gene disorders. Our results

suggest that some genes have retained CGN despite their hypermutability and the often deleterious results on protein function that come from C>T transitions at those sites. As the only amino acid encoded by CpGs at the 1st and 2nd position, studying arginine codon usage provided us with some unique insights into these hotspots for mutation in disease. We speculate that the CGN codons have been anthropologically retained for reasons related to GC content and gene expression, consistent with our observation that genes with higher expression in brain and heart have biased CGN codon usage. Our results provide the possibility of simple gene-level predictions, but by pairing our analysis across highly identical paralogs (such as actin genes) we can see patterns of hypermutable arginine positions associated with disease.

We relied on large public databases, such as ClinVar and gnomAD, for our analyses. On one hand, this allowed us to take a broad, agnostic approach to the study of arginine codon usage. On the other hand, this meant that some of the limitations of these databases also extend to our study. Identifying SNVs as pathogenic, for example, can include the use of damage prediction algorithms, which, depending on their heuristic approach, might more readily identify arginine substitutions as pathogenic compared with other amino acid changes and thus skew the results collected in ClinVar. Constraint metrics provided by gnomAD are limited by the populations included in its database.

Although other amino acids, including cysteine, glutamine, glutamate, tryptophan, and valine, also showed significantly more pathogenic than benign nonsynonymous ClinVar SNVs (Supplementary Table 1), only tryptophan and arginine had disproportionately more ClinVar pathogenic variants than expected relative to the amino acids' frequency throughout coding sequences (Fig. 1b). Unlike arginine, tryptophan is encoded by only one codon that does not involve a CpG dinucleotide.

An outstanding question from our analyses is why such variant-prone CGN arginine codons should persist in the human genome. One speculation is that high GC content might protect against deamination of methylated CpG sites by reducing the rate of DNA melting, thereby decreasing the amount of time spent in a single-stranded state during which CpG sites are more vulnerable to deamination.¹³ Similarly, GC-biased gene conversion could maintain or increase the GC content, and thus CGN codon usage, in certain genomic regions.¹⁴ The positive correlation between CGN codon frequency and GC content observed in our study lends plausibility to both theories.

Classification of genetic variants with regard to their impact on human disease remains a major diagnostic challenge. Here we show that genes with high frequency of CGN codons are enriched for disease loci. Our observations of the relationship between human disease and CGN codon usage could be useful in predicting pathogenicity at the gene and variant level in gene discovery. Moreover, as these sites are highly likely to undergo mutations recurrently in the human population over time, the CGN codons are sites of highly recurrent pathogenic variants. Predicting the most likely mutation sites in disease could prove particularly useful in flagging specific recurrent alleles for oligonucleotide therapy.

SUPPLEMENTARY INFORMATION

The online version of this article (<https://doi.org/10.1038/s41436-020-0813-6>) contains supplementary material, which is available to authorized users.

CODE AVAILABILITY

Scripts used for all analyses are available in Supplementary Materials.

ACKNOWLEDGEMENTS

The authors thank the Genome Aggregation Database (gnomAD) and the groups that provided exome and genome variant data to this resource. A full list of contributing groups can be found at <https://gnomad.broadinstitute.org/about>.

DISCLOSURE

The authors declare no conflicts of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

- Smith T, et al. Extensive variation in the mutation rate between and within human genes associated with Mendelian disease. *Hum Mutat.* 2016;37:488–494.
- Ying H, Huttley G. Exploiting CpG hypermutability to identify phenotypically significant variation within human protein-coding genes. *Genome Biol Evol.* 2011;3:938–949.
- Gao Z, et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc Natl Acad Sci U S A.* 2019;116:9491–9500.
- Karlin S, Mrazek J. What drives codon choices in human genes? *J Mol Biol.* 1996;262:459–472.
- Landrum MJ, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 2016;44:D862–868.
- Zerbino DR, et al. Ensembl 2018. *Nucleic Acids Res.* 2018;46:D754–D761.
- Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536:285–291.
- Yamamoto S, et al. A drosophila genetic resource of mutants to study mechanisms underlying human genetic diseases. *Cell.* 2014;159:200–214.
- GTEX Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013;45:580–585.
- Iossifov I, et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature.* 2014;515:216–221.
- Turner TN, et al. Genomic patterns of de novo mutation in simplex autism. *Cell.* 2017;171:710–e712.
- Sanders SJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature.* 2012;485:237–241.
- Fryxell KJ, Moon WJ. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol Biol Evol.* 2005;22:650–658.
- Pouyet F, Mouchiroud D, Duret L, Semon M. Recombination, meiotic expression and human codon usage. *Elife.* 2017;6:e27344.