ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Cellular Activity of CQWW Nullomer-Derived Peptides

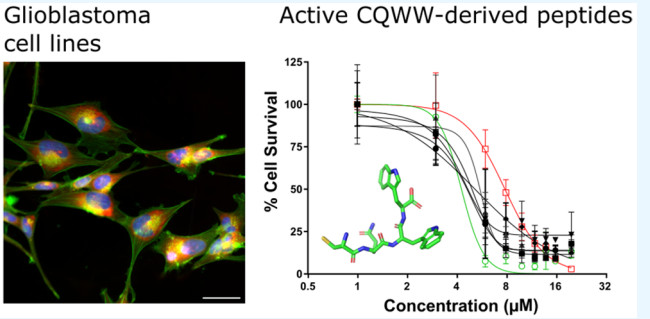
Steven Shave,*,# Rebecka Isaksson,# Nhan T. Pham, Richard J. R. Elliott, John C. Dawson, Julius Soudant, Neil O. Carragher, and Manfred Auer*

Cite This: *ACS Omega* 2025, 10, 6794−6800

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Analysis of observed protein sequences across all species within the UniProtKB/Swiss-Prot data set reveals CQWW as the shortest absent stretch of amino acids. While DNA can be found encoding the CQWW sequence, it has never been observed to be translated or included in manually curated sets of proteins, existing only in predicted, tentative sequences and in a single mature antibody sequence. We have synthesized this "nullomer" peptide, along with 13 derivatives, reversed, truncated, stereoisomers, and alanine-scanning peptides, conjugated to polyarginine stretches to increase cellular uptake. We observed their impact against a healthy neuronal line and six patient-derived glioblastoma cell lines spanning three clinical subtypes. Results reveal IC_{50} values averaging 4.9 μM for inhibition of cell survival across tested oncogenic cell lines. High-content phenotypic analysis of cellular features and reverse-phase protein arrays failed to discern a clear mode of action for the nullomer peptide but suggests mitochondrial impairment through the inhibition of GSK3 and isoforms, supported by observations of reduced mitochondrial stain intensities. With a recent increase in interest in nullomer peptides, we see the results in this study as a starting point for further investigation into this potentially therapeutic peptide class.



INTRODUCTION

The term "nullomers" was introduced in 2007 to refer to DNA sequences absent from the human genome.[1] These nullomers have subsequently been identified across a range of domains and species, finding uses as molecular barcodes in tamper-proof labeling of evidence from crime scenes[2] and in the creation of homology models investigating genome evolution.[3] Vergni and Santoni assigned sequences to a category named "high-order nullomers".[4] These are sequences that remain nullomers when mutations are introduced. In the same work, they analyzed nullomer occurrences in CpG islands, noting them to be present at a higher rate than expected but not high enough to suggest that nullomer formation is driven solely by natural selection. They stated nullomers had "their own peculiar structure and are not simply sequences whose CpG frequency is biased".[4] CpG island nullomer occurrences were, however, given as a potential explanation over natural selection by Acquisti et al.,[5] who argued that the hypermutability of CpG islands contributes to rare sequences becoming nullomers, coupled with the fact that many in humans differ by only one residue, the role of mutation is strengthened in the localization and preservation of nullomers across species.[5] Similarly, Georgakopoulos-Soares et al. use nullomers to build phylogenetic classifiers across vertebrates,[6] while a wider study by Garcia et al. classifies 22 organisms across archaeota, bacteria, and eukaryotes using only the shortest nullomers from each genome.[7] In an approach using codon-translated amino acids, Mouratidis et al.[8] define sets of "quasi-prime" peptides or short peptide k-mers[9,10] as being sequences found in only one species and propose using this data for species identification from biological samples. The idea of searching for absent peptide sequences appears in literature before this terminology referring to DNA was established, with Otaki et al. identifying 12,080 "zero-count" pentamer peptides from 1.5 million protein sequences obtained from public databases.[11] While the absence of some pentamers may be explained by the low occurrence rates of their constituent amino acids, calculations accounting for amino acid propensities demonstrated that many should have been expressed in a theoretical genome. The opposite was also observed, with pentamers composed of low-propensity amino acids being unexpectedly present. Otaki et al. explained this through the evolutionary suppression of these peptide sequences,[11] an explanation that is backed up by Tuller et al. through their discovery that many of the absent pentamer sequences are coded for in noncoding regions of the genome.[12]

Received: September 27, 2024
Revised: January 7, 2025
Accepted: January 23, 2025
Published: February 11, 2025

Attempting to understand potential evolutionary pressure on nullomers, Navon et al. observed expected rates of peptide triplets occurring in the *E. coli* proteome, noting underexpression in only four peptide triplets: CMY, GPP, MWC, and WMC.[13] While examination of triplets undoubtedly leaves out longer interesting sequences, Ung and Winkler identified triplets as occupying a special area of chemical space, possessing optimal ligand efficiency and proposing tripeptide motifs as being the optimal size molecule for biological signaling.[14] Embedding these sequences in GFP and mntA reduced in vivo and in vitro expression levels not only of the proteins themselves but also their unmodified partners during coexpression. This is explained for the CMY and GPP triplets through observed interactions with ribosomal nucleotides A2062 and U2585, known for their involvement in ribosome stalling.[15,16] They note that unlike in *E. coli*, all triplets are observed at expected rates when interrogating the human proteome.[13] In 2019, Mittal et al. reported the counts of all dimer, trimer, tetramer, and pentamer peptides present within UniProtKB/Swiss-Prot, although no synthesis or biological testing was performed on newly discovered nullomer peptides.[17] Later, the propensity of amino acid stretches in ordered versus intrinsically disordered proteins was explored by Mittal et al., leading to the identification of 36 unique tetramer peptides exclusively found within intrinsically disordered proteins[18,19]

Closely related to nullomer peptides, rarely observed amino acid stretches have also been studied by Capone et al., who assert that the minimal sufficient antigenic determinants of a protein could be encoded in just five rarely seen amino acids.[20] Complementing this, Patel et al. noted that these rare sequences could be used to enhance antigen-specific immune responses when dosed alongside adjuvant vaccines.[21] It was also noted by Koulouras et al. that human viruses rarely share human nullomers, facilitating host mimicry and immune evasion,[22] an interesting example of this being highlighted by Silva et al. in their studies on the Ebola virus, identifying human nullomers which consistently appear in two viral Ebola proteins.[23] Rare sequences were observed at higher than expected rates in proto-oncogenes by Trost et al.[24] and Tsiatsianis et al.[25] This points to a potential safety mechanism being inbuilt into these proto-oncogenes, facilitating natural identification and disposal and spurring interest in nullomer epitopes within immunology and oncology for the potential of these rare sequences to aid the immune system in the identification of cancerous cells. In 2012, an early stage drug discovery study by Alileche et al.[26] synthesized pentamer peptide nullomers and discovered two peptides, NWMWC and the permutation WCMNW, that caused mitochondrial impairment in both normal cell lines and cancer cell lines.[26] In a follow-up study, the lethal effect of these pentamers was further explored using the NCI-60 panel[27] containing 60 cell lines derived from human cancers across nine different organs. Both NWMWC and WCMNW were lethal to a high fraction of oncogenic cell lines while not killing the majority of nononcogenic cell lines tested. Notably, the peptides were able to kill both drug-resistant and hormone-resistant prostate and breast cancer cell lines, as well as cancer stem cells,[28] through a mechanism of mitochondrial impairment and ATP depletion.[26] More recently, Ali et al. investigated WCMNW peptide activity in a triple-negative breast cancer mouse model using transcriptomics techniques to reveal the downregulation of key genes involved in the mitochondrial TCA cycle.[29]

Standard medicinal chemistry approaches including conjugation and derivatization of peptides to include non-natural amino acids, cyclization, and other stabilization strategies are now available to overcome many of the limitations present with the use of peptides as therapeutics, improving delivery pharmacokinetics and increasing proteolytic stability.[30,31] Over 80 peptide drugs are currently marketed for a variety of diseases.[32,33] This route from peptide to stable and efficacious treatment suggests the approach of using nullomers as a starting point to find new first-in-class therapeutics is both valuable and viable.

We developed our own implementation of a nullomer and rare sequence discovery algorithm named Aminonaut. This implementation written in Python allows easy integration into existing analysis pipelines and workflows. To facilitate reuse and open science, Aminonaut source code is available on GitHub (https://github.com/stevenshave/Aminonaut) and has also been added to the Python Package Index, allowing installation with a single command to nearly all modern Python environments. We applied Aminonaut to the February 2018 UniProtKB/Swiss-Prot database.[34,35] This article documents our identification and initial biological evaluation of the tetramer nullomer peptide CQWW conjugated to a polyarginine sequence to ensure cellular uptake, with its striking absence begging the question: what specifically makes this sequence "forbidden" in vivo and what might be the biological effects of such a molecule?

RESULTS AND DISCUSSION

Identification and Synthesis of CQWW. We developed the Aminonaut package to specifically look for nullomer and rare protein sequences. Using a *k*-amino acid-sized window, this window is moved across protein sequences, counting occurrences of *k*-mers, where *k* is 2, 3, 4, 5, and 6. Further collation and statistical analysis are also built into the software. We used this software (see Methods) to examine nullomer and rare sequences within the UniProtKB/Swiss-Prot February 2018 release (see Supporting Information). Analysis of pentamer sequences reveals 86,261 nullomer sequences, representing 2.7% out of the total 3.2 million possible pentamers. In line with published results by Mittal et al.,[17] we also identified the single missing tetramer, CQWW, cysteine−glutamine−tryptophan−tryptophan, from the collected tetramer counts. A BLAST search revealed the sequence is coded for in many proposed and hypothetical genomes,[36] ranging from bacteria, oomycetes, to rotifers; however, they are absent from the carefully curated UniProtKB/Swiss-Prot releases, pointing to rare status and absence in proteomes captured in UniProtKB/Swiss-Prot. In 2021, the CQWW sequence was discovered in a human antibody immunoglobulin heavy chain in a study of the autoimmune disease Myasthenia gravis, whereby antibodies target neuromuscular proteins. Analysis of left and right sequence truncates of the CQWW tetrapeptide reveals CQW and QWW present 1251 and 1401 times, respectively, in the data set, well in line with expected occurrence rates. Interestingly, the CQW N-terminal trimer truncate peptide is found in patent literature for antibacterial use against *Listeria* and *Bacillus*.[37] We did not find any biological effects noted for the QWW C-terminal trimer truncate in the literature or patents. CQ, QW, and WW dimers were observed at the expected rates. To find an explanation for the rare status of CQWW, we modeled the peptide (see Methods) to investigate if any intrinsic steric clashes or

https://doi.org/10.1021/acsomega.4c08860
ACS Omega 2025, 10, 6794−6800

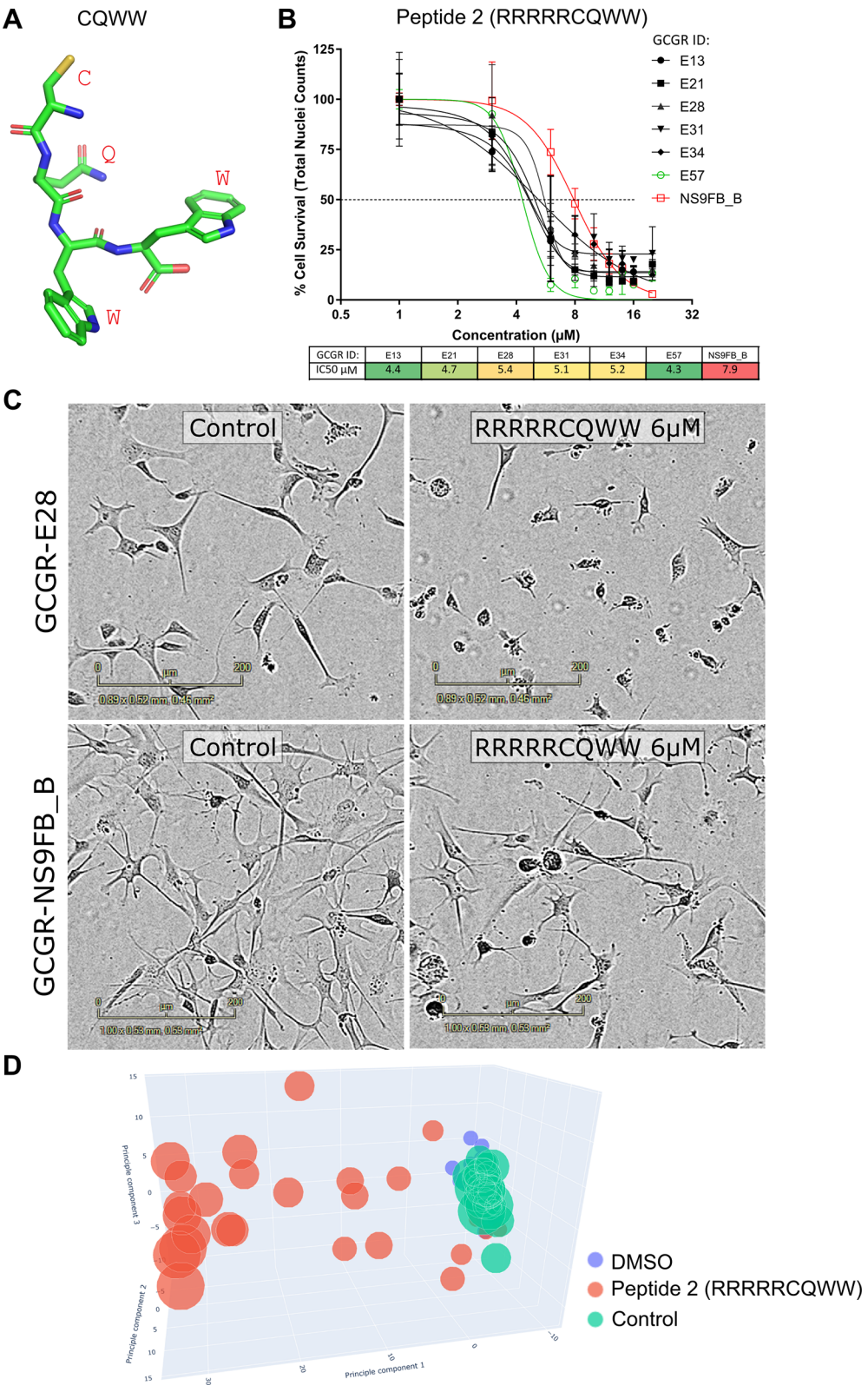


Figure 1. (A) 3D model of the CQWW peptide with residue labels shown in red. (B) Dose-dependent effect of peptide 2 (RRRRRCQWW) on cell survival across a range of cell lines, with fitted IC_{50} values shown below. (C) Representative live cell brightfield images of peptide 2 (RRRRRCQWW, 6 μM) against the glioblastoma stem cell line (GCGR-E28) versus normal neural stem cells (GCGR-NS9FB_B) after 3 h of incubation. (D) 3D principal component analysis (PCA 1, 2, and 3) from multiparametric cell painting image analysis of the GCGR-E13 cell line. A clear dose-dependent phenotypic profile with marker sizes denoting the concentration range of 1−16 μM is seen for peptide 2 (RRRRRCQWW, red) vs polyarginine control (RRRRR, green) and DMSO (blue).

https://doi.org/10.1021/acsomega.4c08860
ACS Omega 2025, 10, 6794−6800

Table 1. Cellular IC_{50}s (μM) as Measured by Nuclei Counts for Nullomer and Derivative Peptides[a]

		Cell line						
Peptide #	Sequence	GCGR-E13	GCGR-E21	GCGR-E28	GCGR-E31	GCGR-E34	GCGR-E57	GCGR-NS9FB_B
2	RRRRR-CQWW	4.4	4.8	5.4	5.1	5.2	4.3	7.9
3	RRRRR-cqww	6.9	7.5	8.4	8.7	5.3	5.4	9.7
4	RRRRR-WWQC	13.1	12.8	10.9	11.8	8.6	9	13.8
5	RRRRR-wwqc	13.6	16.3	13.2	15.7	17.6	9.9	15.9
6	RRRRR-AQWW	150.7	279.2	86	100	100	100	100
7	RRRRR-CAWW	10.6	11.2	9.7	12.1	6	6.3	11.9
8	RRRRR-CQAW	100	75	100	238	100	100	100
9	RRRRR-CQWA	>30	>30	>30	>30	15.76	>30	>30
10	RRRRR-CQW	>100	>100	167.7	>100	153.5	108.1	NA
11	RRRRR-QWW	110.1	158.5	72.6	74.3	114.1	99.5	NA
12	RRRRR-MQWW	117.3	122.1	100.0	59.7	>100	>100	NA
13	RRRRR-FQWW	52.0	81.3	29.6	100.6	>100	55.7	NA
14	RRRRR-HQWW	52.0	132.0	58.0	123.0	>100	119.0	NA
15	RRRRR	30	30	30	30	15.76	30	30

[a]Gradient between blue and red denotes high to low IC_{50}, respectively. Peptides grouped by derivatives (2−5), alanine-scanning peptides (6−9), rare derivative sequences (10−14), and finally control (15).

problematic conformational states were revealed (see Figure 1A). Aqueous solubility prediction was difficult as many online tools require query peptides to be longer than the CQWW tetramer, forcing us to turn to more traditional small-molecule solubility predictors. The SwissADME web service[38] returns a range of predicted solubilities for the free peptide with poor agreement, ranging from 51 nM to 29 mM; however, these techniques are optimized for small molecules, adhering to more traditional medicinal chemistry rules.

The absent tetramer CQWW (peptide 1, Table S3) was synthesized using standard Fmoc peptide synthesis on a solid support (see Supporting Information). Chemical stability was observed via high-performance liquid chromatography (HPLC) (see Supporting Information and Figures S1 and S2 for results) over 72 h in phosphate buffered saline (pH 7.5), resulting in the formation of the disulfide bridged dimer form, confirmed by the addition of TCEP. This disulfide bridge-induced dimer formation was not observed upon the production and evaluation of the polyarginine conjugate RRRRRCQWW (peptide 2, Table S3). For this reason, and to improve peptide cell permeability,[39,40] we produced subsequent peptides in this polyarginine conjugate form, following the approach taken by Alileche et al.[28] of attaching a poly arginine chain to ensure cell penetration, all of which did not dimerize. In addition to peptide 2, we produced a D-stereoisomer form of RRRRRcqww (peptide 3, Table S3), a reversed form of RRRRRWWQC (peptide 4, Table S3), a reversed D-stereoisomer form of RRRRRwwqc (peptide 5, Table S3), alanine-scanning[41] peptides (peptides 6−9, Table S3), and the N-terminal and C-terminal triplet truncates RRRRRCQW and RRRRRQWW (peptides 10 and 11, Table S3). Comparing the observed and expected counts based on either codon frequency or amino acid occurrence rates indicated that a majority of single amino acid mutations of CQWW appear at lower-than-expected frequencies. The highest count being for YQWW (207 occurrences) and the rarest (excluding glutamine) being CMWW (3 occurrences), see Table S1. Variations around the second position (glutamic acid) were found with lower counts than for other positions. To understand the implication of the cysteine residue, we produced low occurrence-count cysteine replacements (see Table S1) using methionine, histidine, and phenylalanine (peptides 12−14, Table S3). In addition to these synthesized peptides, we purchased a polyarginine control peptide, RRRRR (peptide 15, Table S3).

Peptides were tested in an automated image-based high-content[42] cell painting[43,44] assay that was originally developed to explore the mechanism of action of pharmacological and genetic perturbations within cells,[45] as well as a live cell imaging assay against six patient-derived glioblastoma cell lines (GCGR-E13, GCGR-E28, GCGR-E21, GCGR-E57, GCGR-E31, and GCGR-E34), covering classical, mesenchymal, and proneural subtypes, respectively (see Methods and Figure 1B−D), along with a healthy human fetal neuronal stem cell line (GCGR-NS9FB_B).

Cell survival was quantified by nuclei counts from the Hoechst dye-stained nuclei and IC_{50} values calculated along with 95% confidence intervals (Figure 1B; see Table S2 for full details). Table 1 shows a summary of cellular IC_{50}s derived from these nuclei counts.

Alanine-scanning peptides (Table 1, peptides 6−9) clearly demonstrate tolerance for alanine replacement at the second glutamine position, with peptide 7 (RRRRRCAWW) retaining an average IC_{50} of 14.8 μM across cell lines, while alanine replacement in other positions results in IC_{50} values greater than 100 μM (apart from the GCGR-E21 cell line with an IC_{50} of 75.0 μM—see Supporting Information Table S2 for full IC_{50} listings). The phenotypic effect of the peptides was visualized by live cell imaging (Figure 1C, IncuCyte live cell imager) and also quantified via image analysis of cells labeled with the cell

https://doi.org/10.1021/acsomega.4c08860
ACS Omega 2025, 10, 6794−6800

painting reagents (see Supporting Information) followed by image analysis using Cell Profiler (cellprofiler.org). Multiparametric outputs from Cell Profiler were processed using Phenonaut[46] [v 2.0.3] (see Jupyter notebook 'Phenonaut_processing_of_nullomer_data.ipynb' in the Aminonaut repository for methods and code used). Figure 1D shows a 3D principal component plot with a strong, dose-dependent phenotype for the nullomer peptide 2 when compared with the control polyarginine peptide 15 (RRRRR). Peptide 2 also showed approximately 1.5- to 2-fold selectivity for reduction in nuclei counts for glioblastoma stem cells compared to normal neural stem cells (GCGR-NS9FB_B). Peptides 2, 5, and 7 were deemed to have the most merit for mode of action determination using reverse phase protein arrays (RPPAs) (see Supporting Information for methods). Analysis of gene networks showed no clear mode of action, apart from downregulation of GSK-3β (see Supporting Information), which, among other functions, is noted as crucial for mediation of mitochondrial function.[47−50] Analysis of mitochondrial stain intensities captured during the cell painting assay reveals a dose-dependent reduction in intensity for all cell lines (see Supporting Information Figures S3 and S4).

METHODS

Bioinformatics. With an abundance of literature algorithms published to identify nullomers in protein and peptide forms,[1,51−53] we drew on their descriptions to create our own custom implementation, filling a gap in Python bioinformatics tools and better integrating with our existing codebase and workflows. This took the form of a Python (version 3.8.5) program utilizing the NumPy library (version 1.19.1) for efficient array operations. This program was further developed into a suite of tools for interrogation of the UniProtKB/Swiss-Prot database and is available under the open source MIT license as a source code repository at https://github.com/stevenshave/Aminonaut.

After downloading the XML version of the February 2018 UniProtKB/Swiss-Prot database,[34,35] the find_nullomer_motifs.py program was run, passing as arguments the XML file, followed by an output CSV file, and finally, the length of peptides to be counted. This was run for lengths of 2, 3, 4, 5, and 6 amino acids, directing output to different CSV files for each length. The 5- and 6-mers were captured with a ".csv.gz" file extension, directing the program to apply gzip compression.

Peptides were synthesized using a standard Fmoc solid-phase peptide synthesis. Further details are given in Supporting Information.

Modeling of the CQWW peptide was achieved using the PEPstrMOD service[54] with default settings, embedding the peptide between alanine dimers to model a sequence of eight amino acids representative of the CQWW peptide embedded within a protein. Removal of flanking alanine residues and visualization of the remaining CQWW peptide was achieved using PyMol (v 2.5.0a0).

Methods for cell culture and phenotypic profiling are detailed in the Supporting Information accompanying this article.

CONCLUSIONS

In this article, we have demonstrated the steps taken in our identification of the short CQWW peptide nullomer conjugated to a polyarginine sequence and profiled its biological activity using live cell and high-content imaging along with proteomics analysis using RPPA. While mechanism-of-action analysis with RPPA was inconclusive, perturbation of GSK-3 isoforms (see Supporting Information) aligns with the mitochondrial activity noted by Alileche et al.[28] and our observed concentration-dependent reduction in mitochondria stain intensity. Alanine scanning shows a clear tolerance for changes to the glutamine in position 2, with efficacy retained upon replacement with an alanine. Interestingly, the backward and D-stereoisomer sequences retain similar activity across cell lines. The use of peptides for fundamental biology and drug discovery is rapidly increasing due to the size of addressable and explorable peptide chemical space, ease of chemical manufacture or biological expression, and their massive molecular recognition potential. To this end, and with peptidic drugs making up to 7% of new US FDA approvals from 2015 to 2019,[55] we are said to soon be facing "the coming peptide tidal wave".[56] While nullomers and other rare peptides are poorly understood, it is vital that the body of knowledge on these special sequences is expanded and their full potential explored in the search for new first-in-class therapeutics.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c08860.

Methods for chemical synthesis, cell culture, phenotypic screening, and RPPA analysis (PDF)

AUTHOR INFORMATION

Corresponding Authors

Steven Shave − *Edinburgh Cancer Research, Cancer Research UK Scotland Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XR, U.K.; School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, U.K.;* orcid.org/0000-0001-6996-3663; Email: s.shave@ed.ac.uk

Manfred Auer − *School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, U.K.;* orcid.org/0000-0001-8920-3522; Email: manfred.auer@ed.ac.uk

Authors

Rebecka Isaksson − *School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, U.K.; Department of Chemistry, University College London, London WC1H 0AJ, U.K.*

Nhan T. Pham − *Edinburgh Cancer Research, Cancer Research UK Scotland Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XR, U.K.; School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, U.K.; College of Medicine and Veterinary Medicine, University of Edinburgh, Institute for Regeneration and Repair, Edinburgh EH16 4UU, U.K.;* orcid.org/0000-0003-1620-2910

Richard J. R. Elliott − *Edinburgh Cancer Research, Cancer Research UK Scotland Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XR, U.K.*

John C. Dawson − *Edinburgh Cancer Research, Cancer Research UK Scotland Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XR, U.K.*

Julius Soudant − *Edinburgh Cancer Research, Cancer Research UK Scotland Centre, Institute of Genetics and*

https://doi.org/10.1021/acsomega.4c08860
ACS Omega 2025, 10, 6794−6800

Cancer, University of Edinburgh, Edinburgh EH4 2XR, U.K.; Departamento de Farmacologia, Facultad de Medicina, Universidad Autónoma de Madrid, Madrid 28029, Spain; orcid.org/0009-0000-0975-8385

Neil O. Carragher − *Edinburgh Cancer Research, Cancer Research UK Scotland Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XR, U.K.;* orcid.org/0000-0001-5541-9747

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c08860

Author Contributions

#S.S. and R.I. contributed equally to this work. S.S. and M.A. conceived the project. S.S. performed programming and computational analysis, which led to the discovery of the CQWW peptide sequence. The manuscript was written through contributions by S.S., R.I., N.T.P., and M.A. Peptide synthesis was carried out by R.I. with laboratory support provided by N.T.P. Phenotypic screening and image analysis were carried out by R.J.R.E., J.C.D., and N.O.C. J.S. performed RPPA. M.A. obtained and administered funding for the project. All authors were involved in editing the manuscript. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S.S., R.I., and M.A. acknowledge financial support from the Leverhulme Trust (Grant RPG-2019-071). S.S., M.A., and N.T.P. acknowledge financial support from the Scottish Universities Life Sciences Alliance (SULSA- http://www.sulsa.ac.uk) and the Medical Research Council (MRC- www.mrc.ac.uk, J54359) Strategic Grant. M.A. and N.T.P. acknowledge financial support from the Wellcome Trust (Grant 201531/Z/16/Z). S.S. and N.O.C. acknowledge financial support from the Medical Research Council (MRC- www.mrc.ac.uk, W003996). R.J.R.E., J.C.D., and N.O.C. acknowledge financial support from a joint Cancer Research UK (A28596) and The Brain Tumour Charity award (GN419 000676). We wish to thank Gillian Morrison from the Cancer Research UK Glioma Cellular Genetics Resource for provision of human GBM stem cell lines and normal neural stem cells. For the purpose of open access, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

■ ABBREVIATIONS

ATP, adenosine triphosphate; BLAST, basic local alignment search tool; CSV, comma-separated values; DNA, deoxyribonucleic acid; Fmoc, fluorenylmethyloxycarbonyl; GFP, green fluorescent protein; GSK3, glycogen synthase kinase 3; HPLC, high-performance liquid chromatography; NCI, National Cancer Institute; PTC, peptidyl transferase center; RPPA, reverse phase protein array; TCA, tricarboxylic acid cycle; TCEP, tris(2-carboxyethyl)phosphine; US FDA, United States Food and Drug Administration; XML, extensible markup language; mntA, manganese ABC transporter protein A

■ REFERENCES

(1) Hampikian, G.; Andersen, T. Absent sequences: nullomers and primes. *Pac. Symp. Biocomput.* **2006**, *12*, 355−366.

(2) Goswami, J.; Davis, M. C.; Andersen, T.; Alileche, A.; Hampikian, G. Safeguarding forensic DNA reference samples with nullomer barcodes. *J. Forensic Leg Med.* **2013**, *20* (5), 513−519.

(3) Garcia, S. P.; Pinho, A. J. Minimal absent words in four human genome assemblies. *PLoS One* **2011**, *6* (12), No. e29344.

(4) Vergni, D.; Santoni, D. Nullomers and High Order Nullomers in Genomic Sequences. *PLoS One* **2016**, *11* (12), No. e0164540.

(5) Acquisti, C.; Poste, G.; Curtiss, D.; Kumar, S. Nullomers: really a matter of natural selection? *PLoS One* **2007**, *2* (10), No. e1022.

(6) Georgakopoulos-Soares, I.; Yizhar-Barnea, O.; Mouratidis, I.; Hemberg, M.; Ahituv, N. Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. *Genome Biol.* **2021**, *22* (1), 245.

(7) Garcia, S. P.; Pinho, A. J.; Rodrigues, J. M.; Bastos, C. A.; Ferreira, P. J. Minimal absent words in prokaryotic and eukaryotic genomes. *PLoS One* **2011**, *6* (1), No. e16065.

(8) Mouratidis, I.; Chan, C. Y.; Chantzi, N.; Tsiatsianis, G. C.; Hemberg, M.; Ahituv, N.; Georgakopoulos-Soares, I. Quasi-prime peptides: identification of the shortest peptide sequences unique to a species. *NAR:Genomics Bioinf.* **2023**, *5* (2), lqad039.

(9) Kang, D. D.; Froula, J.; Egan, R.; Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **2015**, *3*, No. e1165.

(10) Moeckel, C.; Mareboina, M.; Konnaris, M. A.; Chan, C. S.; Mouratidis, I.; Montgomery, A.; Chantzi, N.; Pavlopoulos, G. A.; Georgakopoulos-Soares, I. A survey of k-mer methods and applications in bioinformatics. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2289−2303.

(11) Otaki, J. M.; Ienaka, S.; Gotoh, T.; Yamamoto, H. Availability of short amino acid sequences in proteins. *Protein Sci.* **2005**, *14* (3), 617−625.

(12) Tuller, T.; Chor, B.; Nelson, N. Forbidden penta-peptides. *Protein Sci.* **2007**, *16* (10), 2251−2259.

(13) Navon, S. P.; Kornberg, G.; Chen, J.; Schwartzman, T.; Tsai, A.; Puglisi, E. V.; Puglisi, J. D.; Adir, N. Amino acid sequence repertoire of the bacterial proteome and the occurrence of untranslatable sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113* (26), 7166−7170.

(14) Ung, P.; Winkler, D. A. Tripeptide motifs in biology: targets for peptidomimetic design. *J. Med. Chem.* **2011**, *54* (5), 1111−1125.

(15) Koch, M.; Willi, J.; Pradere, U.; Hall, J.; Polacek, N. Critical 23S rRNA interactions for macrolide-dependent ribosome stalling on the ErmCL nascent peptide chain. *Nucleic Acids Res.* **2017**, *45* (11), 6717−6728.

(16) Vazquez-Laslop, N.; Ramu, H.; Klepacki, D.; Kannan, K.; Mankin, A. S. The key function of a conserved and modified rRNA residue in the ribosomal response to the nascent peptide. *EMBO J.* **2010**, *29* (18), 3108−3117.

(17) Mittal, A.; Changani, A. M.; Taparia, S. What limits the primary sequence space of natural proteins? *J. Biomol. Struct. Dyn.* **2020**, *38* (15), 4579−4583.

(18) Mittal, A.; Changani, A. M.; Taparia, S.; Goel, D.; Parihar, A.; Singh, I. Structural disorder originates beyond narrow stoichiometric margins of amino acids in naturally occurring folded proteins. *J. Biomol. Struct. Dyn.* **2021**, *39* (7), 2364−2375.

(19) Mittal, A.; Changani, A. M.; Taparia, S. Unique and exclusive peptide signatures directly identify intrinsically disordered proteins from sequences without structural information. *J. Biomol. Struct. Dyn.* **2021**, *39* (8), 2885−2893.

(20) Capone, G.; De Marinis, A.; Simone, S.; Kusalik, A.; Kanduc, D. Mapping the human proteome for non-redundant peptide islands. *Amino Acids* **2008**, *35* (1), 209−216.

(21) Patel, A.; Dong, J. C.; Trost, B.; Richardson, J. S.; Tohme, S.; Babiuk, S.; Kusalik, A.; Kung, S. K.; Kobinger, G. P. Pentamers not found in the universal proteome can enhance antigen specific immune responses and adjuvant vaccines. **2012**.7, e43802, ,

(22) Koulouras, G.; Frith, M. C. Significant non-existence of sequences in genomes and proteomes. *Nucleic Acids Res.* **2021**, *49* (6), 3139−3155.

(23) Silva, R. M.; Pratas, D.; Castro, L.; Pinho, A. J.; Ferreira, P. J. Three minimal sequences found in Ebola virus genomes and absent from human DNA. *Bioinformatics* **2015**, *31* (15), 2421−2425.
(24) Trost, B.; Kanduc, D.; Kusalik, A. Rare peptide segments are found significantly more often in proto-oncoproteins than control proteins: implications for immunology and oncology. *J. R. Soc. Interface* **2009**, *6* (30), 123−127.
(25) Tsiatsianis, G. C.; Chan, C. S.; Mouratidis, I.; Chantzi, N.; Tsiatsiani, A. M.; Yee, N. S.; Zaravinos, A.; Kantere, V.; Georgakopoulos-Soares, I. Peptide absent sequences emerging in human cancers. *Eur. J. Cancer* **2024**, *196*, 113421.
(26) Alileche, A.; Goswami, J.; Bourland, W.; Davis, M.; Hampikian, G. Nullomer derived anticancer peptides (NulloPs): differential lethal effects on normal and cancer cells in vitro. *Peptides* **2012**, *38* (2), 302−311.
(27) Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **2006**, *6* (10), 813−823.
(28) Alileche, A.; Hampikian, G. The effect of Nullomer-derived peptides 9R, 9S1R and 124R on the NCI-60 panel and normal cell lines. *BMC Cancer* **2017**, *17* (1), 533.
(29) Ali, N.; Wolf, C.; Kanchan, S.; Veerabhadraiah, S. R.; Bond, L.; Turner, M. W.; Jorcyk, C. L.; Hampikian, G. 9S1R nullomer peptide induces mitochondrial pathology, metabolic suppression, and enhanced immune cell infiltration, in triple-negative breast cancer mouse model. *Biomed. Pharmacother.* **2024**, *170*, 115997.
(30) Otvos, L.; Wade, J. D. Current challenges in peptide-based drug discovery. *Front Chem.* **2014**, *2*, 62.
(31) Frackenpohl, J.; Arvidsson, P. I.; Schreiber, J. V.; Seebach, D. The outstanding biological stability of β-and γ-peptides toward proteolytic enzymes: an in vitro investigation with fifteen peptidases. *ChemBioChem* **2001**, *2* (6), 445−455.
(32) Muttenthaler, M.; King, G. F.; Adams, D. J.; Alewood, P. F. Trends in peptide drug discovery. *Nat. Rev. Drug Discovery* **2021**, *20* (4), 309−325.
(33) Wang, L.; Wang, N.; Zhang, W.; Cheng, X.; Yan, Z.; Shao, G.; Wang, X.; Wang, R.; Fu, C. Therapeutic peptides: current applications and future directions. *Signal Transduct. Targeted Ther.* **2022**, *7* (1), 48.
(34) Boutet, E.; Lieberherr, D.; Tognolli, M.; Schneider, M.; Bansal, P.; Bridge, A. J.; Poux, S.; Bougueleret, L.; Xenarios, I. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. In *Plant Bioinformatics*; Springer, 2016; pp 23−54.
(35) Bateman, A.; Martin, M. J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480−D489.
(36) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: architecture and applications. *BMC Bioinf.* **2009**, *10*, 421.
(37) Adir, N.; Navon, S.; Swartzmann, T. Anti-microbial peptides and uses of same. WO 2012153337 A3, 2012.
(38) Daina, A.; Michielin, O.; Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017**, *7*, 42717.
(39) Schmidt, N.; Mishra, A.; Lai, G. H.; Wong, G. C. Arginine-rich cell-penetrating peptides. *FEBS Lett.* **2010**, *584* (9), 1806−1813.
(40) Wang, T. Y.; Sun, Y.; Muthukrishnan, N.; Erazo-Oliveras, A.; Najjar, K.; Pellois, J. P. Membrane Oxidation Enables the Cytosolic Entry of Polyarginine Cell-penetrating Peptides. *J. Biol. Chem.* **2016**, *291* (15), 7902−7914.
(41) Morrison, K. L.; Weiss, G. A. Combinatorial alanine-scanning. *Curr. Opin. Chem. Biol.* **2001**, *5* (3), 302−307.
(42) Way, G. P.; Sailem, H.; Shave, S.; Kasprowicz, R.; Carragher, N. O. Evolution and impact of high content imaging. *SLAS Discovery* **2023**, *28* (7), 292−305.
(43) Bray, M.-A.; Singh, S.; Han, H.; Davis, C. T.; Borgeson, B.; Hartland, C.; Kost-Alimova, M.; Gustafsdottir, S. M.; Gibson, C. C.; Carpenter, A. E. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **2016**, *11* (9), 1757−1774.
(44) Cimini, B. A.; Chandrasekaran, S. N.; Kost-Alimova, M.; Miller, L.; Goodale, A.; Fritchman, B.; Byrne, P.; Garg, S.; Jamali, N.; Logan, D. J.; et al. Optimizing the Cell Painting assay for image-based profiling. *Nat. Protoc.* **2023**, *18* (7), 1981−2013.
(45) Gustafsdottir, S. M.; Ljosa, V.; Sokolnicki, K. L.; Anthony Wilson, J.; Walpita, D.; Kemp, M. M.; Petri Seiler, K.; Carrel, H. A.; Golub, T. R.; Schreiber, S. L.; et al. Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One* **2013**, *8* (12), No. e80999.
(46) Shave, S.; Dawson, J. C.; Athar, A. M.; Nguyen, C. Q.; Kasprowicz, R.; Carragher, N. O. Phenonaut: multiomics data integration for phenotypic space exploration. *Bioinformatics* **2023**, *39* (4), btad143.
(47) Yang, K.; Chen, Z.; Gao, J.; Shi, W.; Li, L.; Jiang, S.; Hu, H.; Liu, Z.; Xu, D.; Wu, L. The key roles of GSK-3β in regulating mitochondrial activity. *Cell. Physiol. Biochem.* **2017**, *44* (4), 1445−1459.
(48) Bijur, G. N.; Jope, R. S. Glycogen synthase kinase-3β is highly activated in nuclei and mitochondria. *Neuroreport* **2003**, *14* (18), 2415−2419.
(49) Chiara, F.; Rasola, A. GSK-3 and mitochondria in cancer cells. *Front. Oncol.* **2013**, *3*, 16.
(50) Wang, L.; Li, J.; Di, L. j. Glycogen synthesis and beyond, a comprehensive review of GSK3 as a key regulator of metabolic pathways and a therapeutic target for treating metabolic diseases. *Med. Res. Rev.* **2022**, *42* (2), 946−982.
(51) Xiao, M.; Li, J.; Hong, S.; Yang, Y.; Li, J.; Wang, J.; Yang, J.; Ding, W.; Zhang, L. K-mer Counting: memory-efficient strategy, parallel computing and field of application for Bioinformatics. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; IEEE, 2018; pp 2561−2567.
(52) Pinho, A. J.; Ferreira, P. J.; Garcia, S. P.; Rodrigues, J. M. On finding minimal absent words. *BMC Bioinf.* **2009**, *10* (1), 137.
(53) Falda, M.; Fontana, P.; Barzon, L.; Toppo, S.; Lavezzo, E. keeSeek: searching distant non-existing words in genomes for PCR-based applications. *Bioinformatics* **2014**, *30* (18), 2662−2664.
(54) Singh, S.; Singh, H.; Tuknait, A.; Chaudhary, K.; Singh, B.; Kumaran, S.; Raghava, G. P. PEPstrMOD: structure prediction of peptides containing natural, non-natural and modified residues. *Biol. Direct* **2015**, *10* (1), 73.
(55) de la Torre, B. G.; Albericio, F. Peptide Therapeutics 2.0. *Molecules* **2020**, *25* (10), 2293.
(56) Kruger, R. P. The Coming Peptide Tidal Wave. *Cell* **2017**, *171* (3), 497.

https://doi.org/10.1021/acsomega.4c08860
ACS Omega 2025, 10, 6794−6800