Supporting Information Appendix for

Structural basis for carotenoid cleavage by an archaeal carotenoid dioxygenase

Anahita Daruwalla[1,2], Jianye Zhang[1], Ho Jun Lee[1], Nimesh Khadka[2], Erik R. Farquhar[3,4], Wuxian Shi[3], Johannes von Lintig[2], and Philip D. Kiser[1,5,6,✉]

[1] Department of Physiology & Biophysics, University of California School of Medicine, Irvine, CA, USA

[2] Department of Pharmacology, Case Western Reserve University School of Medicine, Cleveland, OH, USA

[3] National Synchrotron Light Source-II, Brookhaven National Laboratory, Upton, NY, USA

[4] Center for Proteomics and Bioinformatics, Center for Synchrotron Biosciences, Case Western Reserve University School of Medicine, Cleveland, OH, USA

[5] Research Service, VA Long Beach Healthcare System, Long Beach, CA, USA

[6] Department of Ophthalmology, University of California School of Medicine, Irvine, CA, USA

✉ Corresponding Author: Philip D. Kiser, Department of Physiology & Biophysics, University of California School of Medicine, 837 Health Sciences Road, Gillespie Neuroscience Research Facility, Room 2109, Irvine, CA 92617. Tel.: (949) 824-6954; Fax: 949-824-8540; E-mail: pkiser@uci.edu

Contents:

www.pnas.org/cgi/doi/10.1073/pnas.2004116117

Materials and Methods

Reagents – Except as noted below, chemical reagents were purchased from Sigma-Aldrich (St. Louis, MO) or USB Biochemicals (Cleveland, OH) in the highest purity form available. Water from a Milli-Q purification system (resistivity = 18.2 mΩ·cm) (ED Millipore, Billerica, MA) was used to prepare all reagents and buffered solutions.

Archaeal CCD identification and phylogenetic analysis – Archaeal sequences were identified using the BlastP webserver (1) based on their sequence similarity to bovine RPE65 (NCBI accession code NP_776878.1) and *Synechocystis sp.* PCC6803 ACO (NCBI accession code P74334.1). Sequences from major archaeal (super)phyla were selected for study according to the magnitude of their E-values. The distribution of CCDs in archaea was further inspected with the Pfam database (2). A phylogeny was generated based on an alignment of these archaeal CCDs with previously characterized CCDs from bacteria and eukarya, including sequences from plants and animals. CCD protein sequences were collected from the NCBI protein sequence repository and aligned with Clustal Omega using the default parameters (3). A CCD phylogeny was computed using MrBayes v3.2.6. (4) assuming a uniform prior probability, inverse gamma rate of substitution across sites with four rate categories and the WAG substitution matrix (chosen by Markov chain sampling). The posterior probability distribution was estimated by Markov Chain Monte Carlo sampling for 1,098,000 generations with samples taken every 100 generations after a burn-in fraction of 25%, at which point an average standard deviation of split frequencies of 0.001 was reached. Bipartitions with posterior probabilities < 0.5 were collapsed. The majority-rule consensus tree was visualized using iTOL webserver (5). Analysis of the *Ca.* Nitrosotalea devanaterra genome (GenBank ID: LN890280.1) for horizontal gene transfer candidates was conducted with the program DarkHorse (6) using the NCBI Genbank non-redundant protein and taxonomy databases. BLAST queries for all predicted proteins encoded in the genome were carried out using the program Diamond (7). A maximum of 500 target sequences with E-values < 10^{-5} were accepted for each search. DarkHorse was run using a filter threshold of 0.15, which was empirically determined by tracking the maximum candidate set size. In addition to the standard exclusion terms, a single self-exclusion term 'nitrosocosmicus exaquare' was used. This organism was excluded as it was the sole ammonia oxidizing archaeon possessing a CCD with > 50% sequence identity *Nd*CCD and likely was also acquired by horizontal gene transfer.

Protein expression studies – The coding sequences of the archaeal CCDs and *Coxiella brunetii* CCD shown in **Table S1** were synthesized with codons optimized for expression in *E. coli* (GenScript, Piscataway, NJ) and cloned into the pET3a expression vector. The integrity of the expression plasmids was confirmed by Sanger sequencing. Plasmids were transformed into T7 express *Escherichia coli* cells (New England Biolabs) for protein expression. One-liter cultures of LB media (USB Biochemicals) containing 100 μg of ampicillin/mL were grown in a shaker-incubator at 37 °C to an $OD_{600\ nm}$ of 0.5 – 0.6 at which time the culture was supplemented with an additional 100 μg/mL of ampicillin and the temperature lowered to 28 °C. After overnight growth (12–14 h), bacterial pellets for expression analysis were isolated from 1 mL culture samples by centrifugation. The remaining cells were harvested by centrifugation, suspended in 10 mM HEPES-NaOH, pH 7, and stored at −80 °C.

Production of cobalt-substituted NdCCD – Cobalt-substituted *Nd*CCD was generated by cultivating *E. coli* cultures in M9 minimal media (1x M9 salts, 2 mM $MgSO_4$, 0.4% w/v glucose, 0.8% v/v glycerol and 0.1 mM $CaCl_2$) in the presence of cobalt. All glassware was washed with Milli-Q water containing 1 mM EDTA and ultra-low iron content reagents were used whenever possible to reduce iron contamination. *E. coli* colonies transformed with the pET3a-*Nd*CCD plasmid were inoculated into 2 mL of LB media containing 100 μg/mL of ampicillin and grown for ~ 4 h at 37 °C. Cells were collected by centrifugation

at 3,220*g* for 15 min at 37 °C. Cell pellets were then suspended in 1 mL of M9 minimal media and transferred into 1 L of M9 minimal media containing 100 μg/mL ampicillin that had been pre-warmed to 37 °C. The cell culture was grown at 37 °C with 190 rpm shaking to an $OD_{600\ nm}$ of 0.5–0.6. IPTG was then added to a final concentration of 0.1 mM to induce protein expression and the temperature was lowered to 28 °C. $CoCl_2 \cdot 6H_2O$ (35 mg) was then added to the culture. After overnight growth, cells were harvested by centrifugation and resuspended in 25 mL of 10 mM HEPES-NaOH, pH 7.

NdCCD purification – All purification procedures were conducted at 4 °C using a Bio-Rad FPLC system. *E. coli* containing *Nd*CCD were lysed by two passes through a French press. The lysate was clarified by centrifugation at 186,000*g* for 30 min and then diluted two-fold with pure water. The diluted supernatant was then loaded onto a HiTrap Q HP 5 mL anion-exchange column (GE Healthcare) equilibrated with buffer A consisting of 10 mM HEPES-NaOH, pH 7.0. The column was then washed with 10 column volumes of buffer A and bound *Nd*CCD was eluted by linear increase in buffer B containing 10 mM HEPES-NaOH, pH 7, and 1 M NaCl up to 15% followed by a step increase in buffer B to 100% to elute more tightly bound proteins. Fractions containing *Nd*CCD as identified by SDS-PAGE were pooled and concentrated. The sample was then purified by gel filtration chromatography on a Superdex 200 (120 mL) column pre-equilibrated with buffer A (GE Healthcare). In later experiments, material obtained from the initial anion exchange purification was diluted in buffer A and further purified on a 1 mL MonoQ column (0-10% linear gradient of buffer B) to help remove a ~25 kDa contaminant. For apocarotenoid co-purification experiments, a 20 mM ethanolic solution of 3-hydroxy-β-apo-12'-carotenal was added to the protein solution prior to each step of chromatography to a final concentration of 25 μM followed by a 15 min incubation on ice. No precipitation of the apocarotenoid was observed following its addition to the protein solution. Fractions containing purified *Nd*CCD (≥90% pure as judged by SDS-PAGE analysis) were pooled, concentrated to 70 mg/mL, flash frozen, and stored at -80 °C or placed on ice for immediate use. The apparent purity of *Nd*CCD obtained from LB and M9 minimal media were comparable, being >95% pure as judged by Coomassie-stained SDS-PAGE gels. Quantitative amino acid analysis on purified protein samples confirmed the sample purity and provided an experimentally determined extinction coefficient ($\varepsilon_{280\ nm}$) for *Nd*CCD in 10 mM HEPES-NaOH, pH 7 at 22 °C of 74,688 $M^{-1} \cdot cm^{-1}$.

Metal content analysis – Five mg of purified protein was digested in 5 mL of 1% v/v HNO_3 for 3 h at room temperature. Protein precipitate was then removed by filtration of the sample through a 0.22 μm membrane (Millipore). A buffer-only control sample was prepared identically to allow assessment of the background levels of transition metals. The first row transition metal content of the resulting aqueous solutions was then analyzed by inductively coupled plasma-optical emission spectroscopy (ICP-OES) at the University of Minnesota Soil Research Analytical Laboratory.

UV/Vis spectroscopy – UV/Vis spectra of Co-*Nd*CCD samples were recorded at room temperature with a Lambda Bio spectrometer (Perkin Elmer) using micro UV-transparent cuvettes (Perkin Elmer).

HPLC and LC-MS analysis of apocarotenoid cleavage activity – Test substrates (apocarotenoids and retinal) at final concentrations of 50 μM were incubated in the presence and absence of 10 μg of *Nd*CCD in a 500 μL reaction buffer consisting of 10 mM HEPES-NaOH, pH 7 and 0.05 % v/v Triton X-100 for 15 min at 28 °C. Enzymatic tests on *Cb*CCD and CAO1 (8, 9) (used as a negative control) were conducted similarly except that crude supernatant was used as the enzyme source. The reactions were quenched by the addition of 100 μL of methanol and apocarotenoids were extracted by addition of 300 μL of brine and 500 μL of hexanes followed by vigorous shaking. For analysis of apocarotenoids that were copurified with Co-*Nd*CCD, a 20 μL sample of purified sample at a protein concentration of 70 mg/mL was directly added to 100 μL of methanol and then subjected to the same extraction procedure.

The hexane layer containing apocarotenoids was removed and either used directly for normal phase HPLC analyses or dried *in vacuo* and redissolved in methanol for reverse phase LC-MS analysis. Reversed phase LC-MS analysis was carried out on an ACQUITY UPLC system coupled with a QDa mass spectrometry detector (Waters). The samples were separated on a BEH C18/C4 column (2.1 x 50 mm, i.d. 1.7 µm, Waters) with a 5 to 95% acetonitrile in 0.1% formic acid gradient at a flow rate of 0.4 µL/min. Normal phase HPLC analysis was carried out on an Agilent 1100 series chromatography system equipped with a UV/Vis diode array detector and a Zorbax SIL (4.6 x 250 mm, 5 µm, Agilent) column operating with an isocratic liquid phase consisting of either 9:1 or 8:2 hexanes:ethyl acetate for non-hydroxylated and hydroxylated apocarotenoids, respectively.

HPLC analysis of carotenoid cleavage activity – The enzymatic activity of *Nd*CCD towards β-carotene, lycopene and zeaxanthin was carried out both as described above for apocarotenoids as well as with an alternative assay using pre-formed carotenoid-containing micelles and three different test detergents as described previously (10). The latter assay was performed with 4 µg of protein in a 200 µL reaction buffer containing 25 mM HEPES pH 7.4. Detergents (each tested separately) were dissolved in ethanol and mixed with 2 nmol of the respective carotenoid. The mixture was dried in a Speedvac. Carotenoid-detergent micelles were formed by addition of 200 µL of 25 mM HEPES pH 7.4 followed by vortexing for 20 s. Final concentrations of detergent were 0.3 % decyl maltose neopentyl glycol (w/v), 0.3 % N-dodecyl-β-D-maltoside (w/v), or 3% N-octyl-β-D-thioglucopyranoside (w/v). Following addition of *Nd*CCD, the assays were incubated for 12 min at 37 °C under constant shaking (600 rpm). After 12 min, the reaction was stopped by addition of 400 µL of acetone, 400 µL of diethylether, and 100 µL of petroleum ether. The mixture was vortexed for 10 s at highest speed. Then the mixtures were centrifuged for 10 s at 5000*g*. The organic phase which was collected and dried under vacuum at 30 °C in a Speedvac. The sample was dissolved in 150 µL of 9:1 hexanes/ethyl acetate and subjected to HPLC analysis on an 1100 Agilent HPLC series equipped with a diode array detector and normal-phase Zorbax silica column (5 µm, 150 mm; Agilent, Santa Clara, CA, USA). Chromatographic separation was achieved with isocratic flow of 10 - 30% ethyl acetate in hexane at a flow rate of 1.4 ml/min. The enzymatic activity of *Cb*CCD towards β-carotene was measured out as described above for apocarotenoids with CAO1 used as a negative control.

HPLC analysis of phenylpropanoid cleavage activity – The enzymatic activity of *Nd*CCD towards resveratrol and isoeugenol was analyzed as described previously (11) using CAO1 as a positive control. Briefly, test substrates at final concentrations of 50 µM were incubated in the presence and absence of 10 µg of *Nd*CCD or CAO1 in 500 µL of 10 mM HEPES-NaOH, pH 7 reaction buffer for 1 h at 28 °C with 300 RPM shaking. The reactants and products were extracted by addition of 100% ethyl acetate followed by vigorous shaking. Following centrifugation, the organic phase was collected and dried *in vacuo*. Compounds were redissolved in 3:2 hexane:ethyl acetate and separated on a Zorbax SIL (4.6 x 250 mm, 5 µm, Agilent) column with an isocratic liquid phase consisting of 3:2 hexanes:ethyl acetate running at 1.4 mL/min.

Steady-state kinetics – Purified Fe-*Nd*CCD at concentrations of 29.3 µM and 293 µM was prepared in 10 mM HEPES-NaOH, pH 7 for reactions with all-*trans*-8'-apocarotenal and all-*trans*-apo-8'-carotenol respectively. All-*trans*-apo-8'-apocarotenal was dissolved in ethyl alcohol and placed in amber tubes. All-*trans*-β-apo-8'-carotenol was generated from all-*trans*-β-apo-8'-carotenal by $NaBH_4$ reduction. Excess $NaBH_4$ was removed by addition of water. Apocarotenoids were extracted with hexane, dried *in vacuo*, and redissolved in ethanol. Reactions were performed in a total of 70 µL reaction buffer consisting of 20 mM Bis-Tris-HCl, pH 7.0 and 0.05% (w/v) Triton X-100. First, 1.5 µL of substrate was added to each reaction to attain final concentrations ranging from 169 µM to 2 µM for apocarotenal and from 40 µM to

1.6 μM for apocarotenol. Absorbance changes at 472 nm for all-*trans*-β-apo-8'-carotenal and at 445 nm for all-trans-β-apo-8'-carotenol were monitored over 2-5 min with a UV/Vis spectrophotometer immediately following addition of enzyme to the reaction mixture. Reaction rates were derived from least-squares fitting to the initial linear portions of the reaction curves. Absorbance changes were converted to concentration changes through measurement of absorbance standards that were read concurrently with the experimental samples. The initial velocity versus substrate concentration was plotted and fit with the Michaelis-Menton equation in GraphPad to determine the kinetic paramters, K_m and k_{cat}. All measurements were performed in triplicate.

X-ray absorption spectroscopy – Purified Fe-*Nd*CCD at a concentration of ~1.3 mM was prepared in a buffer consisting of 10 mM HEPES-NaOH, pH 7 containing 20% of xylitol. The sample was loaded into a custom-made Delrin cuvette containing a Kapton tape window and then flash cooled in liquid nitrogen. The samples were stored in liquid nitrogen vapor prior to X-ray absorption spectroscopy (XAS) measurements. XAS data were collected at SSRL beamline 7-3 in fluorescence mode using a Canberra 32-element solid state germanium detector. Data were measured in 10 eV steps in the pre-edge region (1 s acquisition time), 0.3 eV steps along the edge (2 s acquisition time), and 0.05 *k* steps in the extended X-ray absorption fine structure (EXAFS) region up to 15 *k* (acquisition time increasing from 2–9 s in a k^2-weighted fashion). Samples were monitored for evidence of radiation damage as indicated by pre-edge or edge shifts. No radiation damage was observed for the samples in this study. The inflection point of the iron foil reference was set to 7112.0 eV. XAS data were processed and averaged using EXAFSPAK (12). Pre-edge peak fitting analysis was carried out with Fityk (13) using a previously reported protocol (9) over the energy range of 7704–7718 eV. EXAFS analysis was performed with Artemis (14). Theoretical phase and amplitude parameters were calculated for a model of the Fe-*Nd*CCD active site using FEFF6L. From these parameters, relevant paths were incorporated into the fitting model and evaluated for the significance of their contributions. For a given shell in all simulations, the coordination number *n* was fixed, while *r* and σ^2 were allowed to float. The amplitude reduction factor S_0^2 was fixed at 0.9, while the edge shift parameter ΔE_o was allowed to float at a single common value for all shells. Histidine multiple scattering was evaluated by representing the imidazole moiety as four sets of grouped paths corresponding to the dominant scattering contributions. The fit was evaluated in k^3-weighted R-space, and fit quality was judged by the *R*-factor and reduced χ^2.

NdCCD crystallization – *Nd*CCD crystallization was performed by the hanging-drop, vapor-diffusion method by mixing 2 μL of purified *Nd*CCD at a concentration of 20 mg/mL with 2 μL of reservoir solutions. Initial showers of small needle-like crystals were observed in condition #15 of the MidasPlus crystallization screen (Molecular Dimensions Inc., Maumee, OH) after 3 days of incubation at 22 °C. These initial conditions were systematically optimized through rational grid screening and alteration of other growth conditions. Optimized crystals grew in the sitting drop format at 22 °C in a crystallant consisting of 0.1 M MES-NaOH, pH 6, 0.1 M NaCl, 22% Sokalan PA 25 CL and 22% xylitol (Fe-*Nd*CCD, 60 mg/mL) or 0.1 M MES-NaOH, pH 6, 0.1 M NaCl, 22% v/v Sokalan PA 25 CL, 5% v/v glycerol and 10-13% xylitol (Co-*Nd*CCD, 52-57 mg/mL). Long, rod-shaped crystals appeared within 2 days and were harvested 1-14 days later depending on the particular experiment. Crystals dimensions at the time of harvest were typically ~ 200 × 200 × 500-1000 μm. Crystals were harvested using dual-thickness MiTeGen loops (Ithaca, NY). Xylitol and glycerol included during crystallization provided sufficient cryoprotection to allow direct freezing in liquid nitrogen. Crystals were stored in liquid nitrogen vapor prior to X-ray data collection.

Diffraction data collection, structure determination, and refinement – Diffraction data on *Nd*CCD crystals were collected at the APS 24-ID-E, NSLS-II 17-ID-1 and 17-ID-2, and SSRL 12-2 beamlines.

Care was taken to orient the long axis of the crystals nearly parallel to the axis of rotation to avoid reflection overlaps. Data were processed with XDS (15) and the reduced data analyzed using phenix.xtriage (16). The Fe-*Nd*CCD crystals diffracted X-rays up to ~2.7 Å resolution with symmetry and systematic absences consistent with space group $P6_{(1,5)}$. Molecular replacement was carried out in MrBUMP (17), using Phaser (18) and Sculptor (19), through the CCP4 online server. The structure was solved by molecular replacement in space group $P6_1$ with 6 molecules in the asymmetric unit using the coordinates of RPE65 (PDB accession code 3KVC) as a search model (20). The initial molecular replacement phases were improved through NCS averaging and histogram matching in Parrot (21). The model was then iteratively improved through density modification, manual and automatic rebuilding in Coot and ARP/wARP respectively, together with reciprocal space refinement in Refmac. Crystals of Co-*Nd*CCD and the Co-*Nd*CCD/apocarotenoid complex diffracted X-rays up to ~2.3 Å resolution and were isomorphous to Fe-*Nd*CCD crystals. The Co-*Nd*CCD structures were solved by direct refinement using the Fe-*Nd*CCD coordinates as an initial model. Manual model adjustments were made using Coot and the structures were refined in Refmac. The Grade Web Server (Global Phasing Ltd) was used to generate coordinate and dictionary files for apocarotenoid ligands. Model stereochemistry was routinely checked after cycles of refinement using the Molprobity (22) server and the quality of the model analyzed using the wwPDB validation server (23). In all of the structures, chains E and F had weaker electron density compared with the other four chains, and except where noted the structural description in the main text is based on an analysis of chains A through D.

Structural analysis and visualization – Structural analyses were carried out using the PISA (24), MOLEonline (25) and PDBsum (26) webservers. Ray-traced structural images were prepared using PyMOL (Schrödinger, LLC).

References

1. Boratyn GM*, et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* 41(Web Server issue):W29-33.
2. El-Gebali S*, et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.* 47(D1):D427-D432.
3. Sievers F*, et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539.
4. Huelsenbeck JP & Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754-755.
5. Letunic I & Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44(W1):W242-245.
6. Podell S & Gaasterland T (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol.* 8(2).
7. Buchfink B, Xie C, & Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12(1):59-60.
8. Diaz-Sanchez V, Estrada AF, Limon MC, Al-Babili S, & Avalos J (2013) The oxygenase CAO-1 of Neurospora crassa is a resveratrol cleavage enzyme. *Eukaryot. Cell* 12(9):1305-1314.
9. Sui X*, et al.* (2017) Structure and Spectroscopy of Alkene-Cleaving Dioxygenases Containing an Atypically Coordinated Non-Heme Iron Center. *Biochemistry* 56(22):2836-2852.
10. Babino D*, et al.* (2015) Characterization of the Role of beta-Carotene 9,10-Dioxygenase in Macular Pigment Metabolism. *J. Biol. Chem.* 290(41):24844-24857.
11. Sui X*, et al.* (2015) Utilization of Dioxygen by Carotenoid Cleavage Oxygenases. *J. Biol. Chem.* 290(51):30212-30223.

12. George GN (2000) EXAFSPAK (Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Stanford, CA).
13. Wojdyr M (2010) Fityk: a general-purpose peak fitting program. *J. Appl. Crystallogr.* 43:1126-1128.
14. Ravel B & Newville M (2005) ATHENA, ARTEMIS, HEPHAESTUS: data analysis for X-ray absorption spectroscopy using IFEFFIT. *J. Synchrotron Radiat.* 12:537-541.
15. Kabsch W (2010) XDS. *Acta Crystallogr. D* 66(Pt 2):125-132.
16. Zwart PH, Grosse-Kunstleve RW, Lebedev AA, Murshudov GN, & Adams PD (2008) Surprises and pitfalls arising from (pseudo)symmetry. *Acta Crystallogr. D* 64(Pt 1):99-107.
17. Keegan RM & Winn MD (2008) MrBUMP: an automated pipeline for molecular replacement. *Acta Crystallogr. D* 64(Pt 1):119-124.
18. McCoy AJ*, et al.* (2007) Phaser crystallographic software. *J. Appl. Crystallogr.* 40(Pt 4):658-674.
19. Bunkoczi G & Read RJ (2011) Improvement of molecular-replacement models with Sculptor. *Acta Crystallogr. D* 67(Pt 4):303-312.
20. Golczak M, Kiser PD, Lodowski DT, Maeda A, & Palczewski K (2010) Importance of membrane structural integrity for RPE65 retinoid isomerization activity. *J. Biol. Chem.* 285(13):9667-9682.
21. Cowtan K (2010) Recent developments in classical density modification. *Acta Crystallogr. D* 66(Pt 4):470-478.
22. Williams CJ*, et al.* (2018) MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* 27(1):293-315.
23. Read RJ*, et al.* (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure* 19(10):1395-1412.
24. Krissinel E & Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* 372(3):774-797.
25. Pravda L*, et al.* (2018) MOLEonline: a web-based tool for analyzing channels, tunnels and pores (2018 update). *Nucleic Acids Res.* 46(W1):W368-W373.
26. Laskowski RA, Jablonska J, Pravda L, Varekova RS, & Thornton JM (2018) PDBsum: Structural summaries of PDB entries. *Protein Sci.* 27(1):129-134.
27. Eme L, Spang A, Lombard J, Stairs CW, & Ettema TJG (2017) Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* 15(12):711-723.
28. McGinnis S & Madden TL (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32(Web Server issue):W20-25.
29. Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28(1):27-30.
30. Sievers F & Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* 27(1):135-145.
31. Gouet P, Robert X, & Courcelle E (2003) ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* 31(13):3320-3323.
32. Evans P (2006) Scaling and assessment of data quality. *Acta Crystallogr. D* 62(Pt 1):72-82.

Supporting Figures

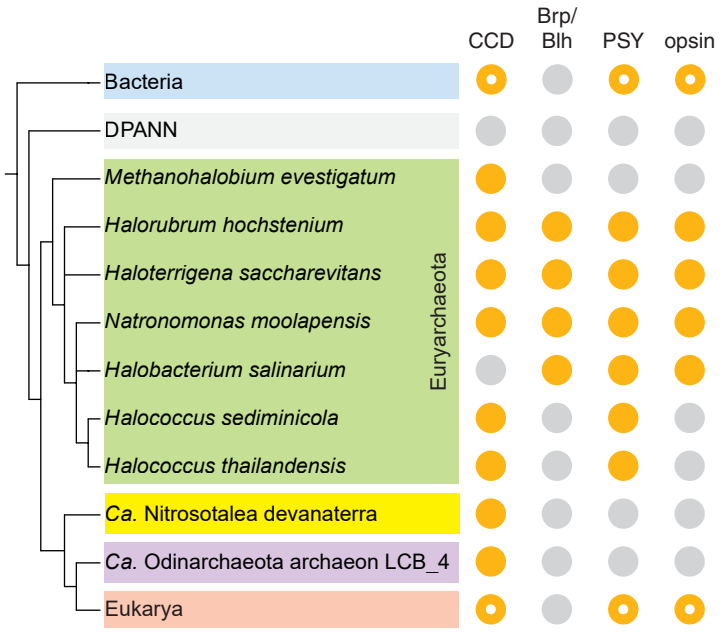

Figure S1 – *Phylogenetic relationships between bacteria, eukarya and select archaea together with the occurrence of CCDs and other genes related to carotenoid metabolism and usage in such organisms.* The tree was constructed according to (27) and the NCBI taxonomy database. Indicated genes were identified using the NCBI BlastP server (28) and the KEGG database (29). Gold and light-grey circles signify the presence or absence of genes encoding the proteins given in the column heads. Hollow gold circles indicate the presence of the indicated gene in some but not all bacteria and eukarya. Brp – bacteriorhodopsin-related protein, Blh – Brp-like homolog, PSY – phytoene synthase.

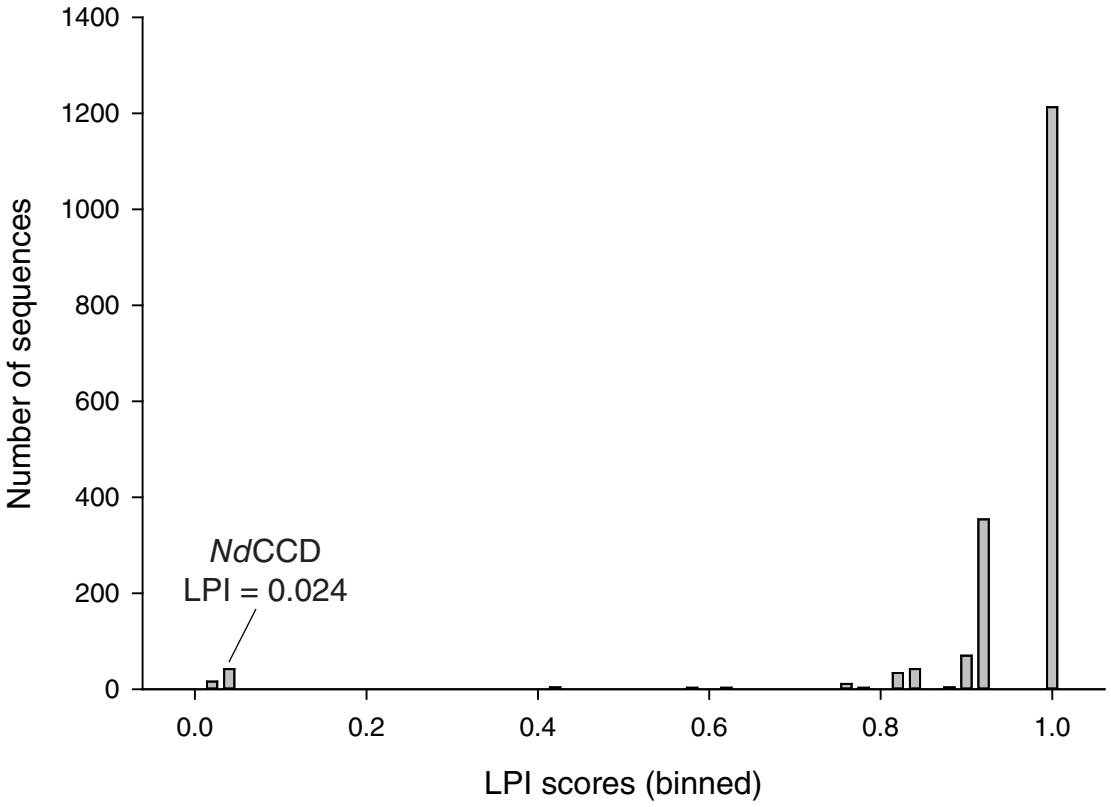

Figure S2 – *Lineage probability index (LPI) score histogram for proteins encoded in the N. devanaterra genome*. Lower LPI scores are indicative of genes most likely to have been acquired by horizontal gene transfer. *Nd*CCD was assigned a score of 0.024 placing it among the most atypical proteins in the *N. devanaterra* genome, consistent with it having been acquired by horizontal gene transfer.

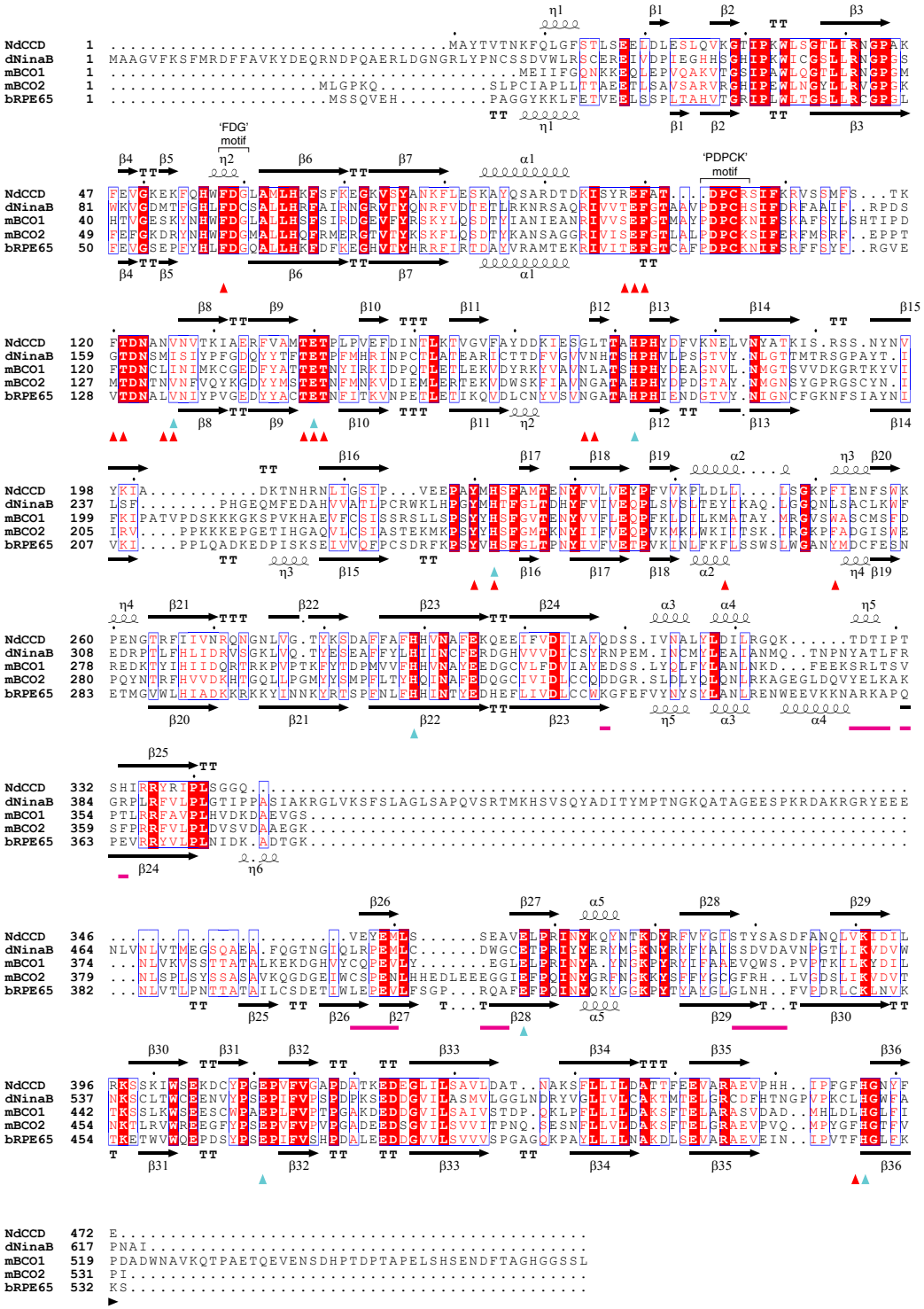

Figure S3 – *Sequence alignment of NdCCD and select metazoan CCDs*. Secondary structure elements for *Nd*CCD and RPE65 are respectively shown on top and bottom of the alignment. Conserved first and second sphere residues are indicated by cyan arrows. Residues observed to interact with β-apo-14'-carotenal bound to the *Nd*CCD active site are marked with red arrows. Residues involved in *Nd*CCD dimer formation, as observed crystallographically, are indicated by magenta lines. The alignment was generated using Clustal Omega (30) and visualized using Espript (31). dNinaB – *Drosophila melanogaster* Neither inactivation nor afterpotential B, mBCO1/2 – *Mus musculus* beta carotene oxygenases 1 and 2, bRPE65 – *Bos taurus* RPE65

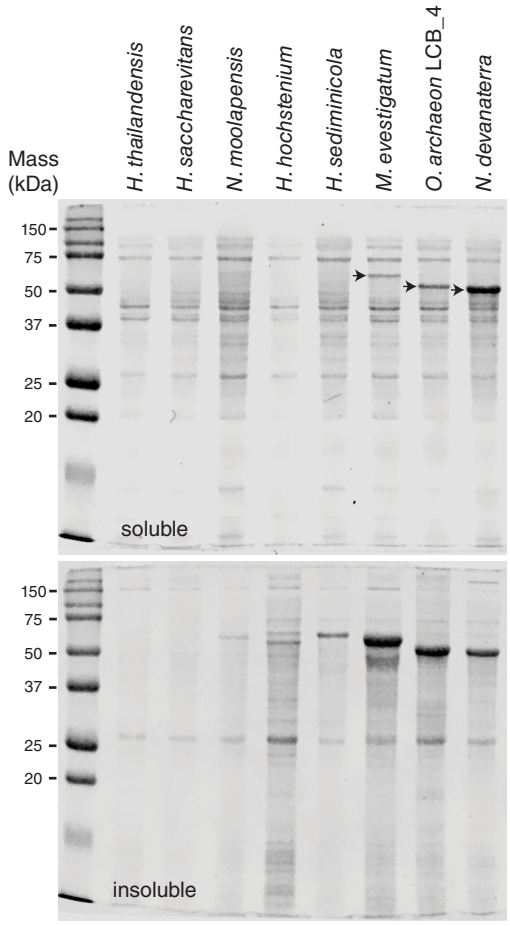

Figure S4 – *Expression of archaeal CCDs in E. coli.* Soluble (top) and insoluble (bottom) proteins following lysis of *E. coli* cells harboring plasmids encoding CCDs from the indicated archaea were separated on 12% SDS-PAGE gels and stained with Coomassie brilliant blue dye. Soluble protein expression was observed for CCDs from *M. evestigatum*, *O. archaeon LCB_4,* and *N. devanaterra* (indicated by black arrows) with the latter protein expressed at the highest level.

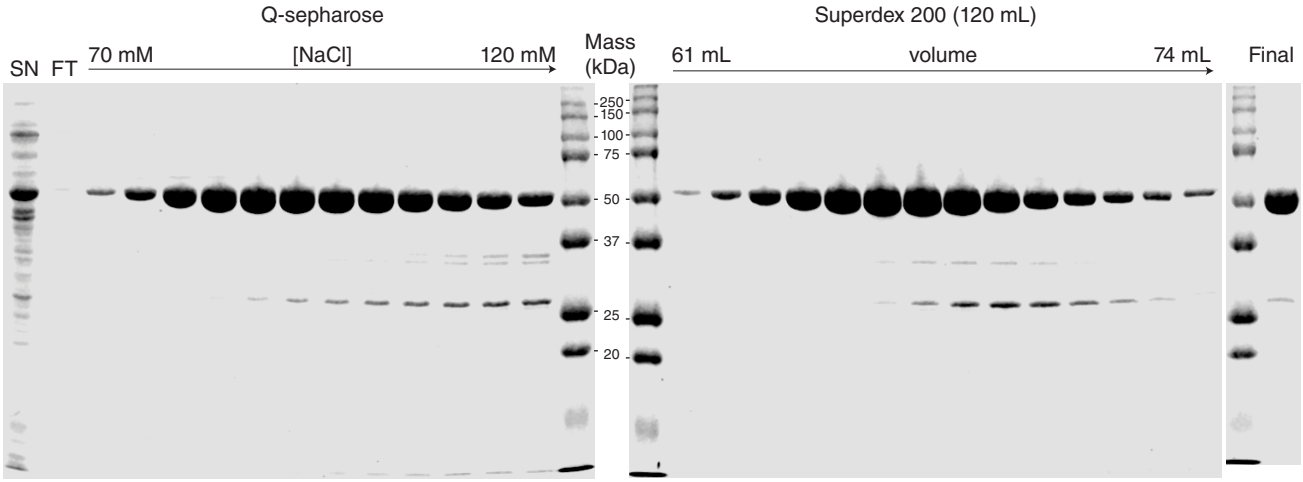

Figure S5 – *Purification of NdCCD.* Untagged *Nd*CCD was expressed in *E. coli* and purified by anion exchange chromatography (Q-sepharose, left-most gel) and gel filtration chromatography (Superdex 200, middle gel). The final preparation (right-most gel) was nearly homogeneous. SN – supernatant, FT – flow-through.

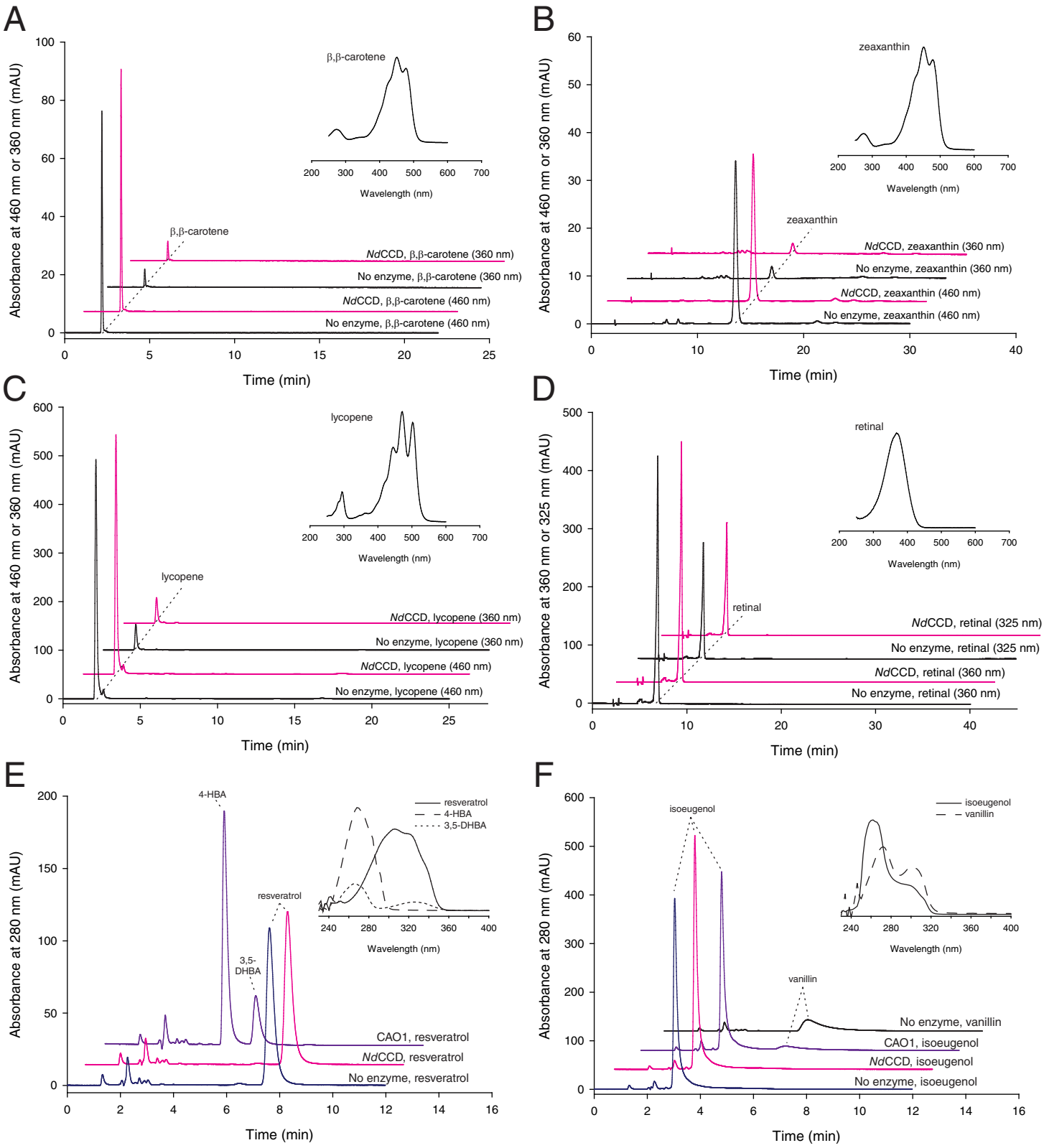

Figure S6 – *Cleavage activity of NdCCD towards C40 carotenoids, retinal, and phenylpropanoids*. NdCCD had no detectable cleavage activity towards **A**) β,β-carotene, **B**) zeaxanthin, **C**) lycopene, **D**) retinal, **E**) resveratrol, or **F**) isoeugenol. In panels E and F, the stilbene-cleaving CCD, CAO1 (8, 9), was used as a positive control for cleavage activity. 4-HBA – 4-hydroxybenzaldehyde, 3,5-DHBA – 3,5-dihydroxybenzaldehyde.

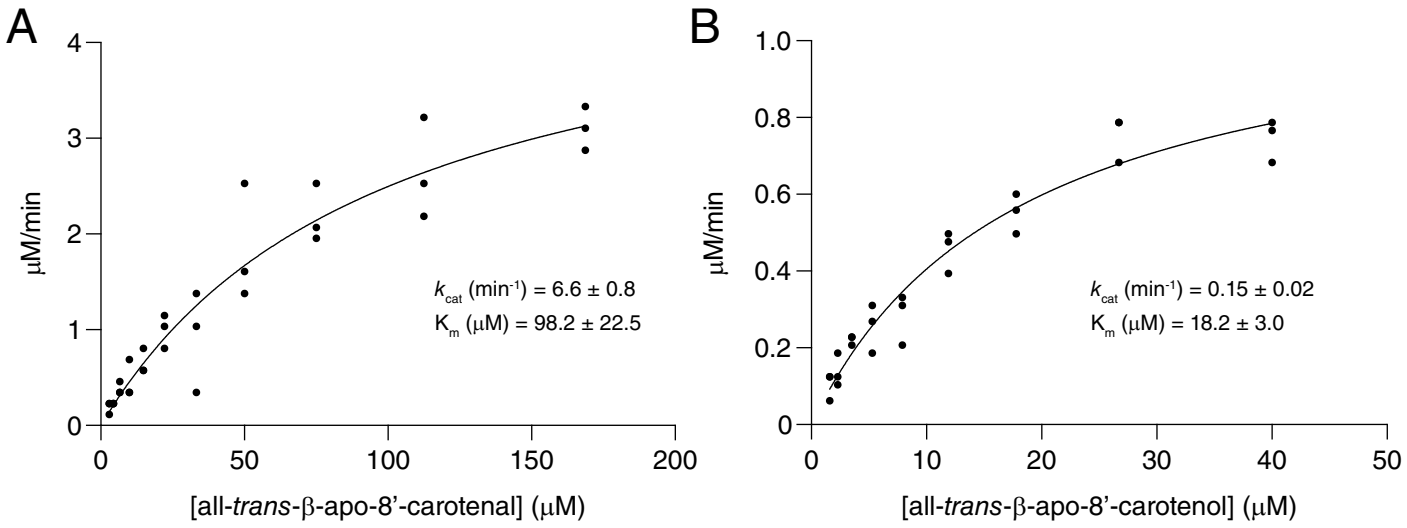

Figure S7 – *Michaelis-Menton kinetics of apocarotenoid cleavage by NdCCD*. **A**) Steady-state cleavage of all-*trans*-β-apo-8'-carotenal. Initial velocities were obtained by measuring the decrease in substrate concentration at 472 nm over time. **B**) Steady-state cleavage of all-*trans*-β-apo-8'-carotenol. Initial velocities were obtained by measuring the decrease in substrate concentration at 445 nm over time. The catalytic efficiency (k_{cat}/K_m) was ~ 8 times higher for apocarotenal versus apocarotenol (0.067 ± 0.017 $min^{-1} \cdot \mu M^{-1}$ versus 0.008 ± 0.002 $min^{-1} \cdot \mu M^{-1}$).

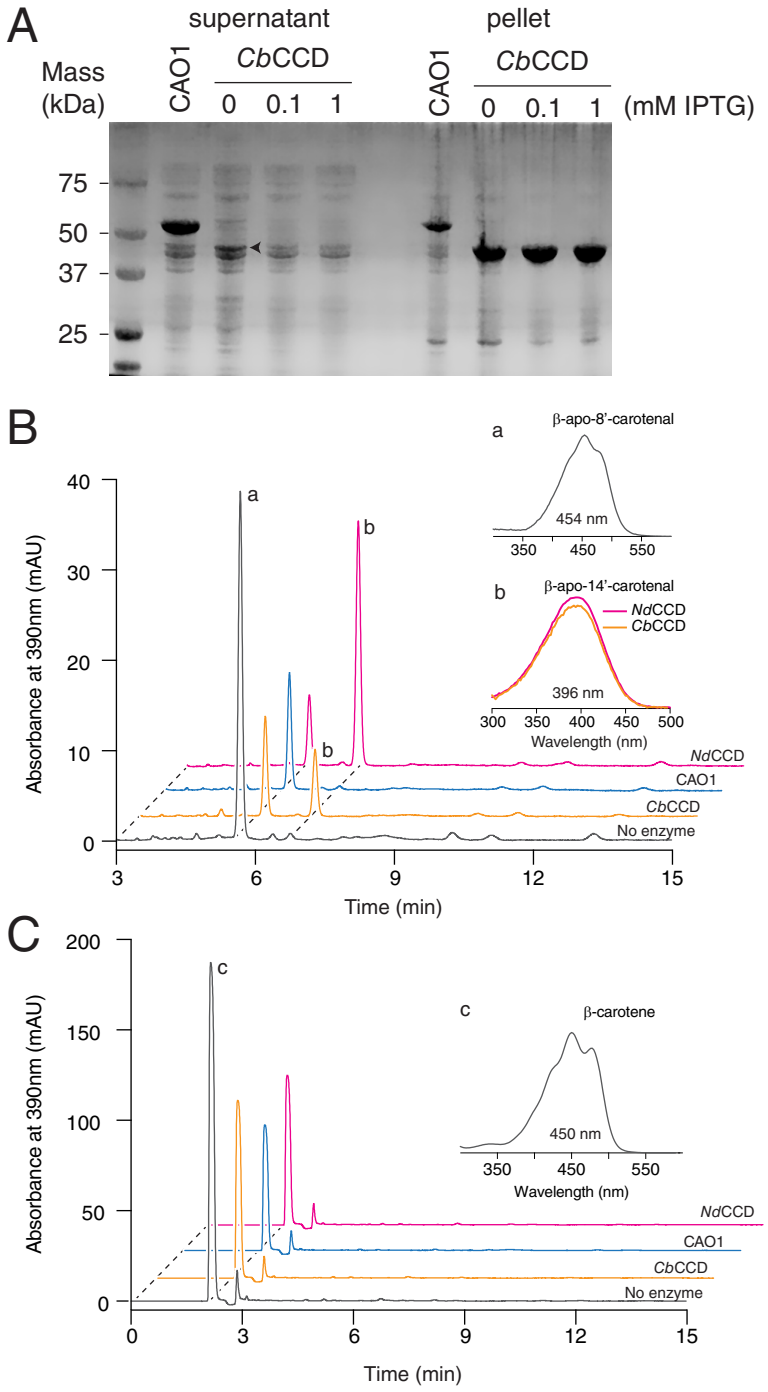

Figure S8 – *Expression of Coxiella burnetii in E. coli and its activity towards β-carotene and β-apo-8'-carotenal.* **A**) Supernatant and pellet fractions from *E. coli* cultures harboring the *Coxiella burnetii* CCD (*Cb*CCD)-pET3a expression plasmid induced with the indicated concentrations of IPTG. Soluble protein was detected only in the uninduced culture (black arrow), while the pellet fractions contained large amounts of insoluble *Cb*CCD. CAO1 was also expressed in *E. coli* as previously described (9) for use as a negative control for carotenoid cleavage activity (8). **B**) *Cb*CCD cleaved β-apo-8'-carotenal (a) to generate β-apo-14'-carotenal as demonstrated by the identical retention time and spectral shape of the product peak (b) as compared to the authentic standard generated by *Nd*CCD. As expected, no cleavage products were formed from CAO1-containing supernatant. **C**) Like *Nd*CCD, *Cb*CCD was unable to cleave β-carotene under standard assay conditions.

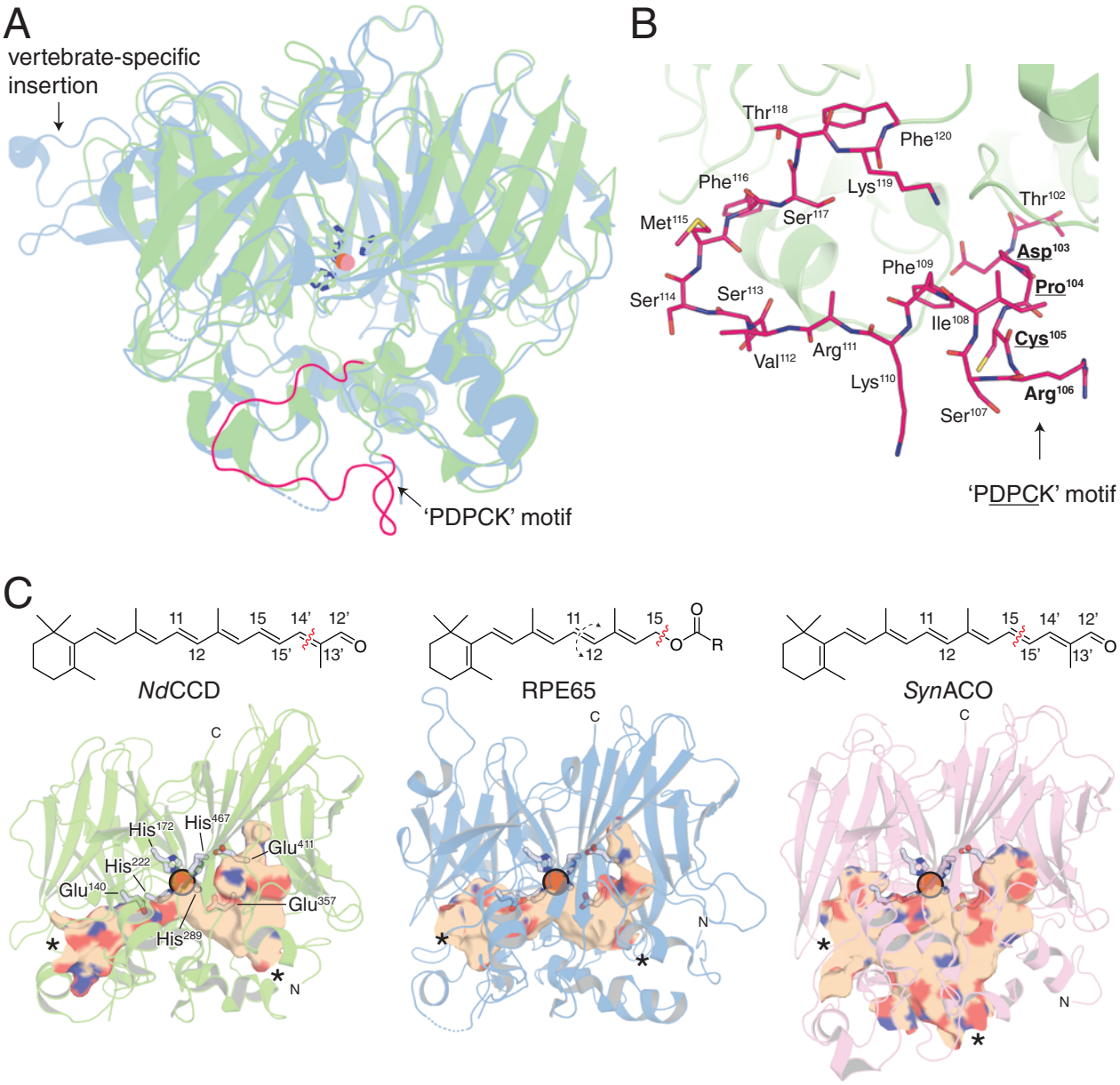

Figure S9 – *Structural comparison of NdCCD to metazoan and bacterial CCDs.* **A**) Superposition of *Nd*CCD (green) with RPE65 (blue, PDB accession code 4RSC) showing their close structural similarity. A vertebrate-specific insertion that mediates dimer formation in RPE65 is labeled on the left side of the panel. The loop containing the 'PDPCK' motif on its N-terminal side in *Nd*CCD is show in pink. The corresponding region of RPE65 is disordered in all structures determined to date. **B**) The 'PDPCK'-containing loop structure is mostly irregular with only three β-turn elements present. Residues in *Nd*CCD exactly matching the motif are bolded and underlined, while the conservative Arg variant is shown in bold. Note the overall hydrophobic and positively charged character for this loop region in *Nd*CCD. The orientation of the loop is the same in panels A and B. **C**) An established substrate for each enzyme is shown on top with the site of cleavage (oxidative or hydrolytic) marked by a wavy red line. The bottom panels show the overall structure of the active site tunnels. The active site of *Nd*CCD (left) resembles that of RPE65 (middle), an enzyme found in vertebrates that cleaves retinyl esters with simultaneous isomerization of the retinoid C11-C12 double bond. By contrast, the active site of *Syn*ACO has a substantially different geometry despite the fact that this enzyme has a substrate specificity similar to that of *Nd*CCD. The asterisks indicate positions where the active site tunnels open to the protein exterior. N and C represent the amino and carboxy termini of each structure.

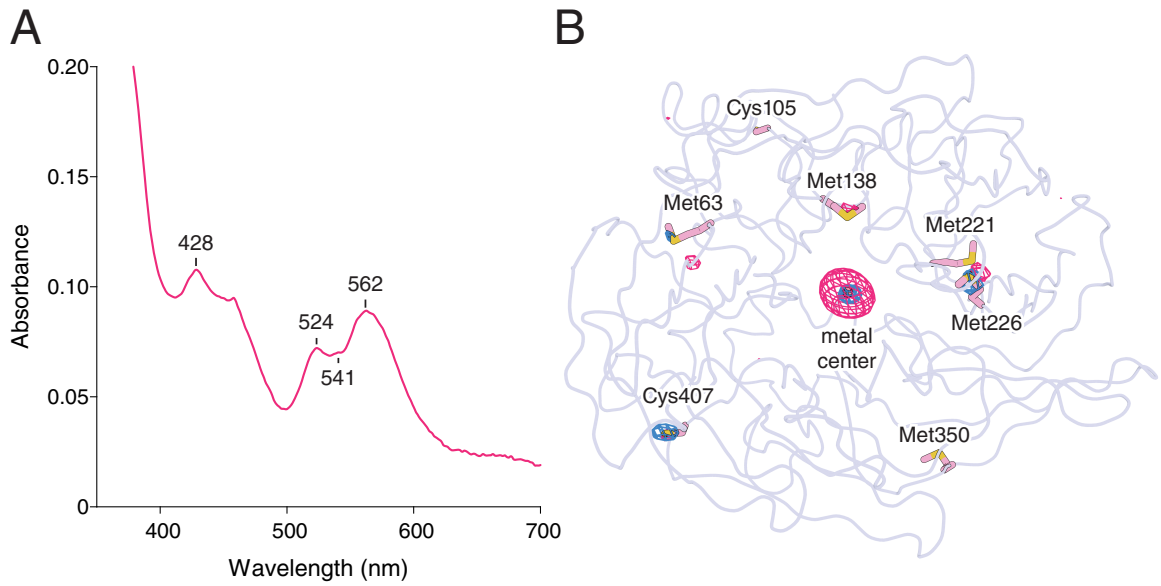

Figure S10 – *Evidence for Co-substitution in purified NdCCD expressed in cobalt-containing minimal media.* **A**) UV/Vis spectrum of Co-substituted *Nd*CCD showing characteristic *d-d* Co^{II} absorbance bands at 500-600 nm. **B**) NCS-averaged imaginary log-likelihood gradient maps computed using amplitudes from Co-*Nd*CCD data sets collected above (magenta mesh) and below (blue mesh) the cobalt K absorption edge. The large difference in scattering at the metal binding site with comparable scattering at sulfur-containing "internal control" residues indicates a preponderance of cobalt bound at the catalytic center.

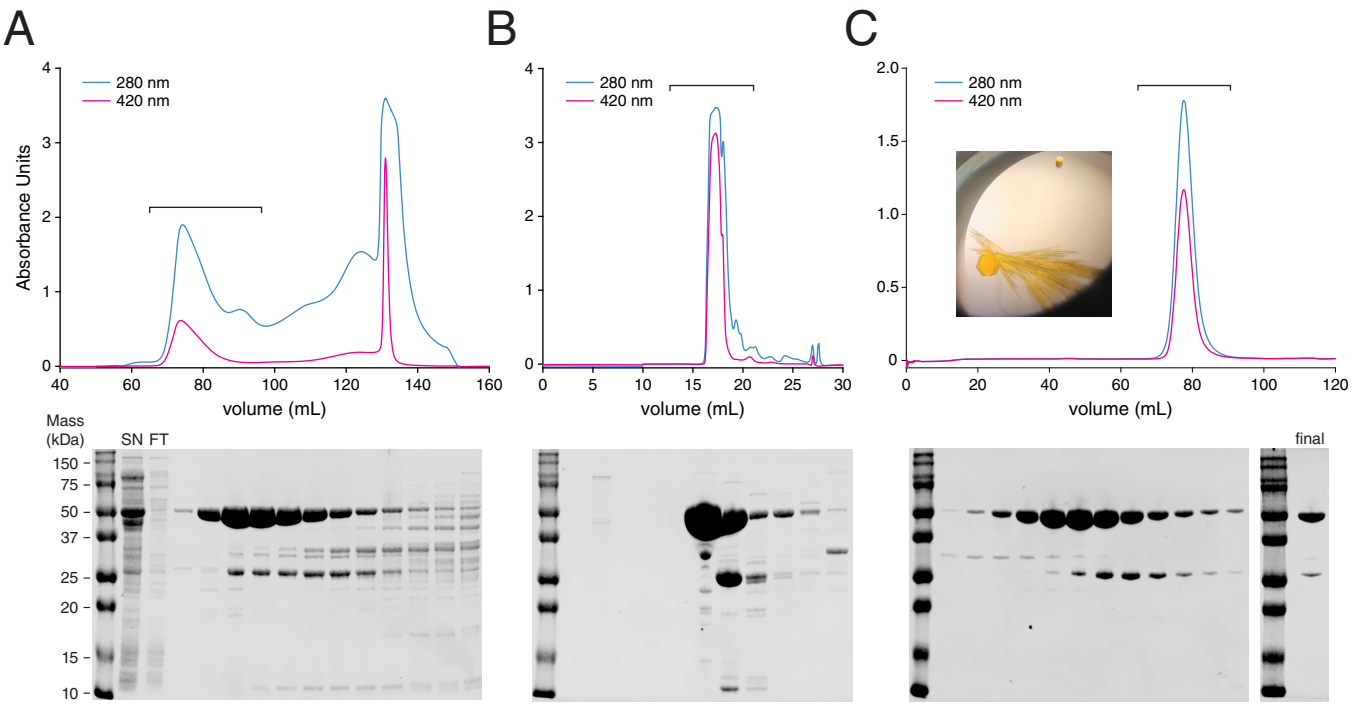

Figure S11 – *Co-purification of 3-hydroxy-β-apo-12'-carotenal with Co-NdCCD.* **A**) Chromatogram and gel from the initial Q-sepharose purification step. **B**) Chromatogram and gel from the subsequent Mono-Q purification step. **C**) Chromatogram and gel from the final gel filtration purification step. The horizontal brackets shown in each chromatogram indicate the material that was analyzed by SDS-PAGE. In all cases, apocarotenoid was observed to copurify with Co-*Nd*CCD as evidenced by the coincident 280 nm absorbance corresponding to *Nd*CCD and the 420 nm absorbance corresponding to the apocarotenoid. The ratio of peak areas in the final gel filtration step indicated a 3-hydroxy-β-apo-12'-carotenal (ε ~75,000 M^{-1} cm^{-1}) to *Nd*CCD (ε ~ 76,000 M^{-1} cm^{-1}) ratio of approximately 0.7. The inset in the right-most chromatogram shows crystals obtained from the final preparation, which exhibited a bright orange color owing to the co-purified apocarotenoid.

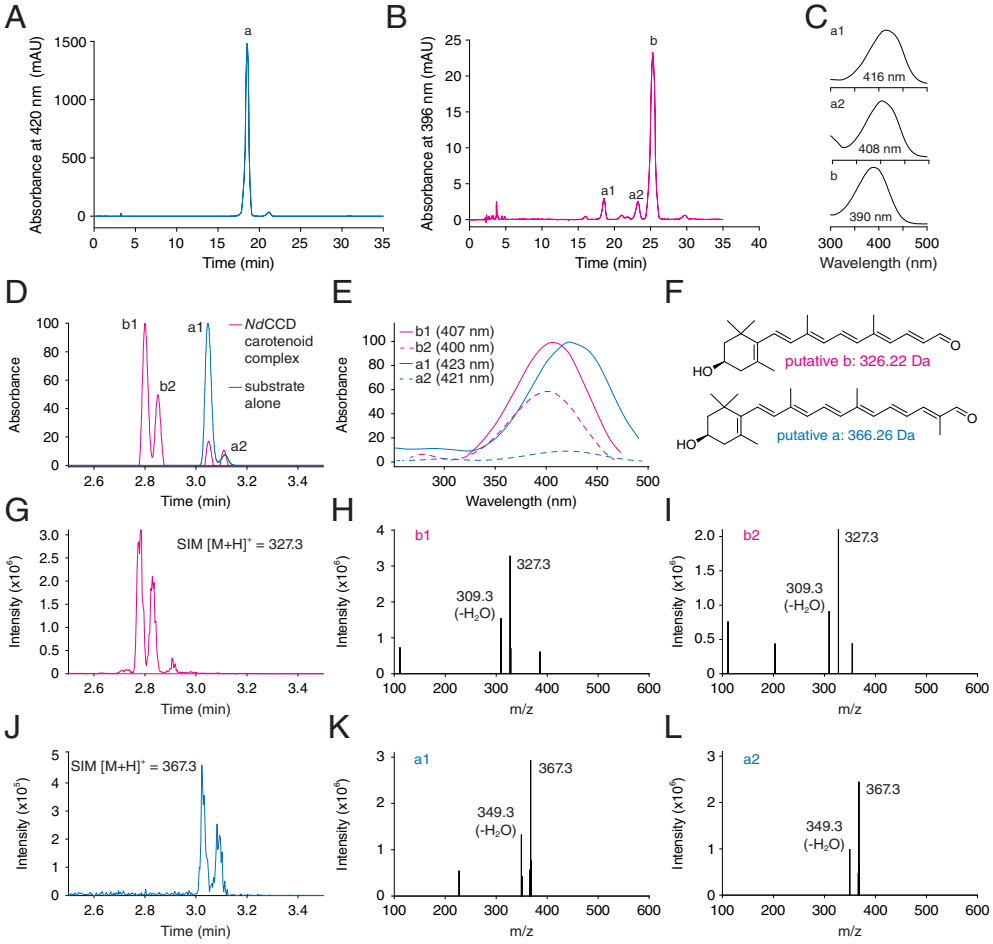

Figure S12 – *HPLC and LC-MS analyses demonstrate a predominance of 3-hydroxy-β-apo-14'-carotenal in Co-NdCCD purified in the presence of 3-hydroxy-β-apo-12'-carotenal.* **A**) Normal phase HPLC analysis of 3-hydroxy-β-apo-12'-carotenal alone. **B**) Normal phase HPLC analysis of apocarotenoids extracted from the *Nd*CCD-apocarotenoid complex obtained by gel filtration chromography. Note the dominance of compound "b" with an absorption spectrum shown in (**C**) consistent with 3-hydroxy-β-apo-14'-carotenal. Compound "a1" represent a minority amount of uncleaved substrate while compound "a2" likely represents isomerized substrate generated during exposure to visible light. The samples were separated with an 80%/20% hexane/ethyl mobile phase. **D)** Reversed-phase HPLC separation of 3-hydroxy-β-apo-12'-carotenal substrate (blue trace) and apocarotenoids extracted from the purified Co-*Nd*CCD-apocarotenoid complex (magenta trace). The apocarotenoids were eluted with a 5 to 95% acetonitrile gradient in the presence of 0.1% formic acid. **E**) Optical absorbance spectra of the peaks shown in panel D. **F**) Putative identifies of compounds 1 and 2 based on their optical absorbance spectra. The small difference in absorbance maxima between "a1" and "a2" as well as "b1" and "b2" indicate that they represent geometric isomers generated by photo- or thermal-isomerization during sample processing. **G**) Selective ion monitoring (SIM) at m/z = 327.3, corresponding to the molecular ion of 3-hydroxy-β-apo-14'-carotenal, showing peaks with retention times identical to that of peak b1 and b2 in panel D. **H)** and **I)** Mass spectra of the SIM peaks in panel G showing a dominance of the m/z = 323.3 species in both cases, confirming that these compounds are geometric isomers. The peaks at 309.3 represent compounds that have undergone fragmentation with loss of water. **J**) Selective ion monitoring (SIM) at m/z = 367.3, corresponding to the molecular ion of 3-hydroxy-β-apo-12'-carotenal, showing peaks with retention times identical to that of peak a1 and a2 in panel D. **K)** and **L)** Mass spectra of the SIM peaks in panel J showing a dominance of the m/z = 367.3 species in both cases, confirming that these compounds are geometric isomers. The peaks at 349.3 represent compounds that have undergone fragmentation with loss of water.

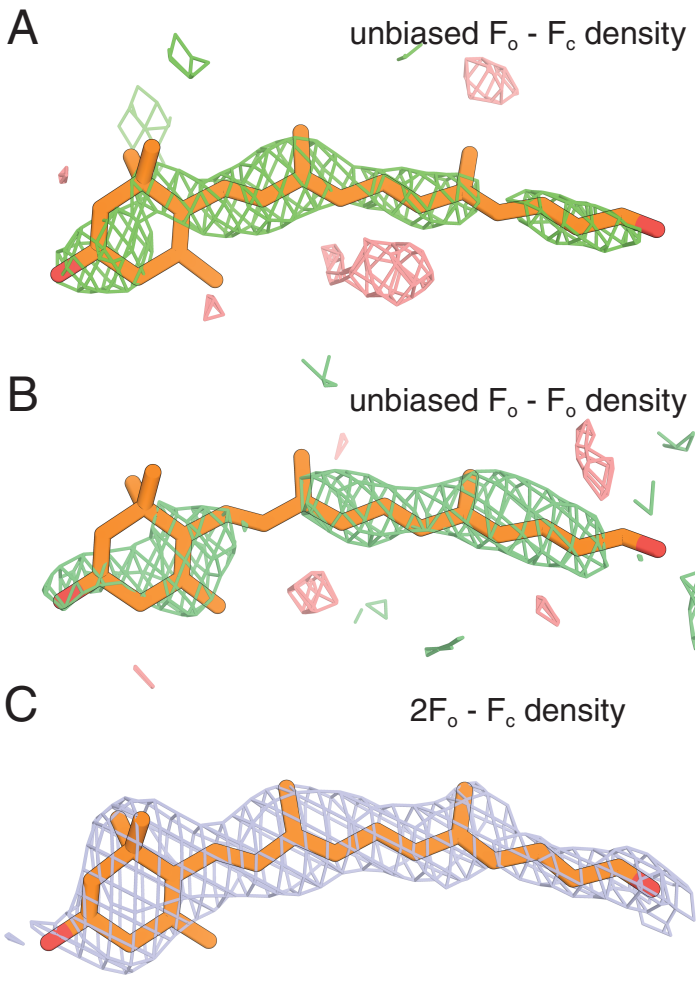

Figure S13 – *Electron density support for the bound apocarotenoid ligand.* **A**) NCS-averaged sigma A-weighted $|F_o|$ - $|F_c|$ electron density map contoured at +5 (green mesh) and -5 (salmon mesh) RMSD. Density is shown within 4 Å of the apocarotenoid. **B**) An NCS-averaged isomorphous difference map computed with $|F_o(\text{ligand})|$ – $|F_o(\text{apo})|$ as amplitudes and $\varphi_{calc}(\text{apo})$ as phases. The map is contoured at +5 (green mesh) and -5 (salmon mesh) RMSD. The R_{iso} between the data sets was 20% overall and 97% in the highest resolution shell. Density is shown within 4 Å of the apocarotenoid. **C**) Sigma A-weighted $2|F_o|$ - $|F_c|$ electron density map (light blue mesh) contoured at 0.8 RMSD and shown within 2 Å of the apocarotenoid. For panels A and C, NCS averaging was performed using only chains A-D as chains E and F were overall more poorly resolved.

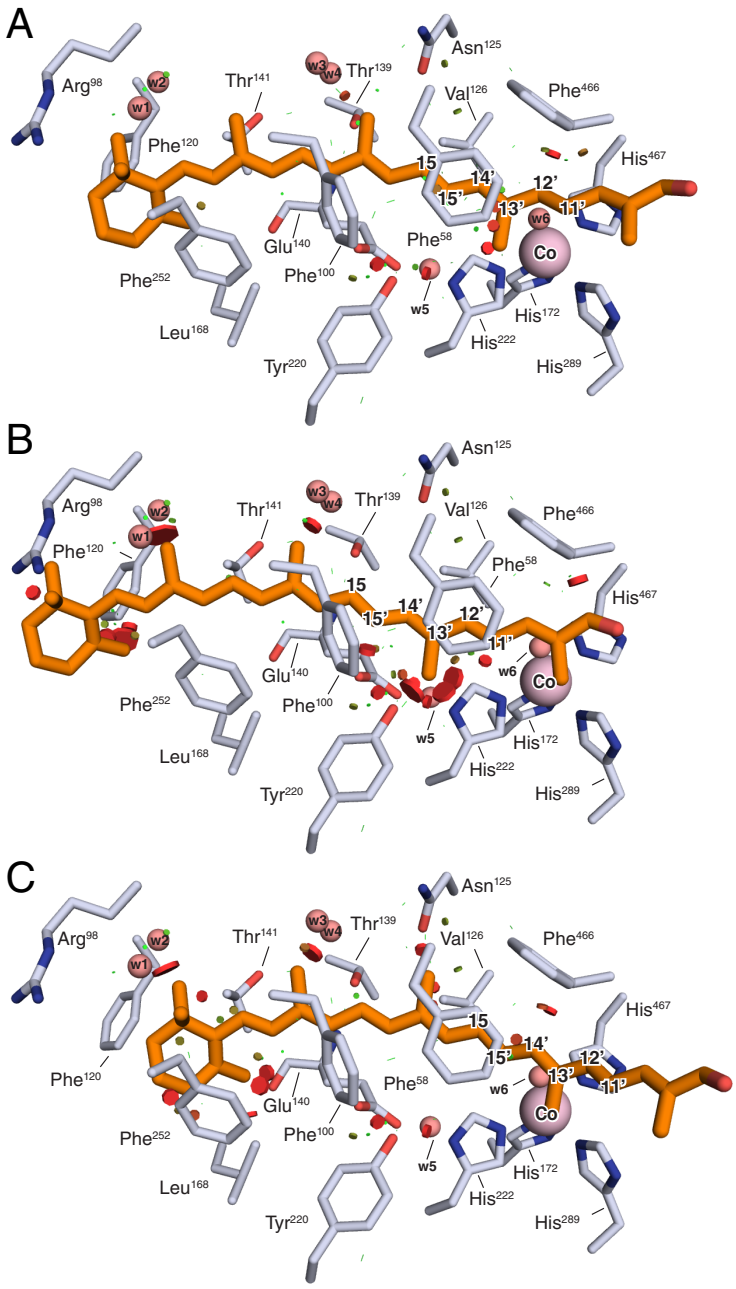

Figure S14 – *Active site determinants of apocarotenoid binding orientation in NdCCD.* **A**) Model of β-apo-8'-carotenal in the active site of NdCCD. The compound was manually docked into the active site as described in the legend to Fig. 7. This docking orientation properly positions the C14'-C13' alkene bond for cleavage. Steric clashes are represented by red disks with the severity of the clash proportional to the size of the disk. This mode of docking shows a minimal number of clashes consistent with an energetically preferred mode of binding. **B**) The apocarotenoid was shifted by one double bond outward to position the C12'-C11' bond for cleavage with slight adjustments made to the chi angles wherever possible to eliminate close interactions. However, this mode of binding results in severe steric clashes with Phe120, Tyr220, His222, as well as active site-bound water molecules indicating it is not energetically favorable. **C**) The apocarotenoid was shifted by one double bond inward (with respect to panel A) to position the C15-C15' bond for cleavage again with appropriate adjustments made to minimize close contacts. This mode of binding also resulted in severe steric clashes with active site residues including Phe252, Glu140, Thr141, and Val126 as well as active site bound water molecules, indicating that it is not energetically preferred. These results help explain the observed C14'-C13' regioselectivity of *Nd*CCD.

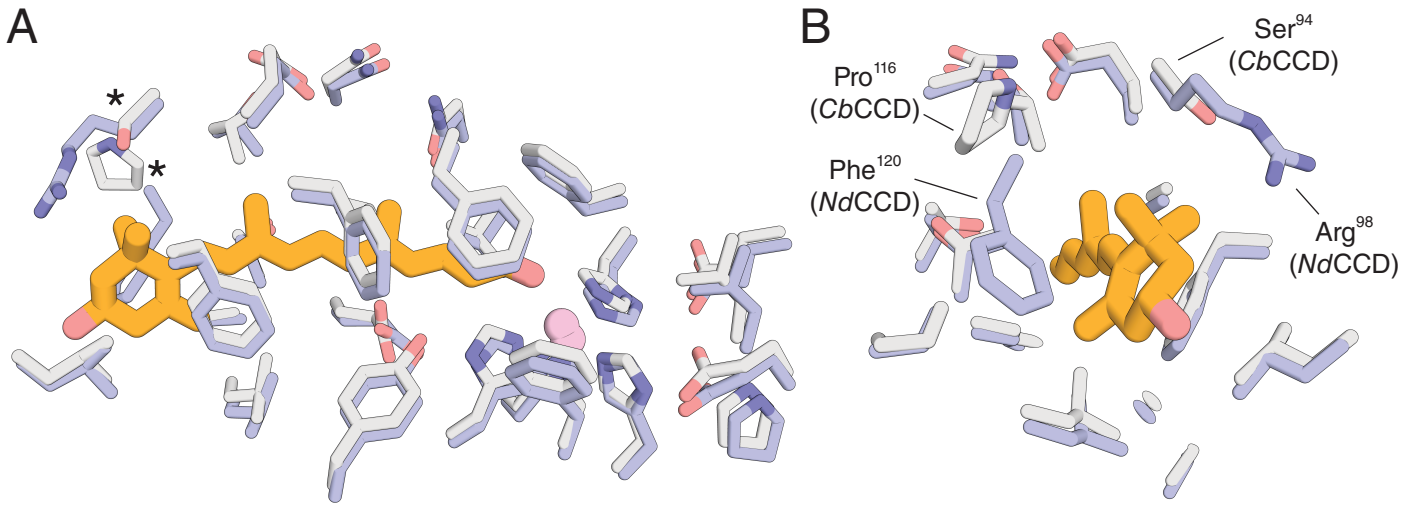

Figure S15 – *Comparison of the active site structure of NdCCD to the active site structure of CbCCD as predicted by homology modeling.* **A**) The central active sites of the two proteins are identical whereas significant differences occur at the substrate entry portal (marked by asterisks). **B**) A view down the substrate entry portal showing the substitution of Phe120 and Arg98 residues in *Nd*CCD with much smaller Pro and Ser residues in *Cb*CCD. These changes result in a significantly more open entry portal for *Cb*CCD as compared to *Nd*CCD. Given the identical cleavage regioselectivity of these two enzymes, it is inferred that a potential bottleneck at the substrate entry portal of *Nd*CCD is not sufficient to determine its cleavage regioselectivity. Instead, the central active site appears to play a dominant role in controlling substrate positioning.

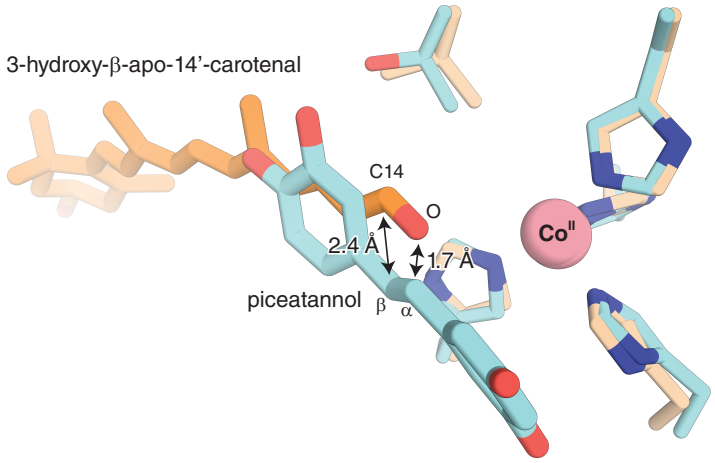

Figure S16 – *Comparison of the scissile bond orientation between carotenoid- and stilbenoid-cleaving CCDs.* The α,β bond of piceatannol (cyan sticks) bound to Co-CAO1 (PDB accession code 5U97) is shifted by ~ 2 Å and oriented differently as compared to the C14'-O bond of 3-hydroxy-β-apo-14'-carotenal (orange sticks), which represents the location of the apocarotenoid scissile bond. This difference could impact how O_2 is bound and activated between the two enzymes.

Supporting Tables

Table S1 – *CCDs investigated in this study*

Organism	NCBI accession code	Number of amino acid residues	Molecular weight (Da)	*E. coli* expression pattern*
Halococcus thailandensis	WP_007741205.1	481	54,071	NE
Haloterrigena saccharevitans	WP_076148579.1	467	52,247	NE
Natronomonas moolapensis	WP_015408604.1	469	52,211	I
Halorubrum hochstenium	WP_008581146.1	480	53,356	I
Halococcus sediminicola	WP_049997164.1	478	53,267	I
Methanohalobium evestigatum	WP_013193561.1	495	56,855	I > S
Candidatus Odinarchaeota archaeon LCB_4	OLS18764.1	493	56,708	I > S
Candidatus Nitrosotalea devanaterra	CUR51878.1	472	53,817	S > I
Coxiella brunetii	WP_005772347.1	470	53,963	I > S

* NE – no expression observed, I – insoluble, S – soluble

Table S2 – *Quantitative amino acid analysis of purified NdCCD*[1].

Amino acid	known number of residues	predicted % molar composition	experimental % molar composition
Asx	51	10.8	10.9
Glx	45	10.4	9.7
Ser	36	7.9	7.7
His	12	2.3	2.6
Gly	23	5.3	4.9
Thr	30	6.2	6.4
Ala	32	7.2	6.9
Arg	18	4.1	3.9
Tyr	24	4.6	5.2
Val	34	6.9	7.3
Met	7	1.2	1.5
Phe	34	7.3	7.3
Ile	30	5.9	6.4
Leu	35	7.9	7.5
Lys	35	7.6	7.5
Pro	20	4.5	4.3

[1] Two Cys and four Trp were not assayed

Table S3 – *First row transition metal content of NdCCD samples*

Expression Media	Protein concentration (mg/mL)	Metal concentration (ppm)						target metal/*Nd*CCD
		Fe	Co	Cu	Mn	Ni	Zn	
LB	2.9	0.68	<0.01	0.01	<0.01	<0.01	0.08	0.9
M9 + Co	2.9	0.10	1.02	0.01	<0.01	<0.01	0.05	1.6
buffer alone	0	0.01	<0.01	0.04	<0.01	<0.01	0.10	N/A

Table S4 - *X-ray diffraction data collection and structure refinement statistics*

Data collection and processing					
Crystal	Fe-*Nd*CCD	Co-*Nd*CCD	Co-*Nd*CCD-product complex	Co-*Nd*CCD post-Co K-edge*	Co-*Nd*CCD pre-Co K-edge*
X-ray source	NSLS-II FMX 17-ID-2	NSLS-II FMX 17-ID-2	NSLS-II FMX 17-ID-2	NSLS-II AMX 17-ID-1	NSLS-II AMX 17-ID-1
Wavelength (Å)	0.979354	0.979364	0.979356	1.606010	1.612290
Space group	$P6_1$	$P6_1$	$P6_1$	$P6_1$	$P6_1$
Unit cell lengths (Å)	a = 107.27, c = 491.11	a = 108.01, c = 491.61	a = 107.25, c = 490.24	a = 107.48, c = 491.89	a = 108.05, c = 493.06
Resolution (Å)†	50-2.68 (2.85-2.68)	50-2.30 (2.44-2.30)	50 – 2.35 (2.49 – 2.35)	50 – 3.20 (3.39 – 3.20)	50 – 3.2 (3.39 – 3.20)
Unique reflections	88,233 (13,897)	142,530 (22,771)	131,487 (21,124)	104,112 (16,658)	105,324 (16911)
Multiplicity	10.5 (10.1)	8.8 (8.9)	10.6 (10.2)	5.4 (5.5)	5.4 (5.5)
Completeness (%)	99.5 (97.5)	99.8 (98.8)	99.9 (99.2)	99.2 (98.4)	99.2 (98.5)
$\langle I/\sigma I \rangle$	11.77 (0.97)	10.59 (0.79)	12.83 (1.05)	7.86 (1.08)	6.95 (0.87)
$R_{merge}I$ (%)	18.5 (195.4)	14.3 (217.3)	13.0 (206.0)	15.1 (122.2)	18.1 (152.5)
$R_{pim}I$ (%)**	6.1 (74.0)	5.1 (88.5)	4.2 (78.7)	7.2 (63.1)	8.7 (79.1)
$CC_{1/2}$ (%)	99.9 (37.3)	99.8 (44.0)	99.9 (44.2)	99.7 (50.3)	99.6 (39.5)
Wilson *B* factor (Å^2)	66	61	62	89	88
Refinement					
Resolution (Å)	49.15 – 2.69	49.44 – 2.30	49.13 – 2.35		
No reflections‡	83,770 (4,373)	135,427 (7,076)	124,967 (6,518)		
R_{work}/R_{free} (%)	25.2/27.1	23.8/25.4	21.3/23.4		
No atoms]	21,655 (14,797)	21,658 (14,893)	21,835 (14,988)		
Protein	21,394 (14,538)	21,188 (14,468)	21,286 (14,476)		
Metal	6 (4) Fe	6 (4) Co	6 (4) Co		
Water	244	460	443		
Ligand	8 BCT	--	96 QVM		
$\langle B$-factor$\rangle$ (Å^2)]	91.1 (78.1)	75.6 (67.8)	83.1 (71.5)		
Protein	91.5 (78.4)	76 (68.2)	83.7 (71.8)		
Metal	63.5 (53.6)	53.7 (47.3)	60.2 (49.8)		
Water	56.1	54.6	59		
Ligand	65.9 BCT	--	77.9 QVM		
RMS deviations					
Bond lengths (Å)	0.002	0.006	0.005		
Bond angles (°)	1.20	1.35	1.30		
Ramachandran plot (% favored/outliers)***	93.8/0	95.9/0	96.5/0		
Molprobity score (%)	100	99	100		
PDB accession code	6VCF	6VCG	6VCH		

* Bijvoet pairs kept separate for calculation of statistics

† Values in parentheses are for the highest resolution shell of data

** Calculated with the program Aimless (32).

‡ Values in parentheses indicate the number of reflections used for cross-validation

] Values in parentheses are for the four well-ordered chains A-D

*** Evaluated using Molprobity (22)

Table S5. *Selected EXAFS fits of Fe^{II}-NdCCD*[a]

	Fe-N/O			Fe-O			Fe•••His					
fit	**n**	**r**	**σ^2**	**n**	**r**	**σ^2**	**n**	**r**	**σ^2**	**ΔE_0**	**χ^2**	**R**
1	4	2.12	1.4							1.91	20.24	0.0482
2	*5*	*2.12*	*2.5*							*1.47*	*12.29*	*0.0292*
3	6	2.12	3.4							1.03	13.63	0.0324
4	4	2.13	1.8	1	2.05	1.1				1.50	16.90	0.0265
5	4	2.14	1.5	2	2.07	3.1				1.17	17.15	0.0269
6	5	2.12	2.5							1.23	25.15	0.1657
7	***5***	***2.13***	***2.5***				***4***	***3.12*** ***3.34*** ***4.33*** ***4.38***	***6.5*** ***2.7*** ***9.0*** ***10.7***	***1.35***	***10.03***	***0.0385***

[a] Fitting range was k = 2.0-14.0 $Å^{-1}$ (resolution = 0.13 Å) with back transform ranges of 1.0-2.2 Å for fits 1-5 and 1.0-3.95 Å for fits 6-7. r is in units of Å; σ^2 is in units of 10^{-3} Å; ΔE_0 is in units of eV; R represents the fractional mis-fit of the data, while χ^2 is the χ^2 fitting metric normalized by the number of independent data points in a given fit.