iScience, Volume *23*

Supplemental Information

Non-full-length Water-Soluble CXCR4^{QTY} and CCR5^{QTY} Chemokine Receptors: Implication for Overlooked Truncated but Functional Membrane Receptors

Rui Qing, Fei Tao, Pranam Chatterjee, Gaojie Yang, Qiuyi Han, Haeyoon Chung, Jun Ni, Bernhard P. Suter, Jan Kubicek, Barbara Maertens, Thomas Schubert, Camron Blackburn, and Shuguang Zhang

Supporting information

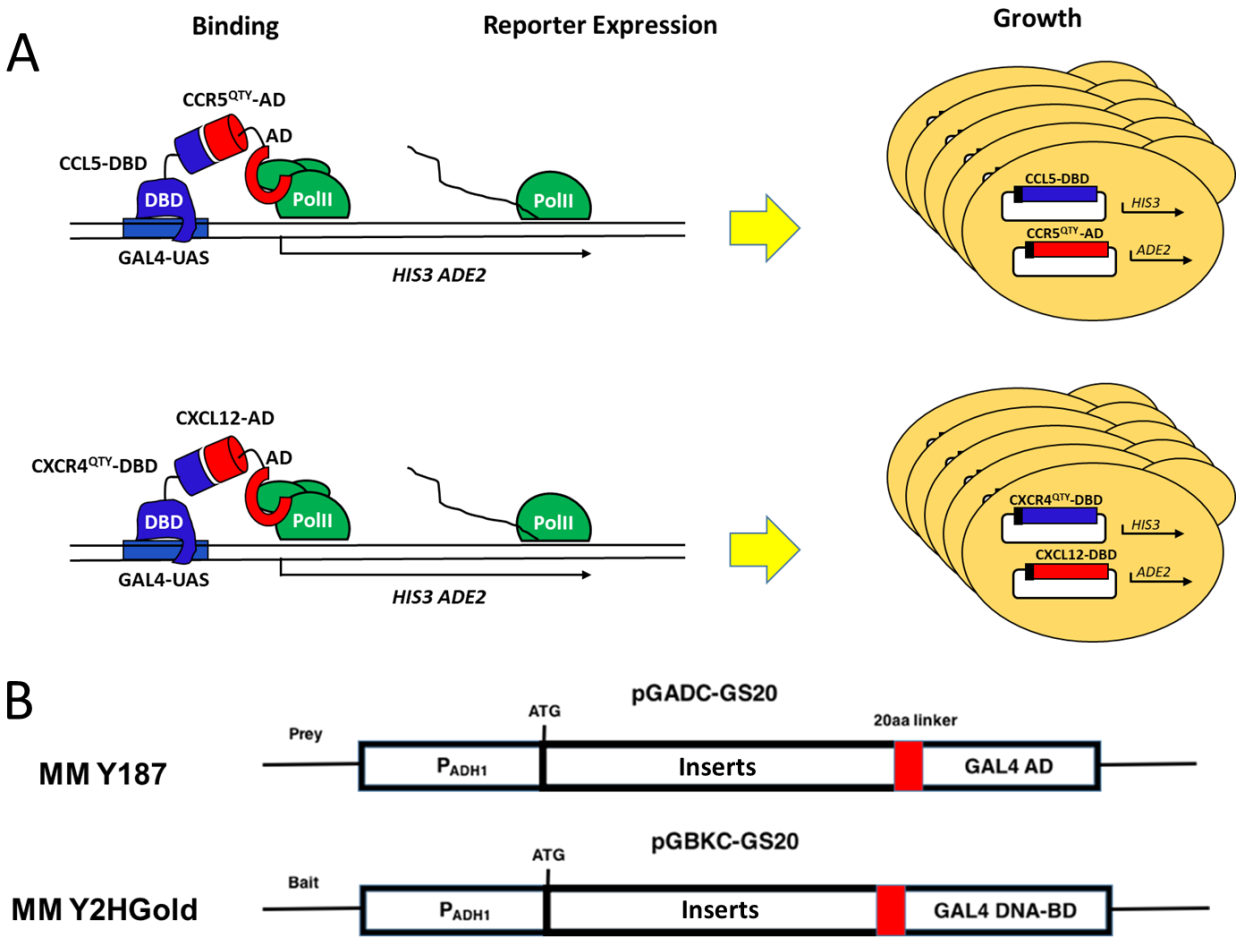

Figure S1. Schematic illustration for Y2H screening, Related to Figure 1. (A) Y2H screening for short CCR5^{QTY} and CXCR4^{QTY} variants. For CCR5^{QTY} screens, CCL5 ligand is in bait orientation in pGBKC-3C vector with CCR5^{QTY} (~3 million variants) in prey orientation in pGADC-2A. For CXCR4^{QTY} screens, CXCL12 ligand is in prey orientation in in pGADC-2A vector and CXCR4^{QTY} in bait orientation in pGBKC-3C. (B) Y2H sequence construct.

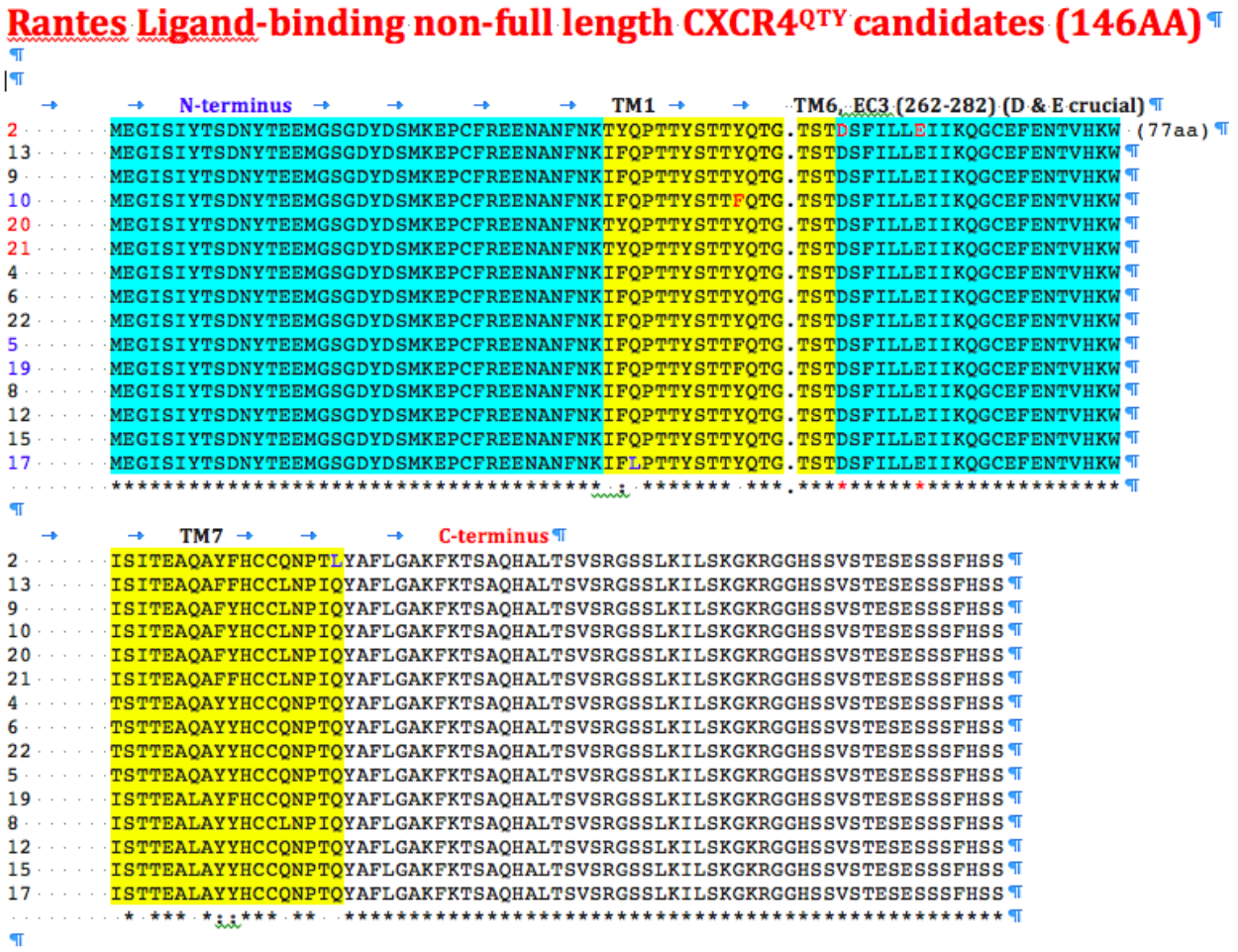

Figure S2. The protein sequences of 15 non-full-length CXCR4^{QTY} variants, Related to Figure 1. These variants were selected through Y2H screen and stringent mating system. Color code: Blue = N-termini and extracellular loop, yellow= transmembrane helical segments, TM= transmembrane. EC = extracellular domain. Both N- and C-terminus remain intact because of using the N- and C-terminal PCR primers.

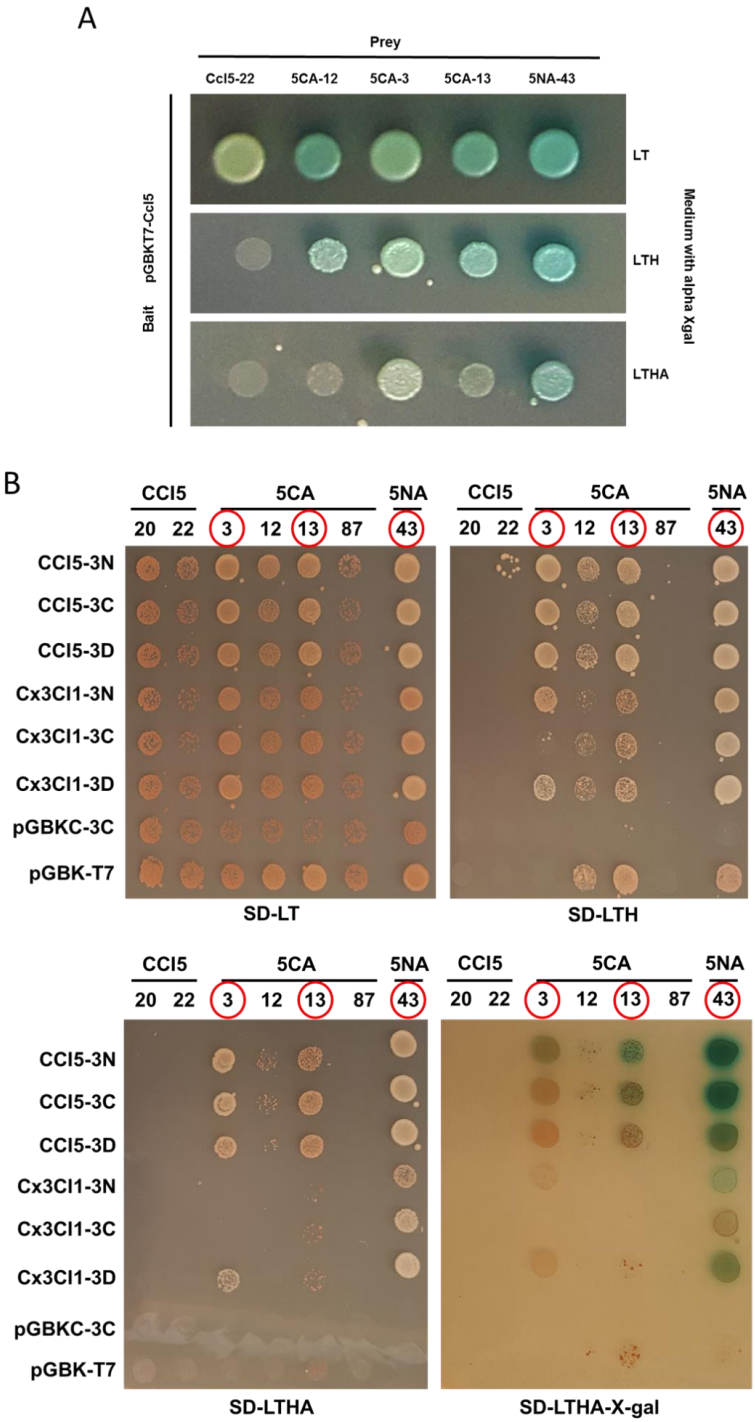

Figure S3. Y2H mating assay of 2 non-full-length CCR5^{QTY} variants, Related to Figure 1. The yeast clone 5NA43 is renamed as SZ218a to be consistent with other names. (A) Qualitative Y2H Interaction Test of CCL5 in Y2H Bait vector and CCR5^{QTY} variants CCL5-22, 5CA-12, 5CA-3, 5CA13, 5NA43 in Y2H Prey vector; (B) Spotting assay series for CCR5^{QTY} variants CCL5-20, 22, 5CA-3, 12, 13, and 5NA-43 and interaction mating with ligands CCL5, CX3CL1, and vector negative controls. Ligands are N-terminal (3N) or C-terminal fusions (3C, 3D) with DNA binding domains (Gal4-DBD).

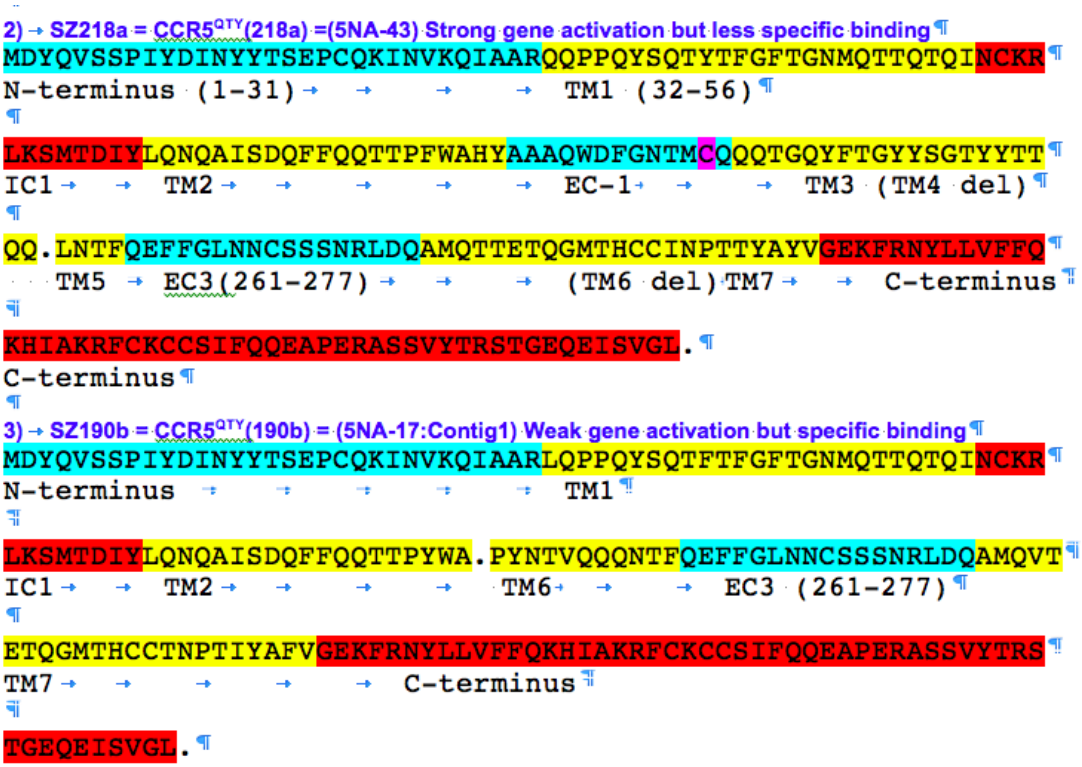

Figure S4. The protein sequences of 2 non-full-length CCR5^{QTY} variants, Related to Figure 2. These variants were selected through Y2H screen and stringent mating system. The proteins were purified, secondary structure was analyzed and ligand binding were studied. Color code: Blue = N-termini and extracellular loop, yellow= transmembrane helical segments, red= intracellular and C-termini. TM= transmembrane. IC= intracellular domain, EC = extracellular domain. Both N- and C-terminus remain intact because of using the N- and C-terminal PCR primers.

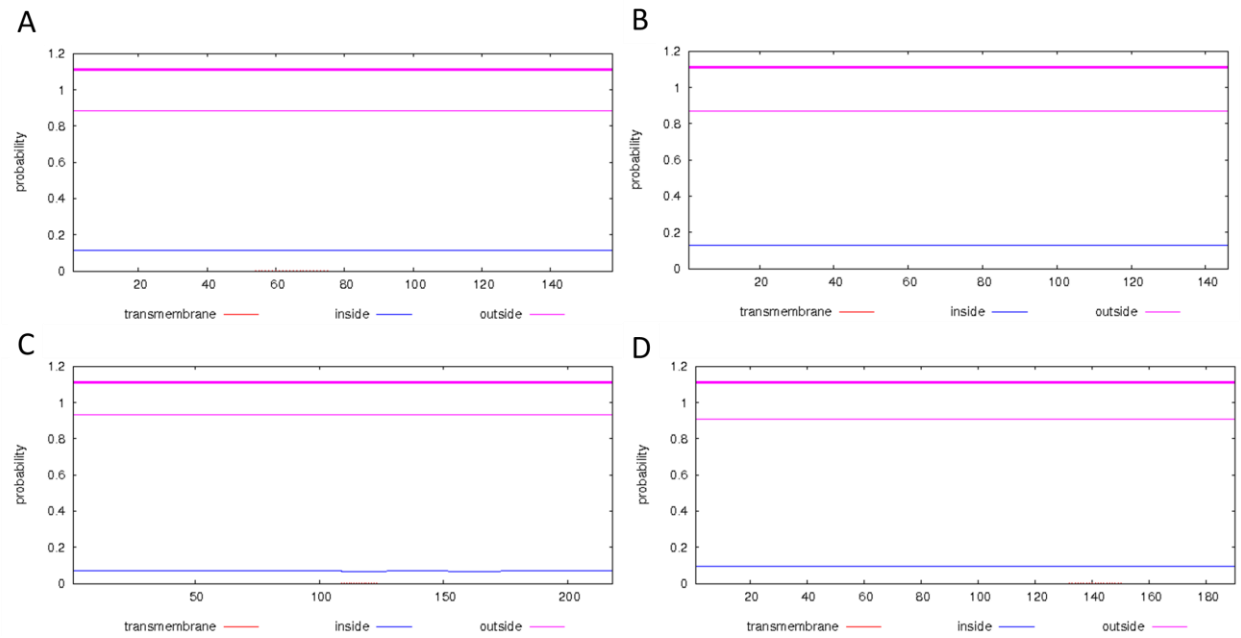

Figure S5. Bioinformatics hydrophobic segment analyses, Related to Figure 2. (A) CXCR4^{QTY}-SZ158a; (B) CXCR4^{QTY}-SZ146a; (C) CCR5^{QTY}-SZ218a; (D) CCR5^{QTY}-SZ190b**.** No hydrophobic TM region can be observed in any of the short variant GPCR^{QTY} proteins. X axis refers to the position of amino acids in the protein from N-terminus to C-terminus. Y-axis refers to the probability of hydrophobic TM segment.

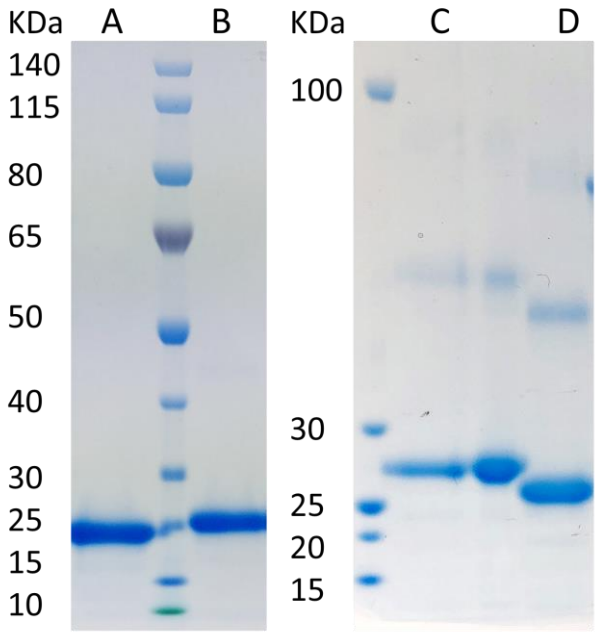

Figure S6. Electrophoresis band of purified non-full-length receptor proteins, Related to Figure 3. (A) CXCR4-SZ158a; (B) CXCR4-SZ146a; (C) CCR5-SZ218a; (D) CCR5-SZ190b. The molecular weight of the ladder is labelled on the left in KDa.

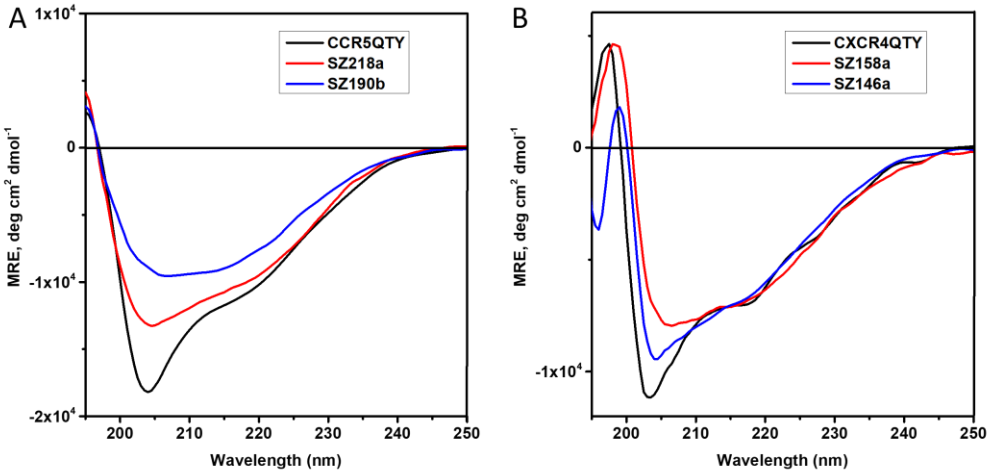

Figure S7. Secondary structure of nfCCR5^{QTY} and nfCXCR4^{QTY} receptors, Related to Figure 3. The Circular dichroism signal between 195nm and 250nm shows characteristic α-helical spectra. Difference in spectra shape is likely to be induced by inter-helix interaction in different proteins.

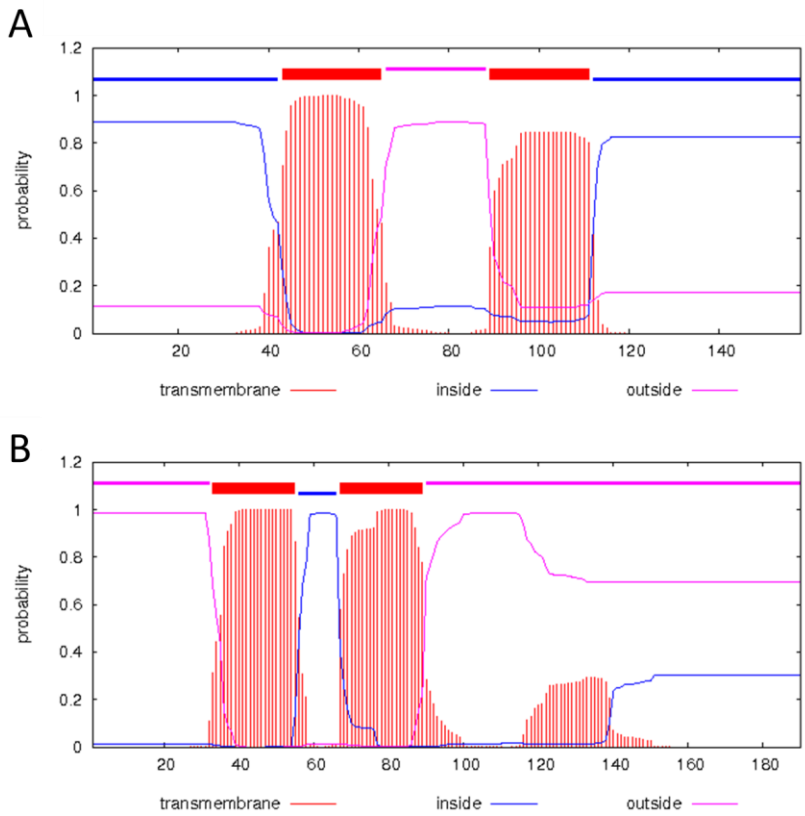

Figure S8. Bioinformatics hydrophobic segment analyses and topology prediction, Related to Figure 4. (A) CXCR4-SZ158a; (B) CCR5-SZ190b. Hydrophobic TM regions were predicted at sequences corresponding to original TM regions before truncation. The likelihood of EC and IC regions are suggested. X axis refers to the position of amino acids in the protein from N-terminus to C-terminus. Y-axis refers to the probability of hydrophobic TM segment.

Figure S9. How many fingers are required to hold a cup, Related to Figure 6. Usually five fingers are used to hold a cup, but a minimum of 2 fingers of various combinations are needed to hold a cup as shown here, although the cup is held less tightly. By analogy, the full-length of all 7TM domains with all 3EC loops are not absolutely required for ligand binding.

TRANSPARENT METHODS

DNA library CCR5 and CXCR4 bioinformatics design

The protein sequences of the human chemokine receptors CCR5 and CXCR4 were obtained from UniProtKB P51681 and UniProtKB P61073. The CCR5QTY and CXCR4QTY DNA libraries were designed and synthesized based on domain shuffling. First, a GPCR protein sequence was divided into 15 fragments based on its 7 transmembrane segments (7TM) and 8 non-transmembrane segments (N-terminal fragment, 3 intracellular loops, 3 extracellular loops and C-terminal fragment). Eight different positional variations were generated by applying the QTY Code. To make different variations, all or part of the changeable amino acid residues in a transmembrane fragment were changed. Only 8 variations were selected for each fragment based on their secondary structure and water solubility calculation results from RaptorX and TMHMM 2.0. Afterward, these fragments were reverse translated to DNA sequences and synthesized by Gen9. All DNA fragments were assembled randomly to form full length or non-full length GPCR genes.

The DNA library sequences were first run through Gen9 (Cambridge, MA, USA) special fragment assembly software to design the library of short fragment DNA. Subsequently, these fragments were made by first synthesizing 200 nucleotides and then assembling them together. Ligand constructs were synthesized by Integrated DNA Technologies.

Yeast 2 hybrid (Y2H) assays.

Y2H interactions were tested in *Saccharomyces cerevisiae* selection strain Y187 (MATα, ura3-52, his3-200, ade2-101, trp1-901, leu2-3, 112, gal4Δ, met-, gal80Δ, MEL1, URA3::GAL1uas-GALTATA-lacZ) and mating partner Y2HGold (MATa, trp1-901, leu2-3, 112, ura3-52, his3-200, gal4Δ, gal80Δ, LYS2::GAL1UAS–Gal1TATA–His3, GAL2UAS–Gal2TATA–Ade2, URA3::MEL1UAS–Mel1TATA, AUR1-C MEL1). Both strains were obtained from Clontech. These strains are effective in minimizing false positive protein interactions and background during a typical GAL4 based Y2H screen. Ligands and receptors were expressed in both strains for interaction testing in different orientations.

In our custom made Y2H vectors, the DNA binding and activation domains are at the C-termini of the Y2H fusion proteins. In pGADC-2A, the insert is separated by a multiple cloning site (MCS) and an HA-tag from the C-terminal GAL4 activation domain (GAL4-AD). In the modified pGADC-GS20 prey vector, the insert is separated from the GAL4-AD by a 20 amino acid polylinker (GS20) enriched in Serine and glycine (SGGGSGGGASSGGGAGGGAS). In the bait vector pGBKC-3C, the insert is separated by a MCS and a Myc-tag from the C-terminal GAL4 DNA binding domain (GAL4-DBD), while pGADC-GS20 contains the GS20 polylinker instead. Fusion protein expression in Y2H vectors is driven by ADH1 promoters. All bait and prey coding sequences are codon optimized for expression in *S. cerevisiae* and preceded by a Kozak sequence. Bait vectors contain the *TRP1* gene and prey vectors the *LEU2* gene for auxotrophic selection.

$CCR5^{QTY}$ variants were cloned via *in vivo* recombination into Y2H prey vector pGADC-2A. The $CCR5^{QTY}$ library was amplified for 9 PCR cycles with primers that anneal at both ends and that

also contain a 35-base overlap to the pGADC-2A target vector. Several aliquots of 5ng library template were amplified using the standard Phusion enzyme protocol (Thermo Fisher Scientific), and purified via gel extraction. For *in vivo* recombination, 2µg of the amplified library were co-transformed with 8µg of BamHI-EcoRI linearized vector pGADC-2A into the host strain Y187 following the library scale protocol for LiAc yeast transformation. Approximately $3x10^6$ primary clones were obtained from this transformation. The library was then expanded in Y187 and used for mating based Y2H assays. The Rantes (CCL5 26-91) gene was cloned into the bait vectors pGBKC-3C, variant pGBKC-3C (C-terminal GAL4-DBD) and pGBKT7 (N-terminal GAL4-DBD) and transformed into Y2HGold.

The $CXCR4^{QTY}$ DNA was amplified for 13 cycles with primers for homologous recombination and cloned into EcoRI-BamHI linearized pGBKC-3C via direct in vivo recombination in Y2HGold strain. The complexity of the library was estimated to be ~3 million primary clones. $CXCL12_{24-88}$ (SDF1α) was cloned into EcoRI-BamHI linearized prey vector pGADC-2A via direct in vivo recombination in Y187 strain for screening and mating retest.

Mating reactions between bait and prey strains were done for ~15 hours on yeast extract peptone with 2% dextrose (YPD), followed by growth on synthetic complete medium with 2% dextrose (SD) medium. Selection for reporter activation was for 3-5 days on stringent (*ADE2* and *HIS3* reporter selection, SD-LTHA) and non-stringent synthetic growth medium (*HIS3* reporter selection, SD-LTH medium). $CXCR4^{QTY}$ and $CCR5^{QTY}$ sequences from selected clones were amplified for sequencing, and plasmids are extracted from selected colonies and retransformed into fresh Y2HGold. 0.5-1 million cells were introduced for mating retests and spotting assays on SD-LTH, SD-LTHA and SD-LT (growth of all diploids). Blue coloration of colonies is observed on medium containing α-Xgal when the *MEL1* reporter is activated.

Bioinformatics of the QTY variants.

Protein properties were calculated based on its primary sequence via the open access web-based tool ExPASy: https://web.expasy.org/protparam/. The existence of hydrophobic patches within the transmembrane region in the variant protein sequences was determined via the open access web-based tool TMHMM Server v.2.0: http://www.cbs.dtu.dk/services/TMHMM-2.0/.

Protein expression, refolding, and purification from SF9 *Cell.*

$nfCCR5^{QTY}$ variant gene sequences selected in the Y2H screen were synthesized with a C-terminal His-tag (Biomatik). Sequences were cloned into a pOET2 transfer vector (Oxford Expression Technologies). The resulting baculovirus preparations were generated using the FlashBacUltra Kit (Oxford Expression Technologies) and amplified to high titer virus stocks. SF9 insect cells (Oxford Expression Technologies) were infected and cultured in 2-liter aerated spinner flasks in serum-free medium (Lonza) for 48 hours post infection at 27°C. Cells were collected by centrifugation at 1,500 rpm and the cell pellet was stored at –80°C.

SF9 Cells were lysed by sonication in PBS buffer, pH7.5, 10mM DTT. No detergent was used. The cells were centrifuged at 20,000×g and the supernatants were subjected to batch binding for 2 hours using a DTT stable Ni-Agarose resin (PureCube 100 INDIGO, Cube Biotech). The bound

His-tagged protein was washed extensively using PBS, pH7.5, with 20mM imidazole. Protein was eluted with PBS, pH7.5, 250mM imidazole. Elution fractions were concentrated with Amicon centrifugal filter units (Merck Millipore) and loaded onto a Superdex 200 gel-filtration column (GE Healthcare). The final protein was eluted in PBS, pH7.5, and was concentrated using Amicon centrifugal filter units (Merck Millipore) to 0.5mg/ml.

Protein expression, refolding, and purification from *E. coli*.

Genes of QTY-modified chemokine receptor proteins were codon-optimized for *E. coli* expression and obtained from Genscript. The genes were cloned into pET20b expression vector with Carbenicillin resistance. The plasmids were reconstituted and transformed into *E. coli* BL21(DE3) strain. Transformants were selected on LB medium plates with 100μg/ml Carbenicillin. *E. coli* cultures were grown at 37°C until the OD600 reached 0.4-0.8, after which IPTG (isopropyl-D-thiogalactoside) was added to a final concentration of 1mM followed by 4-hour expression. Cells were lysed by sonication in B-PER^{TM} protein extraction agent (Thermos-Fisher) and centrifuged (23,000×g, 40min, 4°C) to collect the inclusion body. The biomass was then subsequently washed twice in buffer 1 (50mM Tris.HCl pH7.4, 50mM NaCl, 10mM CaCl2, 0.1%v/v Trition X100, 2M Urea, 0.2μm filtered), once in buffer 2 (50mM Tris.HCl pH7.4, 1M NaCl, 10mM CaCl2, 0.1%v/v Trition X100, 2M Urea, 0.2μm filtered) and again in buffer 1. Pellets from each washing step were collected by centrifugation (23,000×g, 25min, 4°C).

Washed inclusion bodies were fully solubilized in denaturation buffer (6M guanidine hydrochloride, 1×PBS, 10mM DTT, 0.2μm filtered) at room temperature for 1.5 hour with magnetic stirring. The solution was centrifuged at 23,000×g for 40 min at 4°C. The supernatant with proteins was then purified by Qiagen Ni-NTA beads (His-tag) followed by size exclusion chromatography using an ÄKTA Purifier system and a GE healthcare Superdex 200 gel-filtration column. Purified protein was collected and dialyzed twice against renaturation buffer (50mM Tris.HCl pH 9.0, 3mM reduced glutathione, 1mM oxidized glutathione, 5mM ethylenediaminetetraacetic acid, and 0.5M L-arginine). Following an overnight refolding process, the re-natured protein solution was dialyzed against 50mM Tris.HCl pH 9.0 with various arginine content, and filtered through a 0.2μm syringe filter to remove aggregates.

Protein expression in human HEK293T cell.

Genes for full-length CXCR4 (OHu24159) and CCR5 (OHu20119) were directly purchased from Genscript and used as it is. Genes for non-QTY version of CXCR4-SZ158a and CCR5-SZ190b were made by identifying and extracting corresponding regions of NM_003467.3 and NM_000579.3 from Genebank and directly synthesized by IDT (Integrated DNA Technologies) without codon optimization. The genes were cloned into the standard pcDNA3.1 vector backbone under the human CMV promoter for expression. For confocal microscopy, a superfolder-GFP (sfGFP) tag was fused onto the C-terminus of each receptor for visualization.

HEK293T cells were maintained in Dulbecco's modified Eagle's medium supplemented with 100 units ml^{-1} of penicillin, 100 mg ml^{-1} of streptomycin and 10% fetal bovine serum. Receptor plasmids (100 ng) were transfected into cells as duplicates (2 × 500 μl per well in a 24-well plate for confocal microscopy or 2 × 104 μl per well in a 96-well plate for signaling assay) with

Lipofectamine 3000 (Invitrogen) in Opti-MEM (Gibco). Subsequent experiments were carried out after 1 d post-transfection.

MicroScale Thermophoresis.

MicroScale Thermophoresis (MST) is an optical method detecting changes in thermophoretic movement and TRIC of the protein-attached fluorophore upon ligand binding. Active labelled proteins contribute to the thermophoresis signal upon ligand binding. Inactive proteins influence the data as background but not the signals and only data from binding proteins are used to derive the K_d value. Herein ligand binding experiments were carried out with 5nM NT647-labeled protein in respective buffer (nfCXCR4^{QTY}: 50mM Tris-HCl pH 9.0, 100mM Arginine; nfCCR5^{QTY}: 1 X PBS, 10mM DTT) with a gradient of respective ligands in on a Monolith NT.115 pico instrument at 25°C. Synthesized receptors were labeled with Monolith NT™ Protein Labeling Kit RED – NHS (NanoTemper Technologies) so as to obtain unique fluorescent signals. MST time traces were recorded and analyzed to obtain the highest possible signal-to-noise levels and amplitudes, >5 Fnorm units. The recorded fluorescence was plotted against the concentration of ligand, and curve fitting was performed using the K_d fit formula derived from the law of mass action. For clarity, binding graphs of each independent experiment were normalized to the fraction bound (0 = unbound, 1 = bound). MST measurements of SF9 synthesized non-full-length CCR5^{QTY} and non-full-length CXCR4^{QTY} were performed at 2bind GmbH, Regensburg, Germany. MST experiments of *E. coli.* synthesized non-full-length CCR5^{QTY} were performed in the Center for Macromolecular Interactions at Harvard Medical School with 2nd generation Monolith NT™ Protein Labeling Kit RED – NHS.

K_d fitting model:

K_d model is the standard fitting model based on law of mass action.
Curve fit formula:

$$F(c_T) = F_u + (F_b - F_u) * \frac{c_{AT}}{c_A}$$

$$\frac{c_{AT}}{c_A} = fraction\ bound = \frac{1}{2c_A} * (c_T + c_A + K_D - \sqrt{(c_T + c_A + K_D)^2 - 4c_T c_A})$$

F_u: fluorescence in unbound state
F_b: fluorescence in bound state
K_D: dissociation constant, to be determined
c_{AT}: concentration of formed complex
c_A: constant concentration of molecule A (fluorescent), known
c_T: concentration of molecule T in serial dilution

Circular dichroism (CD) measurements.

CD spectra were recorded using JASCO Model J-1500 Circular Dichroism Spectrometer in Biophysical Instrumentation Facility at MIT. The QTY protein sample was dialyzed and refolded into CD buffer (0.05 v/v% TFA, 1mM TCEP). For far UV CD, spectra between 195nm and 250nm were collected with a 0.5nm step size, 1nm bandwidth, and 50nm/min scanning speed in 0.1 cm path length cuvettes. Baselines were measured using dialysis buffer alone without any protein and subtracted from the protein spectra. The baseline-subtracted spectra were scaled to obtain Mean

Residue Ellipticity (MREs) and normalized by protein concentration. The protein concentrations were ~2.4μM, as determined by Nanodrop with calculated extinction coefficient.

Confocal microscopy

HEK293T cells were cultured in glass-bottom 24-well plates at 37 C in Dulbecco's modified Eagle's medium supplemented with 100 units ml^{-1} of penicillin, 100 mg ml^{-1} of streptomycin and 10% fetal bovine serum. Fluorescence imaging was performed using a Nikon Eclipse Ti-E inverted microscope. We used a CSU-W1 spinning disk confocal module, with a 40x 1.15NA Plan Apo long working distance water-immersion objective (Nikon). GFP was excited with a 488 nm laser, with 525/40 emission filter.

Cell-signaling assay

Calcium signaling was monitored by co-transfecting receptors with $G_{\alpha q}$ and loading with a calcium-sensitive fluorescent dye to measure changes in cytosolic calcium concentration after adding ligands. HEK293T cells were maintained in Dulbecco's modified Eagle's medium supplemented with 100 units ml^{-1} of penicillin, 100 mg ml^{-1} of streptomycin and 10% fetal bovine serum. Receptor plasmids (100 ng) were transfected alongside a plasmid expressing the $G_{\alpha q}$ subunit (100 ng) into cells as triplicates (2 × 104 μl per well in a 96-well plate) with Lipofectamine 3000 (Invitrogen) in Opti-MEM (Gibco). 18 hours post-transfection, media was removed and replaced with 100 uL of Rhod-4 dye loading solution (Abcam #112157) according to manufacturer's protocols and incubated at 37 °C for 30 minutes for subsequent calcium measurements.

Calcium signaling in response to 25nM and 100nM ligands for each receptor were monitored by a Tecan Spark microplate reader. The equipment was pre-warmed to 37 °C before measurements. The cells were excited at 540 nm, and emission was monitored at 590 nm. A control sample without ligand was measured in parallel as a reference. Baselines were established before any ligand was added to each well. Data acquisition was immediately started after adding the ligand over a time course of 180s. No pipetting or shaking was conducted intentionally to elongate the signaling time by introducing the factor of ligand diffusion. There is an equipment associated delay of ~ 5s. The plate was rested for an additional 180s and fluorescence reading was collected for another 10 min to observe the recovery of fluorescence to baseline. Three independent biological repeats were conducted to eliminate error and establish statistical significance.

Bioinformatics of naturally existed truncated GPCRs.

The data for pseudogene analysis was retrieved from databases of GENCODE (www.gencodegenes.org), CHESS (ccb.jhu.edu/chess), UniProt (www.uniprot.org), and NCBI's RefSeq (www.ncbi.nlm.nih.gov/refseq). GPCR related pseudogenes were screened by running a Perl script on the Ubuntu (18.04.3 LTS) to only keep the truncated ones left. The web-based server Protter (wlab.ethz.ch/protter/start) was used for analyzing the amino sequences of the screened GPCR pseudogenes and generating protein snake plots. Transmembrane region predictions were also checked with TMHMM-2.0 (www.cbs.dtu.dk/services/TMHMM-2.0) to avoid inaccuracy.