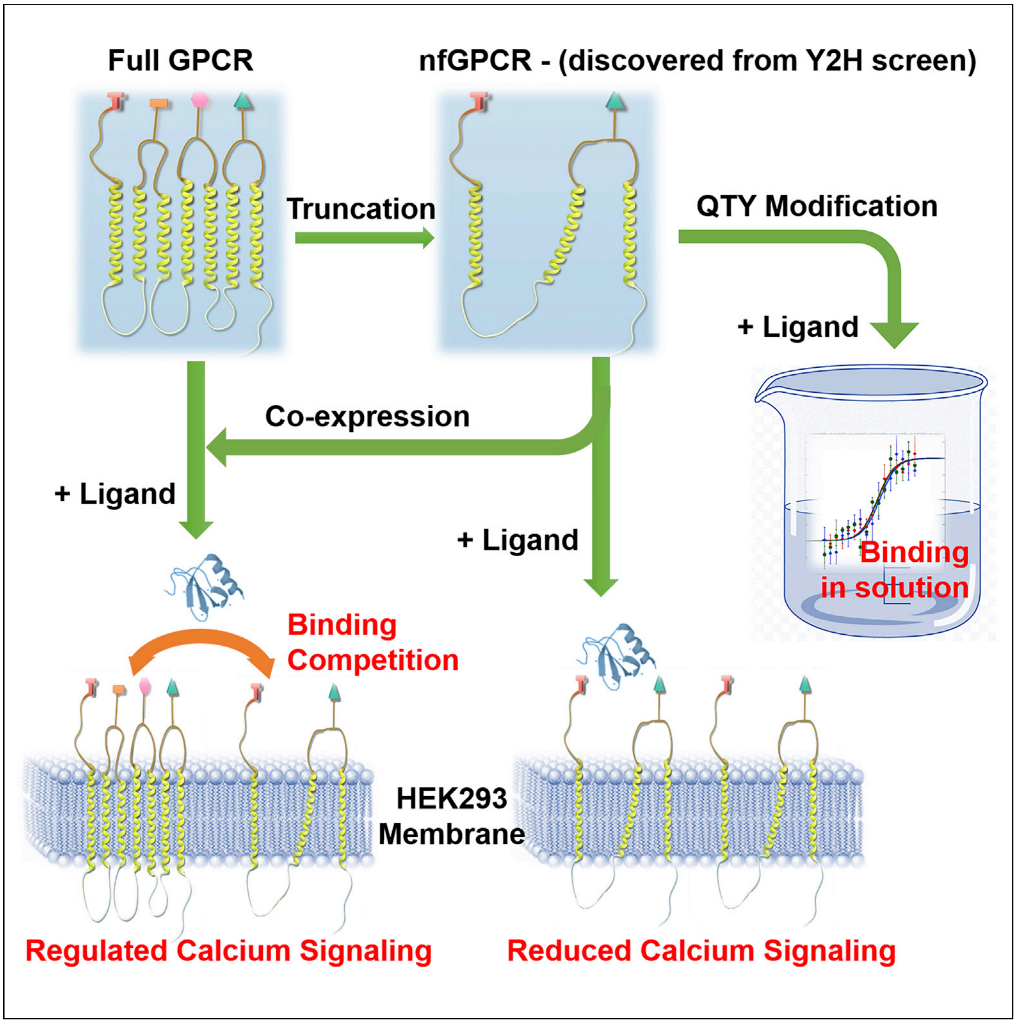
iScience

CellPress
OPEN ACCESS

Article

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, ..., Thomas Schubert, Camron Blackburn, Shuguang Zhang

ruiqing@mit.edu (R.Q.)
shuguang@mit.edu (S.Z.)

HIGHLIGHTS

Y2H screening reveals ligand interaction from truncated CXCR4 and CCR5 in QTY form

Truncated CCR5^{QTY} and CXCR4^{QTY} can be produced in E. coli and bind native ligands

Reconverted receptors localize on membranes and regulate cell signaling in HEK293

Our finding indicates potential presence and function for truncated receptors

Qing et al., iScience 23, 101670
December 18, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.isci.2020.101670

Article

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

Rui Qing,[1,7,8,*] Fei Tao,[2,7] Pranam Chatterjee,[1,6,7] Gaojie Yang,[1] Qiuyi Han,[1] Haeyoon Chung,[1] Jun Ni,[2] Bernhard P. Suter,[3] Jan Kubicek,[4] Barbara Maertens,[4] Thomas Schubert,[5] Camron Blackburn,[6] and Shuguang Zhang[1,*]

SUMMARY

It was posited that functionalities of GPCRs require full-length sequences that are negated by residue deletions. Here we report that significantly truncated nfCCR5^{QTY} and nfCXCR4^{QTY} still bind native ligands. Receptor-ligand interactions were discovered from yeast 2-hybrid screening and confirmed by mating selection. Two nfCCR5^{QTY} (SZ218a, SZ190b) and two nfCXCR4^{QTY} (SZ158a, SZ146a) were expressed in *E. coli*. Synthesized receptors exhibited α-helical structures and bound respective ligands with reduced affinities. SZ190b and SZ158a were reconverted into non-QTY forms and expressed in HEK293T cells. Reconverted receptors localized on cell membranes and functioned as negative regulators for ligand-induced signaling when co-expressed with full-length receptors. CCR5-SZ190b individually can perform signaling at a reduced level with higher ligand concentration. Our findings provide insight into essential structural components for CCR5 and CXCR4 functionality, while raising the possibility that non-full-length receptors may be resulted from alternative splicing and that pseudo-genes in genomes may be present and functional in living organisms.

INTRODUCTION

Alternative RNA splicing generates a diversity of proteins. Such a process increases the total number of proteins from a limited number of genes (Chaudhary et al., 2019), some of which have a variety of functions. These proteins through alternative RNA splicing are sometimes called truncated proteins; namely, they are no longer full length as encoded by the original gene. During the last few decades, researchers have engineered truncated proteins for both scientific studies and biotechnological applications (Den Dunnen and Van Ommen, 1999; Fersht and Winter, 1992; Fersht, 2008). Most such truncated variants studied are water-soluble proteins (Den Dunnen and Van Ommen, 1999; Fersht, 2008). Few truncated membrane receptor proteins have been systematically studied biochemically and biophysically because they require detergents and notoriously difficult to study.

It is generally believed that truncated membrane receptors are no longer functional. One example is the chemokine receptor CCR5Δ32 mutation of the CCR5 co-receptor for HIV entrance into CD4$^+$ and CD8$^+$ T cells (Deng et al., 1996). The CCR5Δ32 mutation has been shown to prevent HIV infection because it has a deletion of 32 DNA base pairs of the EC2 loop between DNA sequences 553 and 585, resulting in a frameshift translation of a truncated receptor (Samson et al., 1996). There are other examples of truncated receptors that lose function including the V2 vasopressin receptor (Zhu and Wess, 1998), the dopamine D3 receptor (Karpa et al., 2000), the mu opiate receptor (Majumdar et al., 2011), the trkB neural receptor (Middlemas et al., 1991), and others (Wise, 2012).

However, some truncated receptors have been shown to have no obvious functional defects (Wise, 2012). For example, the neurotensin receptor with 5TM loops and a long tail is functionally active to form a heterodimer with NTS$_2$ (Perron et al., 2005). Somatostatin receptors sst5TMD4 and sst5TMD5 are 4TM and 5TM truncated mutants, respectively. They are present in normal and tumor tissues (Cordoba-Chacon et al., 2010; Duran-Prado et al., 2009).

[1] Media Lab, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[2] Laboratory of Food Microbial Technology, State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiaotong University, Shanghai 200240, China

[3] Next Interactions, Inc., 2600 Hilltop Drive, Building B, C332, Richmond, CA 94806, USA

[4] Cube Biotech, GmbH, Creative Campus, Alfred-Nobel Strasse 10, 40789 Monheim, Germany

[5] 2bind GmbH, Am BioPark 11, 93053 Regensburg, Germany

[6] The Center for Bits and Atoms, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[7] These authors contributed equally

[8] Lead Contact

*Correspondence: ruiqing@mit.edu (R.Q.), shuguang@mit.edu (S.Z.)

https://doi.org/10.1016/j.isci.2020.101670

Ling et al. deleted 72 amino acids that corresponded to TM1, IC1, TM2, and EC1 of chemokine receptors CCR5 and CXCR4 to produce slightly truncated receptors (Ling et al., 1999). They showed that such truncated receptors still carried out cell signaling in human embryonic kidney (HEK) 293T cells when exposed to their respective ligands, CCL5 (Rantes) and CXCL12 (SDF1α). Thus, it is possible that some truncated receptors may still bind their ligands and carry out signaling despite significant deletions.

The study of membrane receptor proteins requires detergents to prevent their aggregation in aqueous solutions (Lv et al., 2016; Qing et al., 2019; Vinothkumar and Henderson, 2010). Detergent screening is generally a prerequisite to work on membrane proteins *in vitro* (Lin and Guidotti, 2009; Skrzypek et al., 2018).

We previously reported a simple QTY code for systematic membrane protein design. The QTY code substitutes hydrophobic amino acids with hydrophilic ones that are structurally similar but with different chemical properties, so as to design the detergent-free, water-soluble, and functional variants of chemokine receptors (Zhang et al., 2018). The QTY code systematically replaces the hydrophobic amino acids Leu, Val, Ile, and Phe with hydrophilic Gln (Q), Thr (T), and Tyr (Y) in the receptors, particularly in the transmembrane domains, based on their structures and electron density map similarity of amino acids. This approach permits flexibility in designing and studying the physiological and functional properties of these chemokine receptors, while providing extra freedom in their utilization by eliminating the necessity of detergents. The QTY variant of chemokine receptors can be readily produced in multiple hosts and purified without any detergents.

During the screening of gene library design of CCR5^{QTY} and CXCR4^{QTY} in the yeast 2-hybrid (Y2H) system (Figure S1), the yeast colonies that bear CCR5^{QTY} and CXCR4^{QTY} in vectors and their respective ligands CCL5 and CXCL12 in vectors underwent stringent screen and complementary mating tests. We obtained the expected full-length detergent-free CCR5^{QTY} and CXCR4^{QTY} variants. Unanticipated non-full-length, truncated variants of CCR5^{QTY} and CXCR4^{QTY} were also found during DNA sequencing of these yeast colony clones.

We here report that several non-full-length nfCCR5^{QTY} and nfCXCR4^{QTY} chemokine receptors retain ligand binding *in vivo* and *in vitro*. Y2H mating tests revealed many short receptor variants with gene activation via ligand interaction by screening ~3 million gene sequences. We chose two variants of nfCCR5^{QTY}: SZ218a and SZ190b, and two variants of nfCXCR4^{QTY}: SZ158a and SZ146a, to codon-optimize for expression in SF9 insect cells and *E. coli*. These non-full-length chemokine receptors exhibited binding activity *in vitro* with their respective ligands, namely, CCL5 for nfCCR5^{QTY} and CXCL12 for nfCXCR4^{QTY}, albeit with reduced affinity. The nfCCR5^{QTY} and nfCXCR4^{QTY} possessed the N terminus and parts of the EC loops, especially the EC3 loop. The truncated receptors also showed the α-helical structure. Two of the truncated receptors, CCR5^{QTY}-SZ190b and CXCR4^{QTY}-SZ158a, were reconverted to non-QTY forms and expressed in HEK293T cells. Confocal microscopy revealed that these receptors preferentially localized on cell membranes. Signaling assays indicated that truncated receptors negatively regulate ligand-induced signaling of full-length receptors when co-expressed. To our great surprise, CCR5-SZ190b, albeit with large deletion of sequence (190aa/352aa), still carried reduced signaling capability to induce intracellular activity at a higher ligand concentration. Our observations raise the plausibility that some so-called pseudogenes may be present and still active in cells. More systematic analyses will be needed to understand the full biological activity for non-full-length genes and their encoded proteins *in vivo*. Study of these non-full-length receptors can also provide insight into the functionality mechanism for full-length chemokine receptors, which may enable a number of biotechnological, diagnostic, and therapeutic applications.

RESULTS

Y2H Assay for Receptor and Ligand Interactions

The Y2H assay was initially used to study the *in vivo* interactions between different types of full-length QTY variants and their respective ligands. During this process, numerous short length proteins with QTY modification but not full 7TM structures were discovered to exhibit affinity toward the respective ligands.

In Y2H experiments, the ligands and receptors were cloned into custom-made Y2H bait and prey vectors to allow ligand-receptor interactions. The receptor-ligand interaction activates gene transcription, thus enabling yeast cell growth. Only those variants that are folded properly in the intracellular milieu and transported into the yeast nucleus are able to activate gene transcription of the Y2H reporters. The variants were further subjected to control assays to eliminate false-positives. Yeast GAL4 activation and DNA-binding domains are at the C terminus of the fusion proteins, leaving both free receptor and chemokine N terminus. The schematic for Y2H setup is shown in Figure S1.

We screened a library of ~3 million CXCR4^{QTY} receptor variants fused to the C-terminal DNA-binding domain (pGBKC-3C) with CXCR4^{QTY} in bait orientation and CXCL12 in pGADC-2A (C-terminal activation domain) as the prey. Screens were done on stringent medium lacking adenine and histidine and CXCR4^{QTY} library with CXCL12. About 1 in 500 diploid clones activated the HIS3 reporter (0.2%, SD-LTH non-stringent), and about 1 in 25,000 activated HIS3 and ADE2 (0.004%, SD-LTHA stringent). We obtained no selectable clones when the ligand was N-terminally tagged (pGAD-HA).

We picked 22 clones from the CXCR4^{QTY}/CXCL12 screen that grew on high-stringency medium lacking both adenine and histidine and characterized them by colony PCR. Bidirectional sequencing showed 22 selected clones of shortened version of CXCR4 ^{QTY}. Figure S2 shows the 15 variants of non-full length CXCR4 clones, all of which contain N terminus and EC3. On the other hand, PCR of random clones after transformation revealed full-length CXCR4 inserts. Hence, the short CXCR4 fragments were specifically selected in the Y2H screen. These characterized CXCR4^{QTY} clones were retested in a 1:1 mating assay with bait CXCL12 with DBD at C terminus (pGADC-2A), and at N terminus (pGAD-HA), and with the two empty vectors as controls. The absence of Y2H reporter activation in the empty vector controls showed that the interactions of these selected CXCR4^{QTY} clones is specific. The interactions were only observed when the ligand was tagged at the C terminus in the Y2H activation domain (AD). The retest results were very reproducible on different selection media, also showing differential interaction strengths among the CXCR4 ^{QTY} clones.

We characterized the interactions between four individual CXCR4^{QTY} clones CXCR4^{QTY} #1 [renamed as SZ158a], #4 [renamed as SZ146a], #7 [not pursued further], and #22 [renamed as 146b] and CXCL12 ligand in semi-quantitative interaction assays (Figures 1A and 1B). Interactions were tested on non-stringent (SD-LTH) and stringent (SD-LTHA) selection media and were found to be fully dependent on the presence of the ligand. Again, it appears that interactions with some of the CXCR4^{QTY} clones SZ146a and 146b are stronger than with numbers SZ158a and #7. As it turned out, the DNA sequences of clones SZ146a and 146b are identical.

In addition, we also screened a complex library of ~2 million CCR5^{QTY} variants in the Y2H prey vector with its ligand CCL5 as the bait to find variants that bind with high affinity. Several variants were found and plasmids were then isolated from these clones, transformed into fresh Y187 prey strain, and then were subjected to a stringent 1:1 mating test with original ligand constructs (Figure S3A).

Mating interaction assays showed reporter activation when four selected nfCCR5^{QTY} variants 5CA-12, 5CA-3, 5CA-17 (renamed as SZ190b), and 5NA-43 (renamed as SZ218a) were combined with the native ligand CCL5, but not with the full-length CCR5-22 cloneFigure S3. The growth on non-stringent and stringent media, and alpha-galactosidase color formation by MEL1 marker, shows differences in interaction strength in the Y2H interactions with 5NA-43 being the strongest, followed by 5CA-3 and 5CA-13 with 5CA-12 being the weakest. These results were further confirmed in a more detailed assay with differentially tagged bait ligands CCL5, CX3CL1, and empty vector controls showing preferential binding to CCL5 (Figure S3B). Sequence analysis of 5NA-43 (renamed as SZ218a) is presented in Figure S4. Other clones isolated from the screen (5CA-87) did not show growth under these conditions, similar to empty bait vectors without a ligand insert. The interaction was verified by replicate tests.

Having shown that non-full-length CXCR4^{QTY} and CCR5^{QTY} variants can activate Y2H reporters when combined with their respective ligands, we selected several clones that yeast cells formed into colonies on stringent medium plates. DNA from these colonies was purified and then sequenced. These non-full-length DNA sequences were later cloned into both baculovirus expression vector pOET2 to express in SF9 insect cells and also pET20b+ to express in *E. coli*. The proteins from both SF9 insect cells and *E. coli* were affinity purified. Subsequent experiments were carried out to measure the molecular interactions *in vitro* using microscale thermophoresis (MST).

Sequence Alignments of nfCCR5^{QTY} and nfCXCR4^{QTY} Proteins with Full-Length and Native Counterparts

Sequence alignments were performed to show QTY code changes and truncations for nfCCR5^{QTY} and nfCXCR4^{QTY} receptors (Figure 2). Despite the partial inclusion of the gene sequence, all nfCCR5^{QTY} and

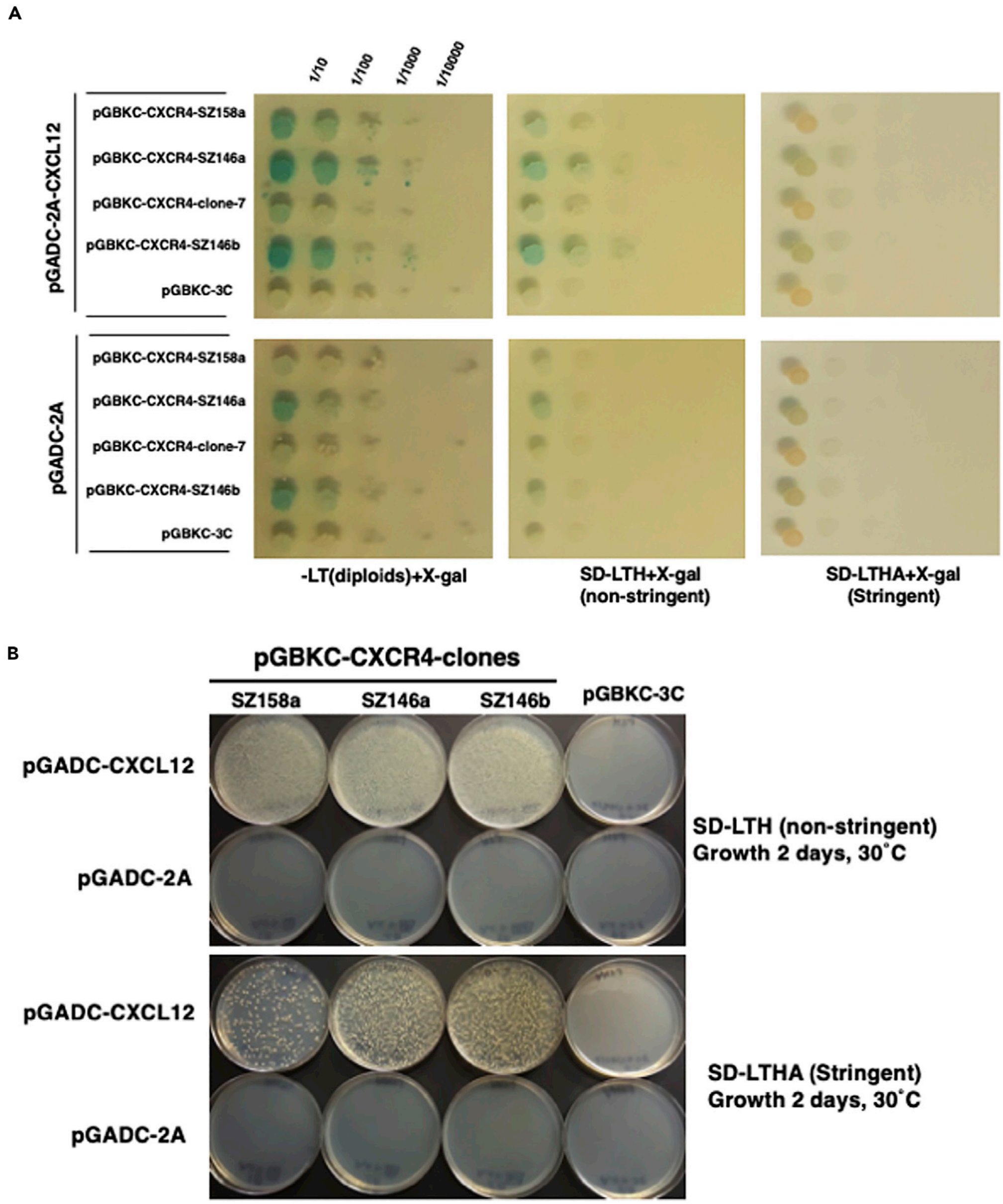


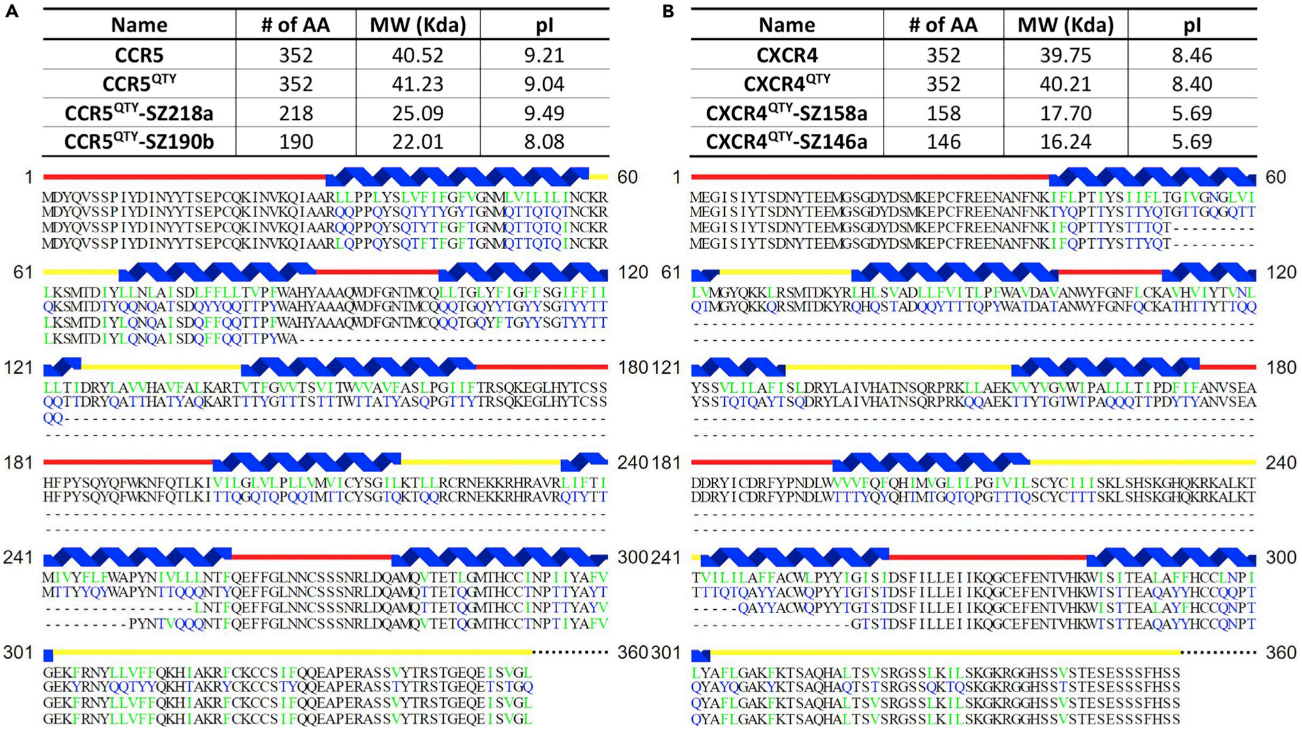
Figure 1. Y2H Mating Assay for nfCXCR4^{QTY}

(A) Yeast colonies were initially picked from plates that allow the activated gene transcription. Cells were spotted in serial 10X dilutions and grown on selective medium with α-Xgal (SD-LT all diploids, SD-LTH non-stringent reporter selection, SD-LTHA stringent reporter selection). The original colonies have been renamed for subsequent studies: #1 (renamed as SZ158a), #4 (renamed as SZ146a), #7 (not pursued further due to weak interaction), and #22 (renamed as SZ146b). SZ146a (#4) and SZ146b (#22) have identical DNA sequence, likely due to PCR amplifications.

(B) The selected clones are further tested on non-stringent (SD-LTH) and stringent (SD-LTHA) selection medium at 30°C for 2 days. If the interactions between the ligand and receptor are strong, these yeast cells will grow, otherwise, cells will not grow. The lower panels are the negative controls without the CXCL12 ligand in the vector, thus no cell growth. Similar screen was carried out for CCR5^{QTY} library with CCL5 ligand.

See also Figures S1–S3.

nfCXCR4^{QTY} proteins strictly follow the rule where only hydrophobic residues in the original transmembrane (TM) regions were replaced by glutamine (Q), threonine (T), and tyrosine (Y). Residues in any other location were untouched, including fragments corresponding to the original N terminus, EC loops, IC loops, and C terminus. The QTY code application diminished the potential existence of hydrophobic



A

Name	# of AA	MW (Kda)	pI
CCR5	352	40.52	9.21
CCR5^{QTY}	352	41.23	9.04
CCR5^{QTY}-SZ218a	218	25.09	9.49
CCR5^{QTY}-SZ190b	190	22.01	8.08

B

Name	# of AA	MW (Kda)	pI
CXCR4	352	39.75	8.46
CXCR4^{QTY}	352	40.21	8.40
CXCR4^{QTY}-SZ158a	158	17.70	5.69
CXCR4^{QTY}-SZ146a	146	16.24	5.69

1 60
MDYQVSSPIYDINYYTSEPCQKINVKQIAARLLPPLYSLVFIFGFVGNMLVILILINCKR
MDYQVSSPIYDINYYTSEPCQKINVKQIAARQQPPQYSQTYTYGYTGNMQTTQTQTNCKR
MDYQVSSPIYDINYYTSEPCQKINVKQIAARQQPPQYSQTYTFGFTGNMQTTQTQINCKR
MDYQVSSPIYDINYYTSEPCQKINVKQIAARLQPPQYSQTFTFGFTGNMQTTQTQINCKR
61 120
LKSMTDIYLLNLAISDLFFLLTVPFWAHYAAAQWDFGNTMCQLLTGLYFIGFFSGIFFII
QKSMTDTYQQNQATSDQYYQQTTPYWAHYAAAQWDFGNTMCQQQTGQYYTGYYSGTYYTT
LKSMTDIYLQNQAISDQFFQQTTPFWAHYAAAQWDFGNTMCQQQTGQYFTGYYSGTYYTT
LKSMTDIYLQNQAISDQFFQQTTPYWA---------------------------------
121 180
LLTIDRYLAVVHAVFALKARTVTFGVVTSVITWVVAVFASLPGIIFTRSQKEGLHYTCSS
QQTTDRYQATTHATYAQKARTTTYGTTTSTTTWTTATYASQPGTTYTRSQKEGLHYTCSS
QQ--
----
181 240
HFPYSQYQFWKNFQTLKIVILGLVLPLLVMVICYSGILKTLLRCRNEKKRHRAVRLIFTI
HFPYSQYQFWKNFQTLKITTQGQTQPQQTMTTCYSGTQKTQQRCRNEKKRHRAVRQTYTT

----
241 300
MIVYFLFWAPYNIVLLLNTFQEFFGLNNCSSSNRLDQAMQVTETLGMTHCCINPIIYAFV
MTTYYQYWAPYNTTQQQNTYQEFFGLNNCSSSNRLDQAMQTTETQGMTHCCTNPTTYAYT
----------------LNTFQEFFGLNNCSSSNRLDQAMQTTETQGMTHCCINPTTYAYV
---------PYNTVQQQNTFQEFFGLNNCSSSNRLDQAMQVTETQGMTHCCTNPTIYAFV
301 360
GEKFRNYLLVFFQKHIAKRFCKCCSIFQQEAPERASSVYTRSTGEQEISVGL
GEKYRNYQQTYYQKHTAKRYCKCCSTYQQEAPERASSTYTRSTGEQETSTGQ
GEKFRNYLLVFFQKHIAKRFCKCCSIFQQEAPERASSVYTRSTGEQEISVGL
GEKFRNYLLVFFQKHIAKRFCKCCSIFQQEAPERASSVYTRSTGEQEISVGL
1 60
MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFLPTIYSIIFLTGIVGNGLVI
MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKTYQPTTYSTTYQTGTTGQGQTT
MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTTQT---------
MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTYQT---------
61 120
LVMGYQKKLRSMTDKYRLHLSVADLLFVITLPFWAVDAVANWYFGNFLCKAVHVIYTVNL
QTMGYQKKQRSMTDKYRQHQSTADQQYTTTQPYWATDATANWYFGNFQCKATHTTYTTQQ

----
121 180
YSSVLILAFISLDRYLAIVHATNSQRPRKLLAEKVVYVGVWIPALLLTIPDFIFANVSEA
YSSTQTQAYTSQDRYLAIVHATNSQRPRKQQAEKTTYTGTWTPAQQQTTPDYTYANVSEA

----
181 240
DDRYICDRFYPNDLWVVVFQFQHIMVGLILPGIVILSCYCIIISKLSHSKGHQKRKALKT
DDRYICDRFYPNDLWTTTYQYQHTMTGQTQPGTTTQSCYCTTTSKLSHSKGHQKRKALKT

----
241 300
TVILILAFFACWLPYYIGISIDSFILLEIIKQGCEFENTVHKWISITEALAFFHCCLNPI
TTTQTQAYYACWQPYYTGTSTDSFILLEIIKQGCEFENTVHKWTSTTEAQAYYHCCQQPT
-----QAYYACWQPYYTGTSTDSFILLEIIKQGCEFENTVHKWISTTEALAYFHCCQNPT
-----------------GTSTDSFILLEIIKQGCEFENTVHKWTSTTEAQAYYHCCQNPT
301 360
LYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
QYAYQGAKYKTSAQHAQTSTSRGSSQKTQSKGKRGGHSSTSTESESSSFHSS
QYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
QYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS

Figure 2. Comparison of Full-Length Native CCR5, CXCR4, Their QTY Variants, and nfCXCR4^{QTY} and nfCCR5^{QTY} Receptors

(A and B) Sequence alignment between (A) native CCR5 (first row), CCR5^{QTY} (second row), nfCCR5^{QTY}-SZ218a (third row), and nfCCR5^{QTY}-SZ190b (fourth row); (B) native CXCR4 (first row), CXCR4^{QTY} (second row), nfCXCR4^{QTY}-SZ158a (third row), and nfCXCR4^{QTY}-SZ146a (fourth row). Substitutions of amino acids are highlighted in different colors. The original hydrophobic L, V, F, and I amino acids are denoted in green; the substitution water-soluble Q, T, and Y amino acids are in blue. The α-helical segments (blue) are shown above the protein sequences, and the external (red) and internal (yellow) loops of the receptors are indicated. Features of native, full-length, and non-full-length QTY chemokine receptors' number of amino acids, pI, and molecular weight are presented.

See also Figures S4–S5.

TM segments in short receptors, as shown in Figure S5. For full-length CCR5^{QTY} and CXCR4^{QTY} proteins, the latest versions with additional QTY modifications in IC loops and C terminus were presented as described in our previous publications (Qing et al., 2019; Zhang et al., 2018). As an additional note, the sequence of *E. coli*-synthesized CCR5^{QTY} contains extra modification in IC loops and C terminus when compared with that of SF9-synthesized CCR5^{QTY} as explained in our previous publication (Qing et al., 2019).

Figure 2A shows the sequence alignments of different variants of CCR5 receptor proteins. From top to bottom, the sequence in each row corresponds to native CCR5, CCR5^{QTY}, CCR5^{QTY}-SZ218a, and CCR5^{QTY}-SZ190b, respectively. CCR5^{QTY}-SZ218a contains N terminus, TM1, IC1, TM2, EC1, part of TM3, part of TM5, EC3, TM7, and C terminus of the full-length protein with deletion of IC2, TM4, IC3, and TM6. CCR5^{QTY}-SZ190b contains N terminus, TM1, IC1, part of TM2, part of TM6, EC3, TM7, and C terminus of the full-length protein with deletion of EC1, TM3, IC2, TM4, EC2, TM5, and IC3.

Figure 2B shows the sequences alignment of different variants of CXCR4 receptor proteins. From top to bottom, the sequence in each row corresponds to native CXCR4, CXCR4^{QTY}, CXCR4^{QTY}-SZ158a, and CXCR4^{QTY}-SZ146a, respectively. CXCR4^{QTY}-SZ158a contains N terminus, part of TM1, part of TM6, EC3, TM7, and C terminus of the full-length protein with deletion of IC1, TM2, EC1, TM3, IC2, TM4, EC2, TM5, and IC3. CXCR5^{QTY}-SZ146a is similar to CXCR4^{QTY}-SZ158a but with less residues in fused TM1 and TM6.

Full-length CCR5^{QTY} and CXCR4^{QTY} exhibit only slight differences in molecular weight (MW) and isoelectric point (pI) compared with native receptors, whereas non-full-length receptors show a large shift of pI

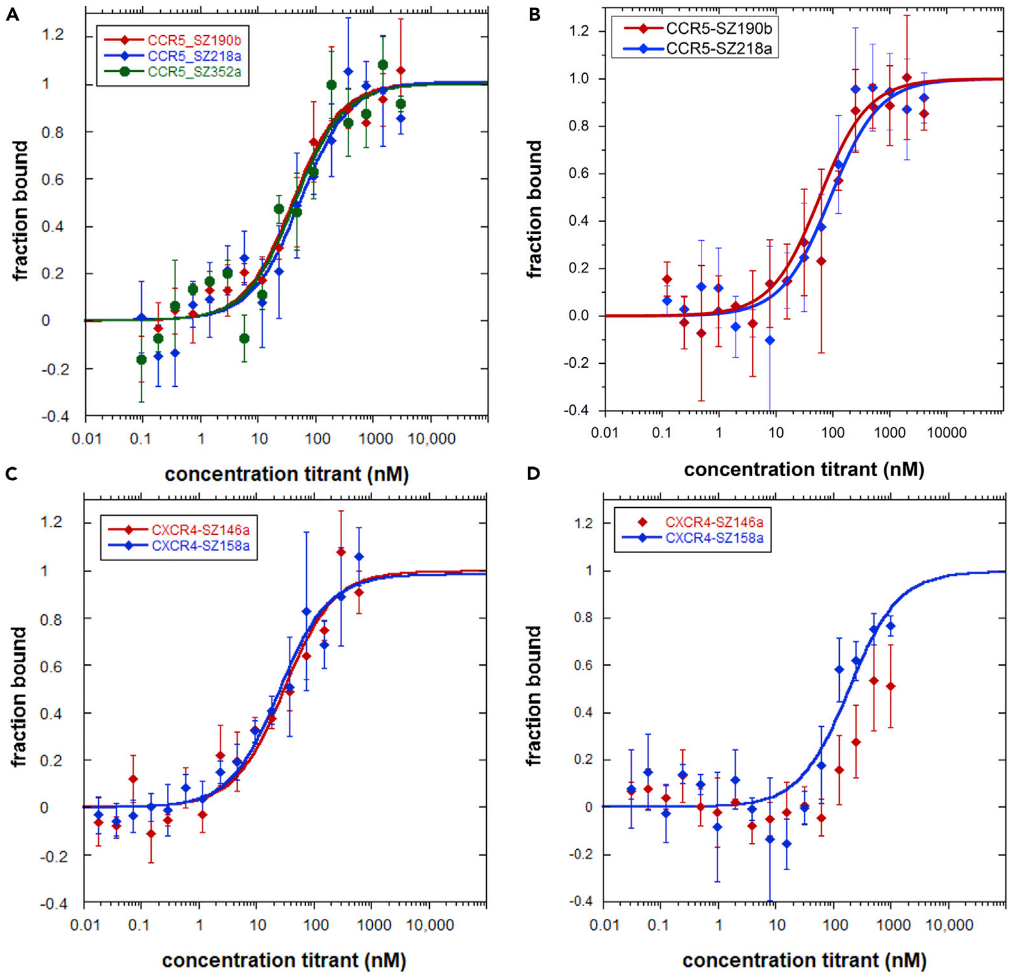


Figure 3. Microscale Thermophoresis (MST) Ligand Binding Measurements

The receptors were labeled with manufacture-provided fluorescent dye. Ligands were obtained commercially and serial diluted in deionized water. Error bars were calculated from three independent repeats of each sample.

(A–D) (A) SF9-synthesized nfCCR5^{QTY} with CCL5, (B) *E. coli*-synthesized nfCCR5^{QTY} with CCL5, (C) *E. coli*-synthesized nfCXCR4^{QTY} with CXCL12, (D) *E. coli*-synthesized nfCXCR4^{QTY} with HIV-1 coat protein gp$_{41-120}$. The K$_d$ value calculated from the graph can be found in Table 1.

See also Figures S6–S7.

value due to residue deletion, which leads to a severe structural change in folded condition. Additionally, exchanging hydrophobic L, I, V, F with hydrophilic Q, T, and Y induces the formation of inter- and intra-helical hydrogen bonds as well as with surrounding water molecules (Qing et al., 2019).

Ligand-Binding Measurement of nfCCR5^{QTY} and nfCXCR4^{QTY} Receptors Expressed and Purified from SF9 Insect Cells and *E. Coli*

The affinity of nfCCR5^{QTY} and nfCXCR4^{QTY} receptors toward their respective ligands CCL5 and CXCL12 were determined using MST (Figure 3). Purified proteins, as shown in Figure S6, were labeled with fluorescent dye through which the changes in their thermophoretic movement and temperature-related intensity changes upon ligand binding were recorded and plotted against ligand concentration (Seidel et al., 2013). No unspecific adhesion or major aggregation of protein was detected during the measurement for SF9-synthesized proteins. Minor ligand-induced aggregation was observed for *E. coli*-synthesized proteins, so the corresponding data were analyzed in early MST time trace. For better visualization, all data were replotted as bound fraction versus concentration. K$_d$ values were calculated using the

	CCL5[a] K_d, nM	CXCL12[a] K_d, nM	gp41-120 K_d, nM
CXCR4 native		~5	~200[b]
CXCR4^{QTY} (*E. coli*)		17.3 ± 4.2	7.0 ± 1.9
CXCR4^{QTY} – SZ158a (*E. coli*)		246.9 ± 62.2	~200
CXCR4^{QTY} – SZ146a (*E. coli*)		301.2 ± 52.2	Not calculatable
CCR5 native	~4		
CCR5^{QTY} (*E. coli*)	6.8 ± 2.0		
CCR5^{QTY} – SZ218a (*E. coli*)	87.7 ± 19.5		
CCR5^{QTY} – SZ190b (*E. coli*)	55.8 ± 8.0		
CCR5^{QTY} (SF9)	41.1 ± 16.8		
CCR5^{QTY} – SZ218a (SF9)	51.7 ± 19.0		
CCR5^{QTY} – SZ190b (SF9)	37.8 ± 11.1		

Table 1. Ligand Binding of Non-full-Length Chemokine Receptors CXCR4^{QTY} and CCR5^{QTY}

[a] CCL5 is also called "Rantes," and CXCL12 is also called "SDF1α" in the literature.

[b] The K_d ~200 nM was measured by a cell-based assay.

manufacturer-provided K_d model, as presented in the Supplemental Information and Transparent Methods.

The ligand binding for nfCCR5^{QTY} receptors expressed in both SF9 cells and *E. coli* were measured (Figures 3A and 3B). Proteins produced from both host systems retain their respective ligand affinity toward CCL5. The ligand-binding measurements were reproducible over several different expressions and purifications. The affinity values obtained for receptors produced in both systems are consistent with each other, with minor variations (Table 1). Full-length CCR5^{QTY} purified from *E. coli* for CCL5 has a higher affinity than CCR5^{QTY} purified from SF9. This is probably due to the enhanced protein stability in an aqueous environment from additional QTY modification in IC loops and C terminus. On the other hand, nfCCR5^{QTY} receptors purified from SF9 exhibit slightly better affinity compared with counterparts purified from *E. coli* in spite of having the same sequence. This might be attributed to the refolding process of receptors purified from *E. coli* where some proteins can misfold into non-functional soluble aggregates and negate the average affinity of the overall system.

The nfCXCR4^{QTY} receptors SZ158a and SZ146a expressed and purified from *E. coli*. were evaluated for CXCL12 ligand binding (Figure 3C). Both non-full-length receptors exhibit 15×–18× decrease in ligand affinity when compared with the full-length CXCR4^{QTY}. The affinity value for CXCR4^{QTY}-SZ158a is similar to that of CXCR4^{QTY}-SZ146a but slightly better. Two proteins differ only in TM1 and TM6 in their protein sequence. It is possible that the higher affinity from CXCR4^{QTY}-SZ158a benefits from the longer α-helical chain between the N terminus and EC3, which increases the adaptability of the receptor structure. Ling et al. showed that EC2 and EC3 are very important but EC1 is not crucial in the HEK293 cell signaling. Our results are consistent with their findings.

In addition, nfCXCR4^{QTY} receptors were tested against HIV1 coat glycoprotein gp$_{41\text{-}120}$ (Figure 3D). Both non-full-length receptors show hints of binding with drastically decreased affinity compared with full-length CXCR4^{QTY}. Our results suggest the essential role that the N terminus and EC3 play in HIV entry into cells.

Secondary Structure Analysis of nfCXCR4^{QTY} and nfCCR5^{QTY} Receptors

The secondary structures of nfCXCR4^{QTY} and nfCCR5^{QTY} receptors were analyzed using circular dichroism (CD) and are presented in Figure S7. Both full-length and non-full-length receptors exhibit a predominantly α-helical spectrum, with characteristic minima located at ~208 and ~222 nm. Considering that both native and QTY chemokine receptor variants contain a large portion of α-helices, the results suggest that these

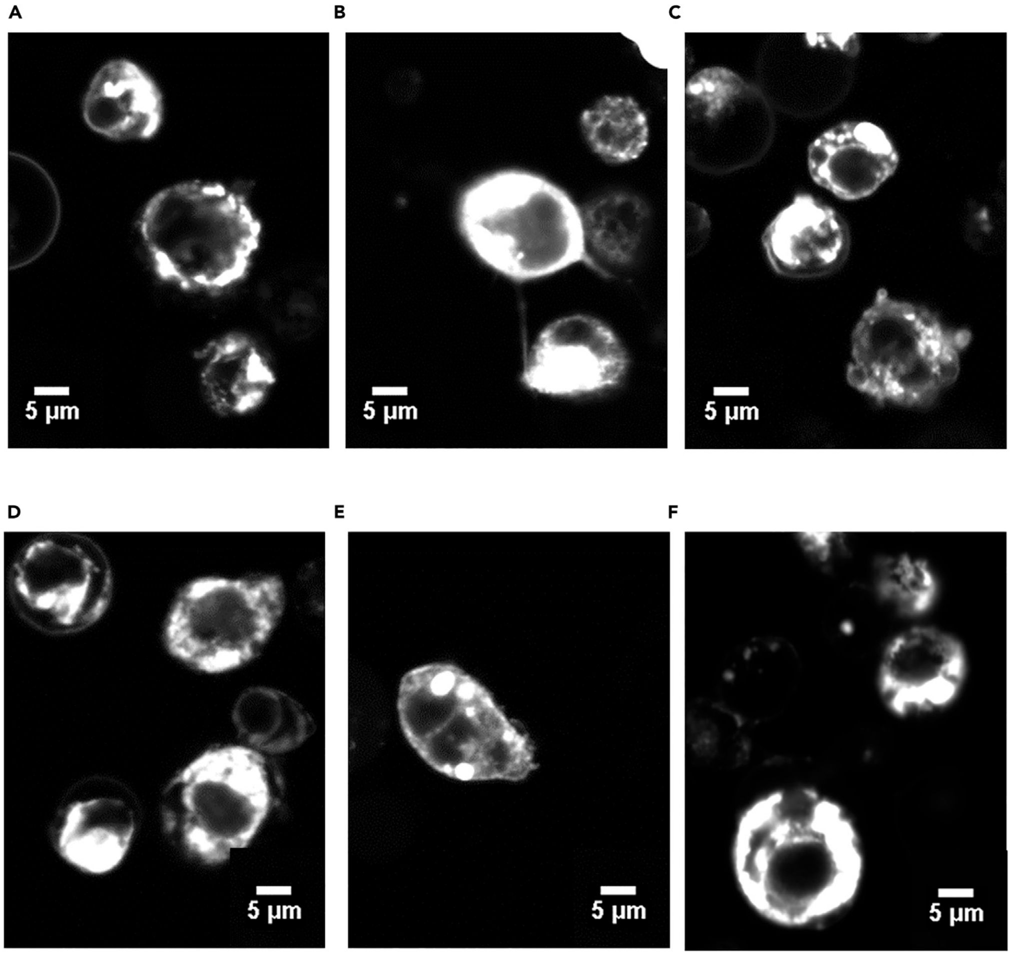


Figure 4. Confocal Images of Native Full-Length and Reconverted Non-QTY CXCR4 and CCR5 Truncated Receptors

Tagged with GFP and expressed in HEK293T cell. All receptors exhibited preferential localization on cell membranes. (A–F) (A) CCR5 full-length, (B) CCR5-SZ190b, (C) CCR5 full-length and CCR5-SZ190b co-transfection, (D) CXCR4 full-length, (E) CXCR4-SZ158a, and (F) CXCR4 full-length and CXCR4-SZ158a co-transfection. Scale bars: 5 µm.
See also Figure S8.

proteins are likely to be folded properly. The CD spectra of full-length QTY receptors correspond well with our previous reports. The non-full-length receptors show slightly different spectra, indicating difference in receptors' inter-helical interactions.

Reconverting nfCXCR4^{QTY} and nfCCR5^{QTY} Receptors to Non-QTY Variants for HEK293T Expression

Having verified the ligand activity of nfCXCR4^{QTY} and nfCCR5^{QTY} receptors in solution, we then asked if such truncations might exist and function in an actual human cell line. Truncated QTY receptors with higher ligand affinities, CCR5^{QTY}-SZ190b and CXCR4^{QTY}-SZ158a, were reconverted to the non-QTY form for gene synthesis and HEK293T expression. Sequences for reconverted non-full-length receptors were identified by extracting DNA sequences corresponding to the truncated protein sequences from GeneBank entry (CXCR4: NM_003467.3; CCR5: NM_000579.3). Full-length CXCR4 and CCR5 genes were directly purchased.

Both full-length and truncated receptors were fused with GFP (green fluorescent protein) on their C terminus and transfected into HEK293T human cells under a human cytomegalovirus promoter (hCMV). The expression and localization of these receptors were visualized using confocal microscopy, as shown in

Figures 4A–4F. Cells transfected with full-length and truncated receptors all show enhanced fluorescence at 525 nm on cell membranes using 488-nm laser excitation. Despite a large deletion of sequences, the non-QTY versions of the nfCXCR4 and nfCCR5 still preferentially localize on the cell membranes, due to the existence of residual hydrophobic TM regions. Co-transfection of both non-full-length and full-length receptors shows similar surface distribution. The localization potentially still enables these truncated receptors to function like membrane proteins. Our observation is consistent with previous report that truncated CXCR4 and CCR5 could insert into the cell membrane (Ling et al., 1999).

TM segments and topology of non-QTY truncated receptors were predicted via TMHMM Server v. 2.0, as shown in Figure S8. The predicted TM regions correspond well with the sequences within TM in original full-length receptors. CXCR4-SZ158a has a predicted intracellular N terminus, whereas the prediction of CCR5-SZ190b agrees with a schematic we have proposed in Figure 6A. Actual topology of these receptors will need to be verified through detailed structural studies.

Cell Signaling of Full-Length and Truncated CXCR4 and CCR5 Receptors

The biological functionality of CXCR4-SZ158a and CCR5-SZ190b in HEK293T cell were determined by carrying out ligand-induced signaling assays using a calcium indicator and comparing with full-length receptors. Plasmids encoding (1) full-length, (2) truncated, or (3) both of the genes were co-transfected with $G_{\alpha q}$ into HEK293T cells. The cells were then stained using calcium-sensitive dye Rhod-4 with excitation/emission wavelength at 540/590 nm and monitored by a plate reader upon application of respective ligands. Free calcium interacts with the fluorophore to cause intensity increase in the well when signaling is triggered by the full-length or truncated receptors upon the addition of the ligands. Amplitude of the fluorescent changes correlates to the ligand-receptor interactions and depends on $G_{\alpha q}$-coupled signaling pathway (Kufareva et al., 2014; Lorenzen et al., 2018). Two concentrations of ligands, 25nM and 100nM, were applied to the wells to determine concentration dependence. The calcium fluorescence change was recorded as a function of time (Figures 5A–5F). Three independent biological repeats of each group were conducted to eliminate error and obtain statistical significance.

Figures 5A–5C show the fluorescence response from CCR5, CCR5-SZ190b, and mix of the two receptors when CCL5 is added. Ligand at two concentrations was added at the time point corresponding to the first dashed line. After 180 s of data acquisition, at the second dashed line, the plate was allowed to rest for an additional 180 s and resume data collection for another 10 min until the fluorescence intensity fully recovered to baseline level. Full-length CCR5 exhibits the most significant fluorescence change at both of the ligand concentrations (Figure 5A). In ideal conditions, the time-dependent fluorescent change should present a bell-like shape (Caers et al., 2014). The different trends in Figure 5A (100 nM) and Figure 5A (25 nM) resembles the different stages in an ideal response model and are likely due to the ligand diffusion at different concentrations, which can be the rate-limiting factor.

Despite a large deletion of sequences, CCR5-SZ190b also processes a discernable ligand-induced calcium signaling at 100 nM CCL5 concentration, albeit with much lower calcium response intensity. The signaling is not triggered at 25 nM ligand concentration, as shown in Figure 5B, where no fluorescence change was observed. When CCR5 and CCR5-SZ190b are co-expressed, the fluorescence changing profile at 100 nM CCL5 strongly resembles that of CCR5 at 25 nM CCL5, showing an increasing trend on fluorescence, but with an intensity similar to that of the full-length CCR5. The hysteresis effect might suggest that, when truncated CCR5 co-exists on the cell membrane with CCR5, it can act like a ligand sink to negatively regulate the binding event between CCR5 and CCL5, prolonging the reaction without diminishing it.

Figures 5D–5F show the calcium fluorescence response from CXCR4, CXCR4-SZ158a, and a mix of the two receptors when CXCL12 is added. No discernable signaling response is observed except for CXCR4 with 100 nM CXCL12 added. Such signaling is again negated when CXCR4 is co-expressed with CXCR4-SZ158a. There are dubious fluorescence increases for CXCR4-SZ158a and co-expressed cultures when 25 nM CXCL12 is added, but data is not significant enough to draw a conclusion.

Taken together, this set of data shows that certain truncated receptors, such as CCR5-SZ190b, can carry out limited signaling function at high ligand concentrations when individually expressed. However, both the truncated receptors preferentially behave like a ligand sink and negatively regulate binding between ligand and full-length receptors when co-expressed. Such signal-regulating effects were commonly

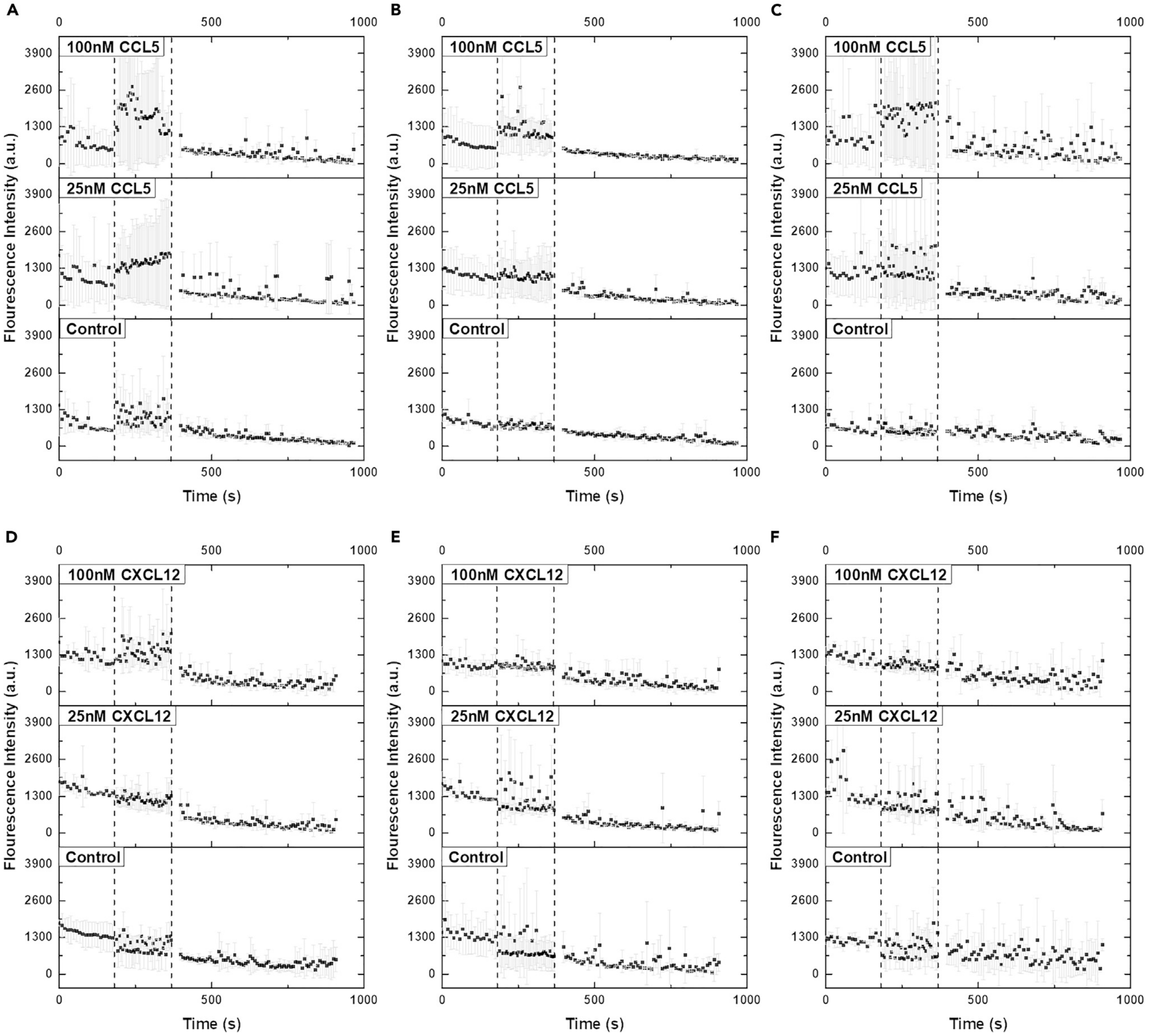


Figure 5. Ligand-Induced Calcium Signaling Assays of Native Full-Length and Reconverted Non-QTY CXCR4 and CCR5 Truncated Receptors Co-transfected with $G_{\alpha q}$ in HEK293T Cells

(A–F) (A) CCR5 full-length, (B) CCR5-SZ190b, (C) CCR5 full-length and CCR5-SZ190b co-transfection, (D) CXCR4 full-length, (E) CXCR4-SZ158a, and (F) CXCR4 full-length and CXCR4-SZ158a co-transfection. Signaling was monitored by a calcium-sensitive dye Rhod-4. Fluorescence readings were recorded as a function of time. Two concentrations of respective ligand, 100nM (upper panel) and 25nM (middle panel), are added to the stained cells and compared with a negative control sample (lower panel). Two events were marked with dashed lines in the graph, representing the addition of ligand (first dashed line) and resting of plate (second dashed line), respectively. The fluorescence intensities were averaged from three independent experiments and normalized to baseline for illustration purposes. Error bars were shown for individual data points.

observed on soluble single transmembrane cytokine receptors (Heaney and Golde, 1998) and human-engineered "decoy receptors" (Kariolis et al., 2014; Wise, 2012). The signaling assays infer that, if these non-full-length receptors are indeed present in living organisms, they can provide an additional level of regulatory functions by affecting the binding and signaling of the full-length receptors.

Schematic Representations for nfCXCR4^{QTY} and nfCCR5^{QTY} Receptors

A schematic representation of how nfCXCR4^{QTY} and nfCCR5^{QTY} receptors interact with their respective ligands is shown in Figure 6A. Tamamis et al. simulated the ligand interaction of native CXCR4 and CCR5

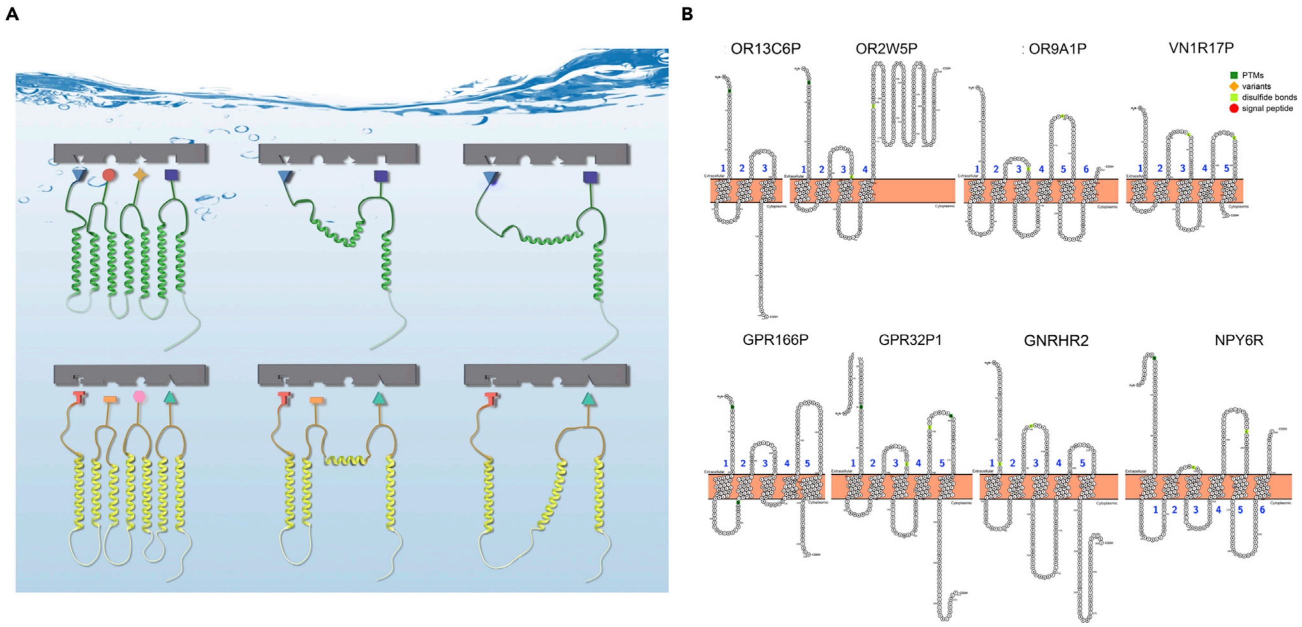


Figure 6. Schematic Illustration of and Data Mining of Non-Full-Length Receptors

(A) Schematic illustration of possible ligand interaction of non-full-length CXCR4 and CCR5 receptors. The ligand-binding motif in N terminus and 3 EC loops are simplified and represented with cartoon blocks. Y2H screen indicated EC3 to be an essential part for CXCL12 binding by nfCXCR4^{QTY} receptors SZ146a and SZ158a. The inter-connect coil between the N terminus and EC3 only slightly reduced the ligand affinity. For nfCCR5^{QTY}, even though SZ218a contains ligand-binding motif EC2 loop, the inter-coil in between EC1 and EC3 may cause undesired spacing between the functional sites and rendered a slightly reduced affinity compared with EC3 containing SZ190b. In these cases, EC3 loop is required for both nfCXCR4^{QTY} and nfCCR5^{QTY} ligand binding as first identified in Y2H *in vivo* selections.

(B) Truncated or mutated GPCRs without full 7TM. Eight truncated GPCRs are mined from the genome database that include three olfactory receptors and one vomeronasal receptor, GNRHR2 (gonadotropin releasing hormone receptor 2), NPY6R (putative neuropeptide Y receptor type 6), and putative GPCRs of unknown function. Common GPCRs have 7TM, but these truncated GPCRs with various deletions have 3TM, 4TM, 5TM, and 6TM. They are presumed to be non-functional. However, no experiments have been carried out to test their biological function. It is plausible that some of them may be still able to bind their respective ligands and carry out signaling in cells.

See also Figure S9.

proteins (Tamamis and Floudas, 2014a, b). The primary interaction between receptors' EC components with ligands were suggested with key residues referenced. We use graphical illustrations to represent the motifs on N terminus and EC loops that are responsible for ligand binding. In spite of a large deletion of a primary sequence, many of the key motifs are still present in the non-full-length variant chemokine receptors, rendering ligand affinity. Both non-full-length CXCR4^{QTY}-SZ158a and SZ146a are similar in retention of their EC components, thus they have close affinity values. Interestingly, CCR5^{QTY}-SZ190b shows slightly higher ligand affinity compared with CCR5^{QTY}-SZ218a, with one less EC loop. One possible explanation could be the orientation of TM3+TM5. The rigidity of the non-parallel α-helix may prevent EC3 from forming a proper binding pocket with N terminus and EC1 against its ligand.

Truncation of the receptors also affects the signaling capability for the non-QTY version of the receptors. For CXCR4-SZ158a, it is likely that there is only one TM segment, and the lack of any intracellular loop would be insufficient to carry out signaling activity. For CCR5-SZ190b, there may likely be three TM α-helices and one intracellular loop (IC1) that might still be capable of performing reduced signaling albeit with significant deletion. Such a hypothesis will be verified in our further research.

Data Mining for Native Non-Full-length GPCR

In light of our finding on non-full-length chemokine receptors with significant deletions of residues, we carried out a data mining search for native GPCRs with alternative splicing or frameshift mutations that resulted in truncations (Figure 6B). Eight truncated GPCRs are mined from the genome databases that include three olfactory receptors, one vomeronasal receptor, GNRHR2 (gonadotropin-releasing hormone receptor 2), NPY6R (putative neuropeptide Y receptor type 6), and two putative GPCRs of unknown function. Common GPCRs have 7TM domains, but these truncated GPCRs with various deletions have 3TM, 4TM, 5TM, and 6TM domains. They are presumed to be non-functional and thus neglected directly. No systematic efforts have been made to study their biological function. It is plausible that some of them may be still able to bind their respective ligands and perform certain functions in cells. Systematic experiments will be needed to test and verify if some may be still functional and retain biological relevance.

DISCUSSION

The discovery of truncated chemokine receptors nfCCR5^{QTY} and nfCXCR4^{QTY} stimulated us to ask new questions about the deterministic factors for native GPCR functionality: if similar non-full-length receptors exist in humans and how they interact with the full-length receptors and if they have some regulatory activities. It has been suggested that certain truncated receptors can form dimers or oligomers to hinder the transport of full-length receptors to the cell surface (Wise, 2012).

Elucidating the Essential Components for Binding Events

Numerous efforts were devoted to identifying the key residues responsible for ligand binding and HIV infection for CXCR4 and CCR5 receptors. Researchers used either mutation-based methods (Abrol et al., 2014; Brelot et al., 2000; Choi et al., 2005; Howard et al., 1999; Lopalco, 2010; Wescott et al., 2016) or computer simulations (Abrol et al., 2014; Tamamis and Floudas, 2014a, b) to reveal specific amino acids without which the activity of the proteins will be severely hindered. Our approach is complementary with a mutation-based analysis and cross-referenced with computer simulations. Proteins are analyzed by fractions through which the essential components can be identified. For instance, EC3 of CXCR4 is identified as a key component for CXCL12 binding as only receptors with EC3 are observed with gene activation in Y2H assays. A simple analogy is shown in Figure S9 where not all five fingers are necessary to hold a teacup.

Implications and Future Studies of Truncated Membrane Receptors

Our observation of non-full-length functional CCR5 and CXCR4 variants raises more questions than it provides answers. Some questions are as follows. (1) Are there DNA sequences specifically coding for non-full-length receptors in all genomes? (2) Are they capable of performing regulatory functions *in vivo* at another level? (3) What are the smallest functional receptors that can exist *in vivo*? (4) Are they synthesized and subsequently cleared?

It is plausible that there are a few means of generating non-full-length receptors and proteins in general through (1) alternative RNA splicing (Ambros, 2004; Chaudhary et al., 2019), (2) SINE and LINE transposon insertions and deletions (Adams et al., 1980; Cordaux and Batzer, 2009; Deininger et al., 1981; Ewing and Kazazian, 2011; Singer, 1982; Vassetzky and Kramerov, 2013; Wicker et al., 2007), (3) frameshift mutations resulting in premature translational termination, and (4) non-AUG translation initiation (Ghosh et al., 1967; Kearse and Wilusz, 2017). Many gene identification bioinformatics search for receptors and proteins with AUG as the translational initiation, and most experiments probe for RNA, rather than proteins. Therefore, it is plausible that such non-full-length proteins may have been overlooked.

Suggested Systematic Experimental Studies

To identify and study non-full-length receptors, or more generally, non-full-length proteins, alternative methods are required to find, experimentally characterize, and finally understand their possible biological functions. These methods include (1) performing RNA sequencing using long-read sequencing technology to seek corresponding transcripts and identify non-AUG initial codons, (2) isolating proteins from 1D and 2D membrane protein-specific gels (Carrette et al., 2006; Santucci et al., 2015; Westermeier, 2014) combined with mass spectroscopy identifications, and (3) generating specific monoclonal antibodies (mAbs) for particular regions of known proteins as probes to find non-full-length proteins in various cellular regions and all tissues of every cell cycle as function of time, for example, generation of mAbs (Hashimoto et al., 2018; Huang et al., 2016) for membrane receptors of every intracellular and extracellular loop and N and C termini. The final method is (4) isolating proteins from 1D and 2D protein gels to carry out single protein molecule sequencing using the latest aerolysin nanopore method (Ouldali et al., 2020).

Think Differently and Ask Unusual Questions

Before microRNAs were unexpectedly discovered (Lee et al., 1993; Wightman et al., 1993), such small RNA species and other noncoding RNAs were also overlooked. Since then, microRNAs have been found to be indispensable in every aspect for biological regulations, especially for highly evolved biological systems. Recently, a number of mini-proteins or micro-proteins, previously identified as peptides or small open reading frames, have been found to play a very important role in all aspects of biological regulation (Anderson et al., 2015; Bhati et al., 2018; Camarero, 2017; Carvunis et al., 2012; D'Lima et al., 2017; Delcourt et al., 2018; Graeff et al., 2016; Ingolia et al., 2011; Kageyama et al., 2011; Orr et al., 2019; Saghatelianr and Couso, 2015; Singh et al., 2019; Staudt and Wenkel, 2011). Our unexpected discovery of truncated membrane receptor variants in this study may thus alert us again to venture beyond current paradigms to discover, characterize, and design proteins.

Limitation of the Study

Here we discovered non-full-length CCR5 and CXCR4 chemokine receptors that still retain ligand affinity and partial signal transduction capability. As this study is based on the convenient QTY design code and Y2H screening system *in vitro*, the truncation of the full-length receptors *in vivo* resulted from alternative splicing or pseudo-genes might differ in exact sequence. Herein, as suggested in the Discussion section, a systematic study of the human genome is required to identify actual truncated genes that serve the function. Details of experiments are proposed in the Discussion section. Nevertheless, our work provides valuable insights on the structure-function relation of non-full-length receptors and their possible regulatory functions in living organism.

On the other hand, due to the ongoing pandemic of COVID-19 and institution closedown, we were only able to access a regular Tecan Spark microplate reader for the calcium signaling experiments. The unit lacks an auto-injection module, which is essential to continuously monitor ligand-induced fluorescence change. A hardware-enforced 5-s delay is estimated between the addition of the ligand and the start of measurement. As the ligand-induced calcium signaling is a transient process, such delay negates our ability to evaluate the whole period of ligand-receptor interaction. We expect that a plate reader with auto-injection and shaking module would be able to fully elucidate how non-full-length receptor responds to ligand under different condition. In the current study, we introduced ligand diffusion factor by not pipetting or shaking the plate to prolong the response time. We also conducted three independent biological repeats to average out the intensity fluctuations to establish statistical significance of experiments. Albeit not in ideal conditions, we consider our current approach illustrative enough to reveal the potential functions of these non-full-length receptors *in vivo*.

Resource Availability

Lead Contact

Further information and requests for materials should be directed to lead contact, Rui Qing, Ruiqing@mit.edu.

Materials Availability

All new unique genes generated in this study are available from the lead contact with a completed materials transfer agreement. Genes for native and QTY version of non-full-length CXCR4 and CCR5 chemokine receptors will also be deposited on Addgene for research use.

Data and Code Availability

This published article includes all datasets generated or analyzed during this study.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101670.

ACKNOWLEDGMENTS

This work was primarily funded by OH2 Laboratories and the MIT Center for Bits and Atoms Consortium that includes Bay Valley Innovation Center (Shanghai). We also thank Dorrie Langsley for carefully helping English editing.

AUTHOR CONTRIBUTIONS

R.Q., F.T., P.C., and S.Z. designed the experiments. R.Q., P.C., G.Y., Q.H., H.C., J.N., B.S., J.K., B.M., T.S., C.B. performed the experiments, R.Q., F.T., P.C., G.Y., B.S., and S.Z. wrote the paper.

DECLARATION OF INTERESTS

This research was in part funded by an MIT startup OH2 Laboratories. S.Z. fully discloses that he has a minority stake in this startup for the invention.

Received: April 14, 2020
Revised: August 12, 2020
Accepted: October 8, 2020
Published: October 28, 2020

REFERENCES

Abrol, R., Trzaskowski, B., Goddard, W.A., Nesterov, A., Olave, I., and Irons, C. (2014). Ligand- and mutation-induced conformational selection in the CCR5 chemokine G protein-coupled receptor. Proc. Natl. Acad. Sci. U S A *111*, 13040–13045.

Adams, J.W., Kaufman, R.E., Kretschmer, P.J., Harrison, M., and Nienhuis, A.W. (1980). A family of long reiterated DNA-sequences, one copy of which is next to the human beta-globin gene. Nucleic Acids Res. *8*, 6113–6128.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350–355.

Anderson, D.M., Anderson, K.M., Chang, C.L., Makarewich, C.A., Nelson, B.R., McAnally, J.R., Kasaragod, P., Shelton, J.M., Liou, J., Bassel-Duby, R., et al. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. Cell *160*, 595–606.

Bhati, K.K., Blaakmeer, A., Paredes, E.B., Dolde, U., Eguen, T., Hong, S.Y., Rodrigues, V., Straub, D., Sun, B., and Wenkel, S. (2018). Approaches to identify and characterize microProteins and their potential uses in biotechnology. Cell. Mol. Life Sci. *75*, 2529–2536.

Brelot, A., Heveker, N., Montes, M., and Alizon, M. (2000). Identification of residues of CXCR4 critical for human immunodeficiency virus coreceptor and chemokine receptor activities. J. Biol. Chem. *275*, 23736–23744.

Caers, J., Peymen, K., Suetens, N., Temmerman, L., Janssen, T., Schoofs, L., and Beets, I. (2014). Characterization of G Protein-coupled receptors by a fluorescence-based calcium mobilization assay. J. Vis. Exp. *28*, e51516.

Camarero, J.A. (2017). Cyclotides, a versatile ultrastable micro-protein scaffold for biotechnological applications. Bioorg. Med. Chem. Lett. *27*, 5089–5099.

Carrette, O., Burkhard, P.R., Sanchez, J.C., and Hochstrasser, D.F. (2006). State-of-the-art two-dimensional gel electrophoresis: a key tool of proteomics research. Nat. Protoc. *1*, 812–823.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., et al. (2012). Proto-genes and de novo gene birth. Nature *487*, 370–374.

Chaudhary, S., Khokhar, W., Jabre, I., Reddy, A.S.N., Byrne, L.J., Wilson, C.M., and Syed, N.H. (2019). Alternative splicing and protein diversity: plants versus animals. Front. Plant Sci. *10*, 708.

Choi, W.T., Tian, S.M., Dong, C.Z., Kumar, S., Liu, D.X., Madani, N., An, J., Sodroski, J.G., and Huang, Z.W. (2005). Unique ligand binding sites on CXCR4 probed by a chemical biology approach: implications for the design of selective human immunodeficiency virus type 1 inhibitors. J. Virol. *79*, 15398–15404.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703.

Cordoba-Chacon, J., Gahete, M.D., Duran-Prado, M., Pozo-Salas, A.I., Malagon, M.M., Gracia-Navarro, F., Kineman, R.D., Luque, R.M., and Castano, J.P. (2010). Identification and characterization of new functional truncated variants of somatostatin receptor subtype 5 in rodents. Cell. Mol. Life Sci. *67*, 1147–1163.

D'Lima, N.G., Ma, J., Winkler, L., Chu, Q., Loh, K.H., Corpuz, E.O., Budnik, B.A., Lykke-Andersen, J., Saghatelian, A., and Slavoff, S.A. (2017). A human microprotein that interacts with the mRNA decapping complex. Nat. Chem. Biol. *13*, 174–180.

Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T., and Schmid, C.W. (1981). Base sequence studies of 300 nucleotide renatured repeated human DNA clones. J. Mol. Biol. *151*, 17–33.

Delcourt, V., Staskevicius, A., Salzet, M., Fournier, I., and Roucou, X. (2018). Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA. Proteomics *18*, e1700058.

Den Dunnen, J.T., and Van Ommen, G.J.B. (1999). The protein truncation test: a review. Hum. Mutat. *14*, 95–102.

Deng, H.K., Liu, R., Ellmeier, W., Choe, S., Unutmaz, D., Burkhart, M., DiMarzio, P., Marmon, S., Sutton, R.E., Hill, C.M., et al. (1996). Identification of a major co-receptor for primary isolates of HIV-1. Nature *381*, 661–666.

Duran-Prado, M., Gahete, M.D., Martinez-Fuentes, A.J., Luque, R.M., Quintero, A., Webb, S.M., Benito-Lopez, P., Leal, A., Schulz, S., Gracia-Navarro, F., et al. (2009). Identification and characterization of two novel truncated but functional isoforms of the somatostatin receptor subtype 5 differentially present in pituitary tumors. J. Clin. Endocr. Metab. *94*, 2634–2643.

Ewing, A.D., and Kazazian, H.H. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res. *21*, 985–990.

Fersht, A., and Winter, G. (1992). Protein engineering. Trends Biochem. Sci. *17*, 292–294.

Fersht, A.R. (2008). From the first protein structures to our current knowledge of protein folding: delights and scepticisms. Nat. Rev. Mol. Cell Biol. *9*, 650–654.

Ghosh, H.P., Soll, D., and Khorana, H.G. (1967). Studies on polynucleotides .67. Initiation of protein synthesis in vitro as studied by using ribopolynucleotides with repeating nucleotide sequences as messengers. J. Mol. Biol. *25*, 275–&.

Graeff, M., Straub, D., Eguen, T., Dolde, U., Rodrigues, V., Brandt, R., and Wenkel, S. (2016). MicroProtein-mediated recruitment of CONSTANS into a TOPLESS trimeric complex represses flowering in arabidopsis. PLoS Genet. *12*, e1005959.

Hashimoto, Y., Zhou, W., Hamauchi, K., Shirakura, K., Doi, T., Yagi, K., Sawasaki, T., Okada, Y., Kondoh, M., and Takeda, H. (2018). Engineered membrane protein antigens successfully induce antibodies against extracellular regions of claudin-5. Sci. Rep. *8*, 8383.

Heaney, M.L., and Golde, D.W. (1998). Soluble receptors in human disease. J. Leukoc. Biol. *64*, 135–146.

Howard, O.M.Z., Shirakawa, A.K., Turpin, J.A., Maynard, A., Tobin, G.J., Carrington, M., Oppenheim, J.J., and Dean, M. (1999). Naturally occurring CCR5 extracellular and transmembrane domain variants affect HIV-1 co-receptor and ligand binding function. J. Biol. Chem. *274*, 16228–16234.

Huang, R., Kiss, M.M., Batonick, M., Weiner, M.P., and Kay, B.K. (2016). Generating recombinant antibodies to membrane proteins through phage display. Antibodies *5*, 11.

Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802.

Kageyama, Y., Kondo, T., and Hashimoto, Y. (2011). Coding vs non-coding: translatability of short ORFs found in putative non-coding transcripts. Biochimie *93*, 1981–1986.

Kariolis, M.S., Miao, Y.R., Ii, D.S.J., Kapur, S., Mathews, I.I., Giaccia, A.J., and Cochran, J.R. (2014). An engineered Axl 'decoy receptor' effectively silences the Gas6-Axl signaling axis. Nat. Chem. Biol. *10*, 977–983.

Karpa, K.D., Lin, R.W., Kabbani, N., and Levenson, R. (2000). The dopamine D3 receptor interacts with itself and the truncated D3 splice variant D3nf: D3-D3nf interaction causes mislocalization of D3 receptors. Mol. Pharmacol. *58*, 677–683.

Kearse, M.G., and Wilusz, J.E. (2017). Non-AUG translation: a new start for protein synthesis in eukaryotes. Gene Dev. *31*, 1717–1731.

Kufareva, I., Stephens, B.S., Holden, L.G., Qin, L., Zhao, C.X., Kawamura, T., Abagyan, R., and Handel, T.M. (2014). Stoichiometry and geometry of the CXC chemokine receptor 4 complex with CXC ligand 12: molecular modeling and experimental validation. Proc. Natl. Acad. Sci. U S A *111*, E5363–E5372.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. Cell *75*, 843–854.

Lin, S.H., and Guidotti, G. (2009). Purification of membrane proteins. Method Enzymol. *463*, 619–629.

Ling, K., Wang, P., Zhao, J., Wu, Y.L., Cheng, Z.J., Wu, G.X., Hu, W., Ma, L., and Pei, G. (1999). Five-transmembrane domains appear sufficient for a G protein-coupled receptor: functional five-transmembrane domain chemokine receptors. Proc. Natl. Acad. Sci. U S A *96*, 7922–7927.

Lopalco, L. (2010). CCR5: from natural resistance to a new anti-HIV strategy. Viruses *2*, 574–600.

Lorenzen, E., Ceraudo, E., Berchiche, Y.A., Rico, C.A., Fürstenberg, A., Sakmar, T.P., and Huber, T. (2018). G protein subtype–specific signaling bias in a series of CCR5 chemokine analogs. Sci. Signal. *11*, eaao6152.

Lv, X.C., Liu, J.L., Shi, Q.Y., Tan, Q.W., Wu, D., Skinner, J.J., Walker, A.L., Zhao, L.X., Gu, X.X., Chen, N., et al. (2016). In vitro expression and analysis of the 826 human G protein-coupled receptors. Protein Cell *7*, 325–337.

Majumdar, S., Grinnell, S., Le Rouzic, V., Burgman, M., Polikar, L., Ansonoff, M., Pintar, J., Pan, Y.X., and Pasternak, G.W. (2011). Truncated G protein-coupled mu opioid receptor MOR-1 splice variants are targets for highly potent opioid analgesics lacking side effects. Proc. Natl. Acad. Sci. U S A *108*, 19778–19783.

Middlemas, D.S., Lindberg, R.A., and Hunter, T. (1991). Trkb, a neural receptor protein-tyrosine kinase - evidence for a full-length and 2 truncated receptors. Mol. Cell Biol. *11*, 143–153.

Orr, M.W., Mao, Y., Storz, G., and Qian, S.B. (2019). Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res. *48*, 1029–1042.

Ouldali, H., Sarthak, K., Ensslen, T., Piguet, F., Manivet, P., Pelta, J., Behrends, J.C., Aksimentiev, A., and Oukhaled, A. (2020). Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. Nat. Biotechnol. *38*, 176–181.

Perron, A.L., Sarret, P., Gendron, L., Stroh, T., and Beaudet, A. (2005). Identification and functional characterization of a 5-transmembrane domain variant isoform of the NTS2 neurotensin receptor in rat central nervous system. J. Biol. Chem. *280*, 10219–10227.

Qing, R., Han, Q., Skuhersky, M., Chung, H., Badr, M., Schubert, T., and Zhang, S. (2019). QTY code designed thermostable and water-soluble chimeric chemokine receptors with tunable ligand affinity. Proc. Natl. Acad. Sci. U S A *116*, 25668–25676.

Saghatelianr, A., and Couso, J.P. (2015). Discovery and characterization of smORF-encoded bioactive polypeptides. Nat. Chem. Biol. *11*, 909–916.

Samson, M., Libert, F., Doranz, B.J., Rucker, J., Liesnard, C., Farber, C.M., Saragosti, S., Lapoumeroulie, C., Cognaux, J., Forceille, C., et al. (1996). Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature *382*, 722–725.

Santucci, L., Bruschi, M., Ghiggeri, G.M., and Candiano, G. (2015). The latest advancements in proteomic two-dimensional gel electrophoresis analysis applied to biological samples. Methods Mol. Biol. *1243*, 103–125.

Seidel, S.A.I., Dijkman, P.M., Lea, W.A., van den Bogaart, G., Jerabek-Willemsen, M., Lazic, A., Joseph, J.S., Srinivasan, P., Baaske, P., Simeonov, A., et al. (2013). Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions. Methods *59*, 301–315.

Singer, M.F. (1982). Sines and lines - highly repeated short and long interspersed sequences in mammalian genomes. Cell *28*, 433–434.

Singh, D.R., Dalton, M.P., Cho, E.E., Pribadi, M.P., Zak, T.J., Seflova, J., Makarewich, C.A., Olson, E.N., and Robia, S.L. (2019). Newly discovered micropeptide regulators of SERCA form oligomers but bind to the pump as monomers. J. Mol. Biol. *431*, 4429–4443.

Skrzypek, R., Iqbal, S., and Callaghan, R. (2018). Methods of reconstitution to investigate membrane protein function. Methods *147*, 126–141.

Staudt, A.C., and Wenkel, S. (2011). Regulation of protein function by 'microProteins. EMBO Rep. *12*, 35–42.

Tamamis, P., and Floudas, C.A. (2014a). Elucidating a key anti-HIV-1 and cancer-associated Axis: the structure of CCL5 (rantes) in complex with CCR5. Sci. Rep. *4*, 5447.

Tamamis, P., and Floudas, C.A. (2014b). Elucidating a key component of cancer metastasis: CXCL12 (SDF-1 alpha) binding to CXCR4. J. Chem. Inf. Model. *54*, 1174–1188.

Vassetzky, N.S., and Kramerov, D.A. (2013). SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. *41*, D83–D89.

Vinothkumar, K.R., and Henderson, R. (2010). Structures of membrane proteins. Q. Rev. Biophys. *43*, 65–158.

Wescott, M.P., Kufareva, I., Paes, C., Goodman, J.R., Thaker, Y., Puffer, B.A., Berdougo, E., Rucker, J.B., Handel, T.M., and Doranz, B.J. (2016). Signal transmission through the CXC chemokine receptor 4 (CXCR4) transmembrane helices. Proc. Natl. Acad. Sci. U S A *113*, 9928–9933.

Westermeier, R. (2014). Looking at proteins from two dimensions: a review on five decades of 2D electrophoresis. Arch. Physiol. Biochem. *120*, 168–172.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. *8*, 973–982.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern-formation in C. elegans. Cell *75*, 855–862.

Wise, H. (2012). The roles played by highly truncated splice variants of G protein-coupled receptors. J. Mol. Signal. *7*, 13.

Zhang, S.G., Tao, F., Qing, R., Tang, H.Z., Skuhersky, M., Corin, K., Tegler, L., Wassie, A., Wassie, B., Kwon, Y., et al. (2018). QTY code enables design of detergent-free chemokine receptors that retain ligand-binding activities. Proc. Natl. Acad. Sci. U S A *115*, E8652–E8659.

Zhu, X.Y., and Wess, J. (1998). Truncated V2 vasopressin receptors as negative regulators of wild-type V2 receptor function. Biochemistry *37*, 15773–15784.

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

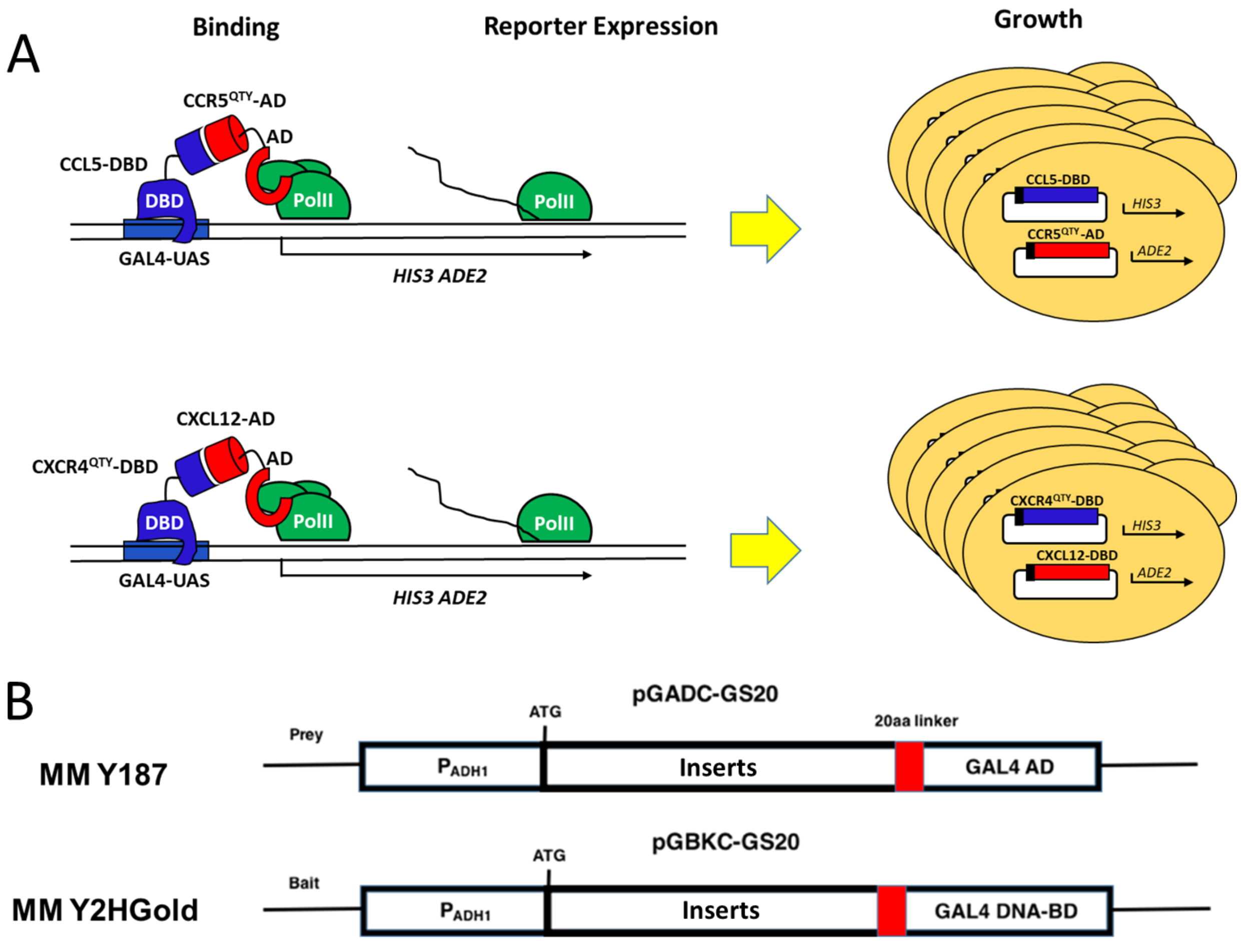
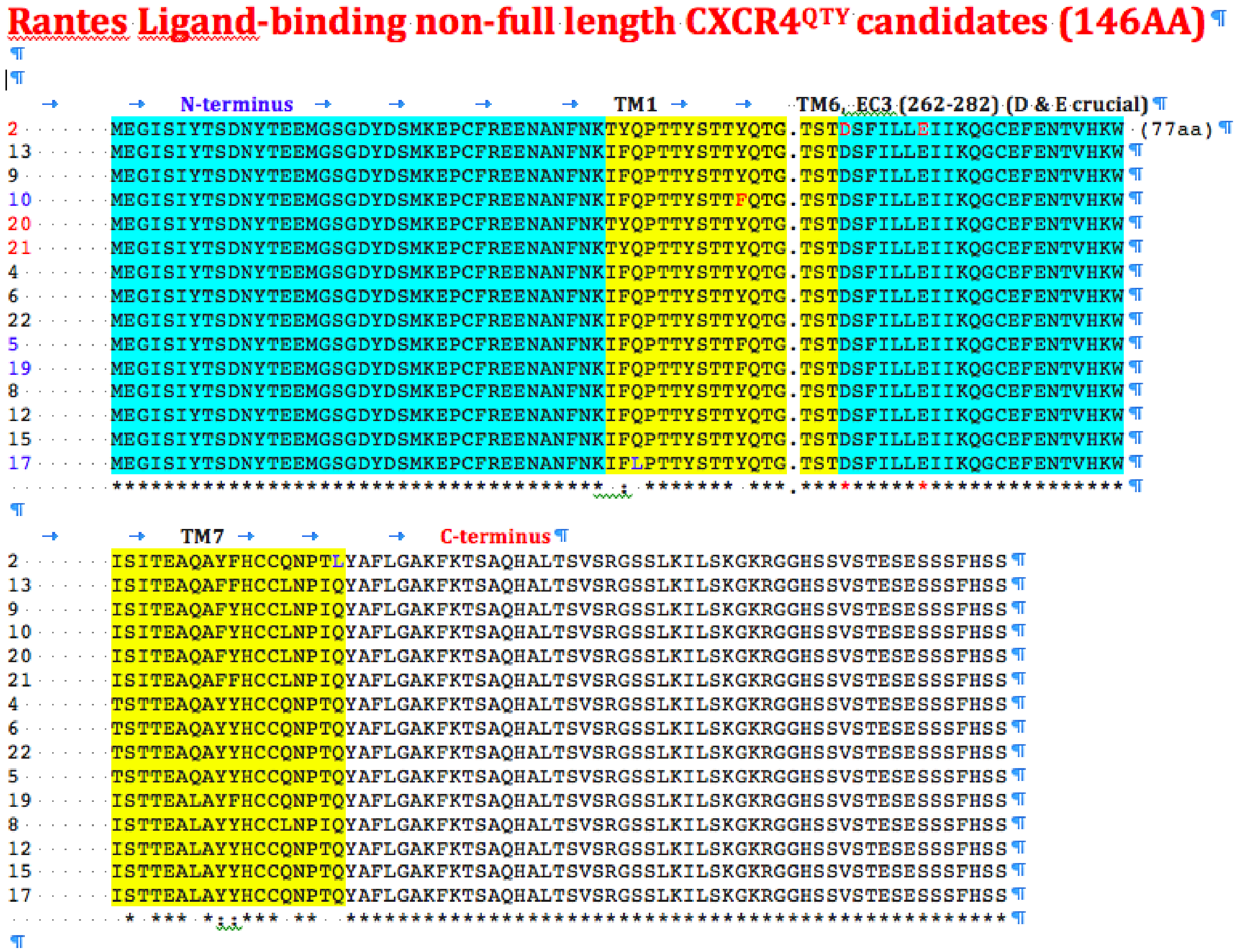


Figure S1. Schematic illustration for Y2H screening, Related to Figure 1. (A) Y2H screening for short CCR5^{QTY} and CXCR4^{QTY} variants. For CCR5^{QTY} screens, CCL5 ligand is in bait orientation in pGBKC-3C vector with CCR5^{QTY} (~3 million variants) in prey orientation in pGADC-2A. For CXCR4^{QTY}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



Rantes Ligand-binding non-full length CXCR4^{QTY} candidates (146AA)

```
                   N-terminus                                 TM1                   TM6, EC3 (262-282) (D & E crucial)
2        MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKTYQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW (77aa)
13       MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
9        MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
10       MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTFQTG.TSTDSFILLEIIKQGCEFENTVHKW
20       MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKTYQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
21       MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKTYQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
4        MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
6        MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
22       MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
5        MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTFQTG.TSTDSFILLEIIKQGCEFENTVHKW
19       MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTFQTG.TSTDSFILLEIIKQGCEFENTVHKW
8        MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
12       MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
15       MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFQPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
17       MEGISIYTSDNYTEEMGSGDYDSMKEPCFREENANFNKIFLPTTYSTTYQTG.TSTDSFILLEIIKQGCEFENTVHKW
         **************************************  : ******* *** .*************************

                   TM7                    C-terminus
2        ISITEAQAYFHCCQNPTLYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
13       ISITEAQAFFHCCLNPIQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
9        ISITEAQAFYHCCLNPIQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
10       ISITEAQAFYHCCLNPIQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
20       ISITEAQAFYHCCLNPIQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
21       ISITEAQAFFHCCLNPIQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
4        TSTTEAQAYYHCCQNPTQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
6        TSTTEAQAYYHCCQNPTQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
22       TSTTEAQAYYHCCQNPTQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
5        TSTTEAQAYYHCCQNPTQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
19       ISTTEALAYFHCCQNPTQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
8        ISTTEALAYYHCCLNPIQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
12       ISTTEALAYYHCCQNPTQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
15       ISTTEALAYYHCCQNPTQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
17       ISTTEALAYYHCCQNPTQYAFLGAKFKTSAQHALTSVSRGSSLKILSKGKRGGHSSVSTESESSSFHSS
          * *** *::*** **  ***************************************************
```

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.

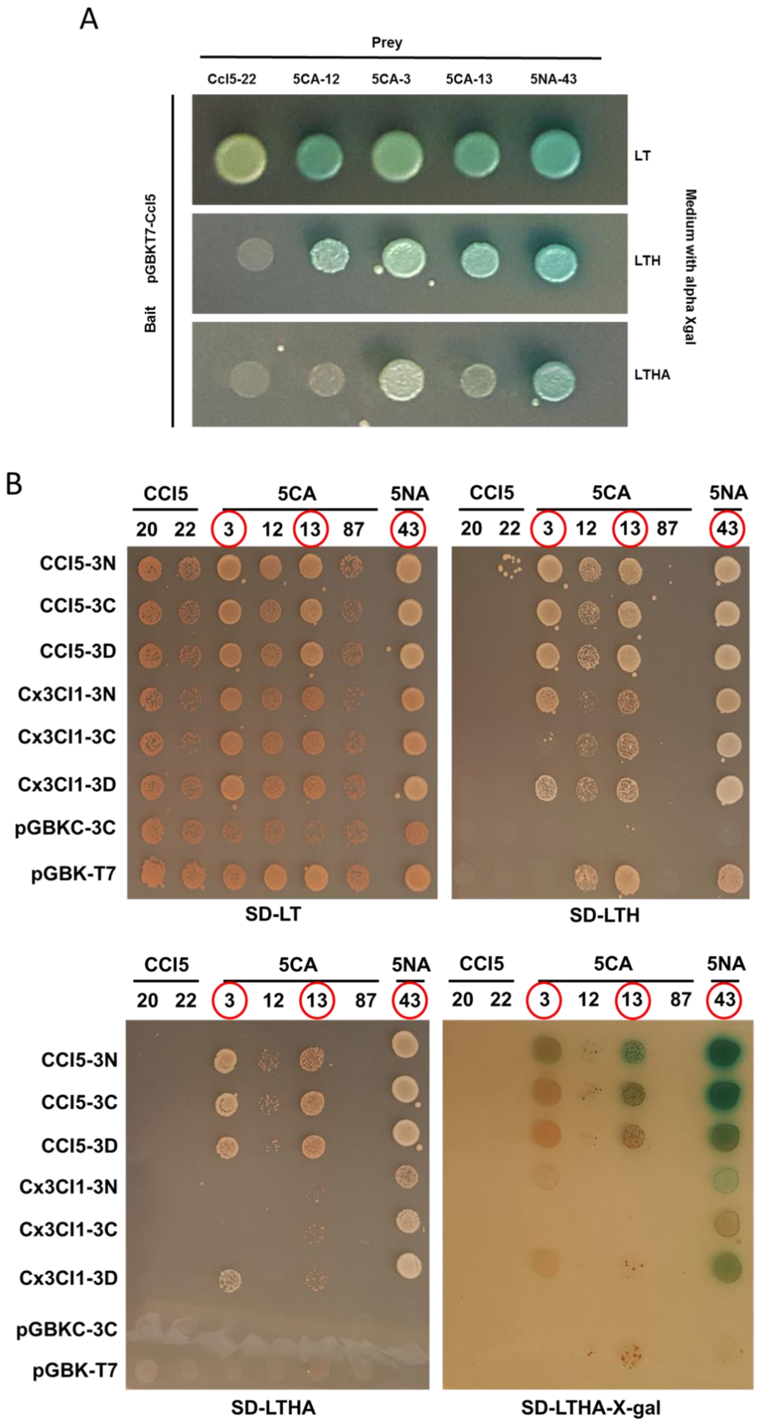
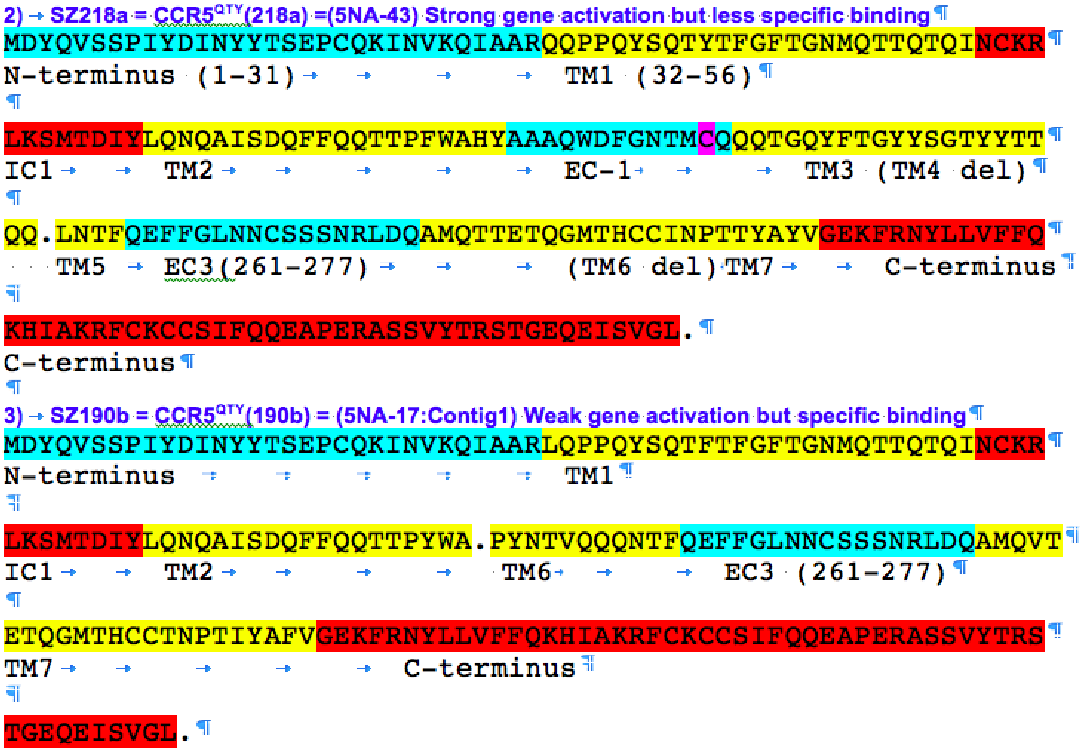


Figure S3. Y2H mating assay of 2 non-full-length CCR5^{QTY} variants, Related to Figure 1. The yeast clone 5NA43 is renamed as SZ218a to be consistent with other names. (A) Qualitative Y2H Interaction Test of CCL5 in Y2H Bait vector and CCR5^{QTY} variants CCL5-22, 5CA-12, 5CA-3, 5CA13, 5NA43 in Y2H 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).



2) SZ218a = CCR5^{QTY}(218a) =(5NA-43) Strong gene activation but less specific binding

```
MDYQVSSPIYDINYYTSEPCQKINVKQIAARQQPPQYSQTYTFGFTGNMQTTQTQINCKR
N-terminus (1-31)                TM1 (32-56)

LKSMTDIYLQNQAISDQFFQQTTPFWAHYAAAQWDFGNTMCQQQTGQYFTGYYSGTYYTT
IC1       TM2                    EC-1          TM3 (TM4 del)

QQ.LNTFQEFFGLNNCSSSNRLDQAMQTTETQGMTHCCINPTTYAYVGEKFRNYLLVFFQ
   TM5   EC3(261-277)            (TM6 del)TM7       C-terminus

KHIAKRFCKCCSIFQQEAPERASSVYTRSTGEQEISVGL.
C-terminus
```

3) SZ190b = CCR5^{QTY}(190b) = (5NA-17:Contig1) Weak gene activation but specific binding

```
MDYQVSSPIYDINYYTSEPCQKINVKQIAARLQPPQYSQTFTFGFTGNMQTTQTQINCKR
N-terminus                       TM1

LKSMTDIYLQNQAISDQFFQQTTPYWA.PYNTVQQQNTFQEFFGLNNCSSSNRLDQAMQVT
IC1       TM2                TM6             EC3 (261-277)

ETQGMTHCCTNPTIYAFVGEKFRNYLLVFFQKHIAKRFCKCCSIFQQEAPERASSVYTRS
TM7                     C-terminus

TGEQEISVGL.
```

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.

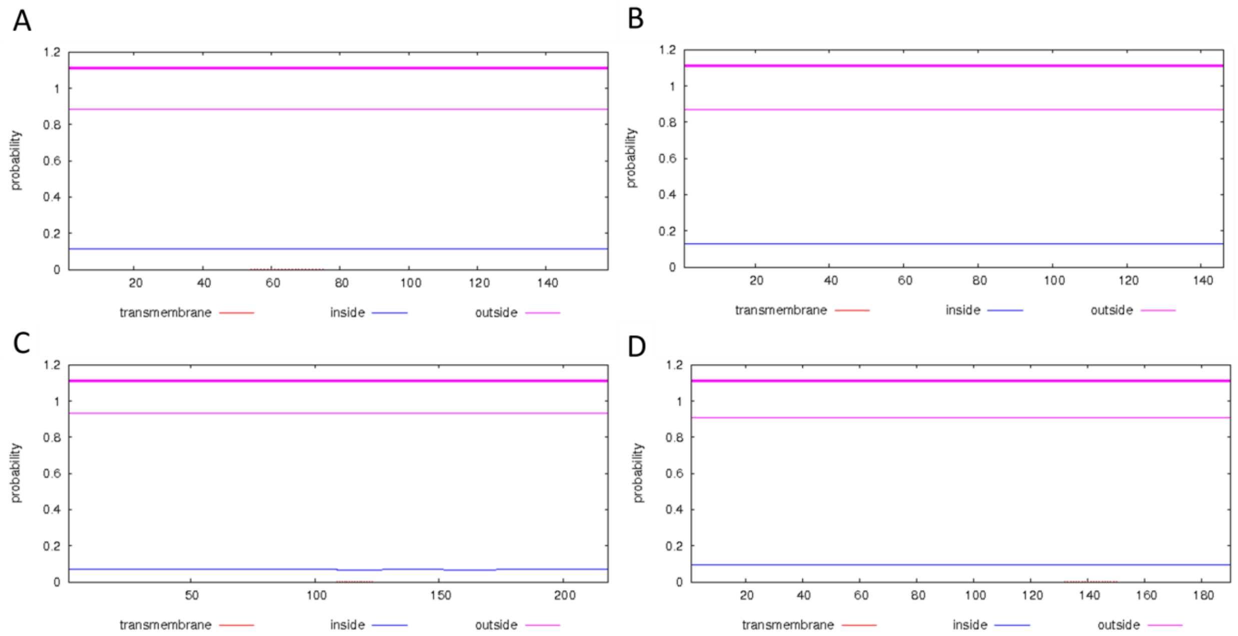


Figure S5. Bioinformatics hydrophobic segment analyses, Related to Figure 2. (A) CXCR4^{QTY}-SZ158a; (B) CXCR4^{QTY}-SZ146a; (C) CCR5^{QTY}-SZ218a; (D) CCR5^{QTY}-SZ190b**.** No hydrophobic TM region can be observed in any of the short variant GPCR^{QTY} proteins. X axis refers to the position of amino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

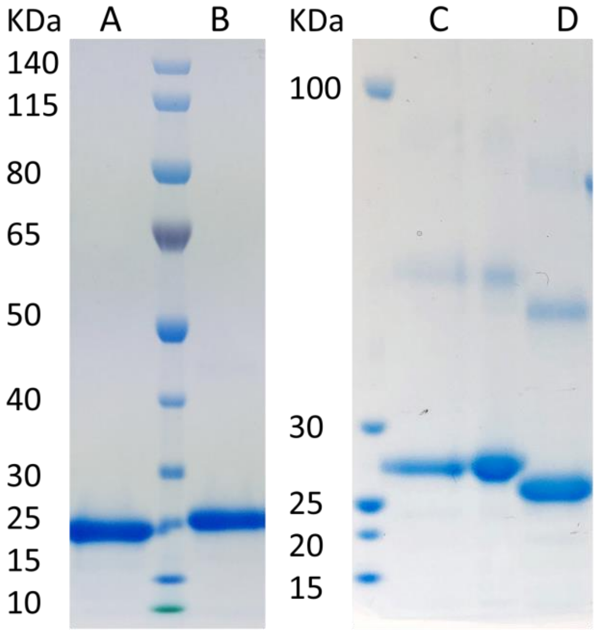


Figure S6. Electrophoresis band of purified non-full-length receptor proteins, Related to Figure 3. (A) CXCR4-SZ158a; (B) CXCR4-SZ146a; (C) CCR5-SZ218a; (D) CCR5-SZ190b. The molecular weight of the ladder is labelled on the left in KDa.

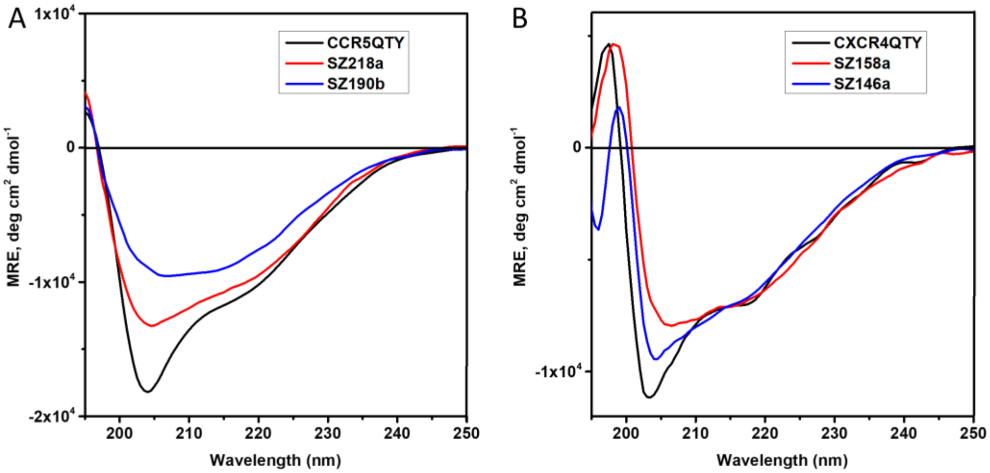


Figure S7. Secondary structure of nfCCR5^{QTY} and nfCXCR4^{QTY} receptors, Related to Figure 3. The Circular dichroism signal between 195nm and 250nm shows characteristic α-helical spectra. Difference in spectra shape is likely to be induced by inter-helix interaction in different proteins.

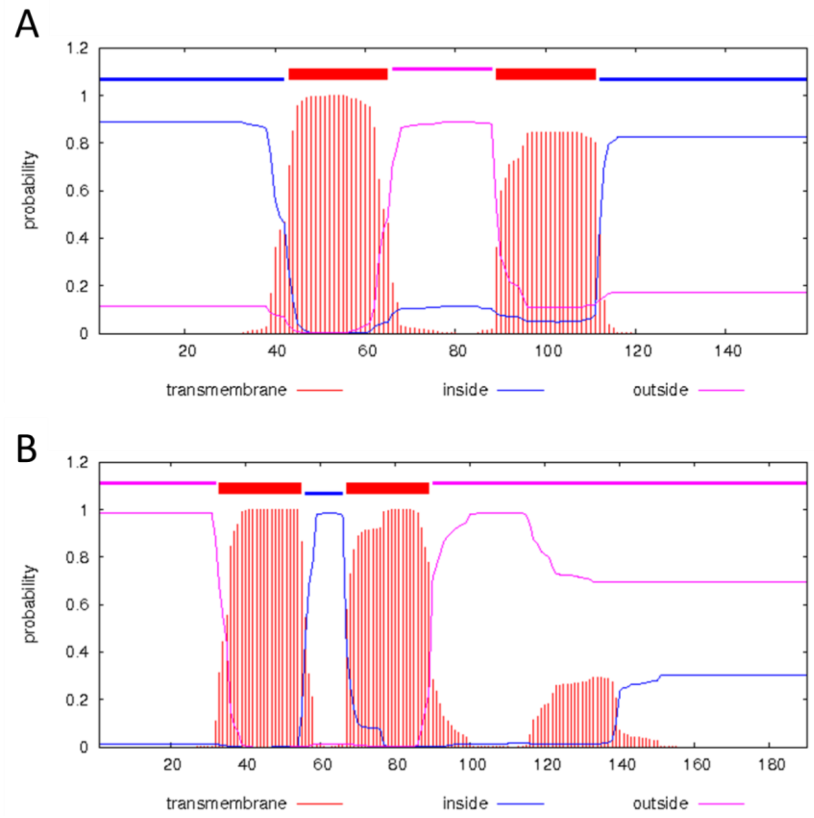


Figure S8. Bioinformatics hydrophobic segment analyses and topology prediction, Related to Figure 4. (A) CXCR4-SZ158a; (B) CCR5-SZ190b. Hydrophobic TM regions were predicted at sequences corresponding to original TM regions before truncation. The likelihood of EC and IC regions are suggested. X axis 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.

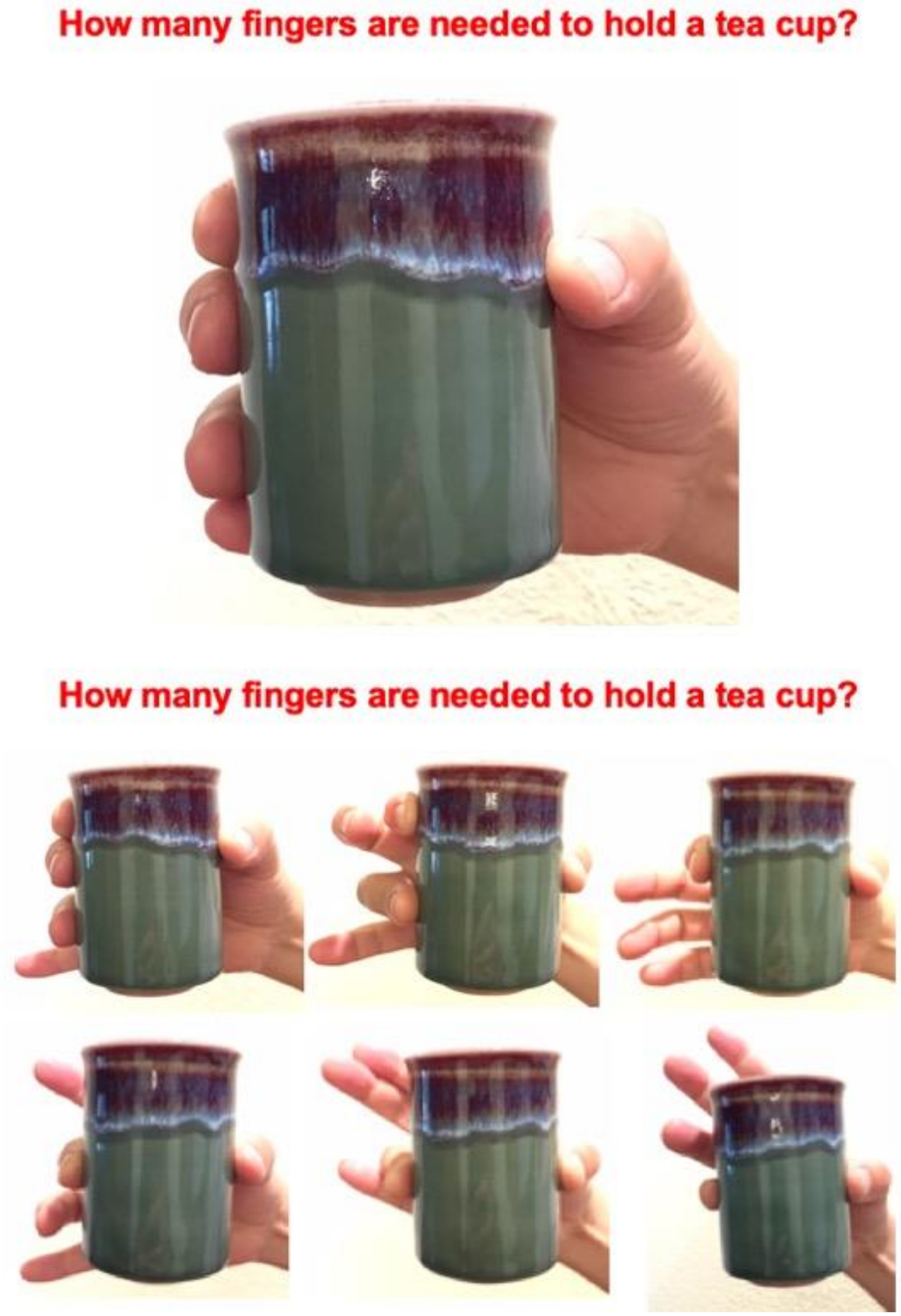

How many fingers are needed to hold a tea cup?
How many fingers are needed to hold a tea cup?

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.