Supporting Information

Schlinkmann et al. 10.1073/pnas.1202107109

SI Text

Position Nomenclature. Positions of rat neurotensin receptor 1 (rNTR1) are identified by their residue type, using the one-letter amino acid code, and by their sequential rNTR1-position (e.g., N82). The Ballesteros–Weinstein nomenclature is given as a superscript, in this example N82$^{1.50}$, with the first number (before the period) identifying the transmembrane helix (1–7), and the second number (after the period) describing the position within the helix, where 1.50 is assigned to the most conserved position of helix 1. See Ballesteros and Weinstein (1) for details. Thus, N82A$^{1.50}$ denotes a mutation from Asn to Ala in position 82, located in helix 1, very conserved, and located near the middle of the helix.

Starting Construct. The starting construct NTR1-D03 (abbreviated D03) comprises amino acids 43–424 of rNTR1 (UniProtKB release no. P20789) fused between the C terminus of maltose binding protein and the N terminus of thioredoxin to enhance expression in *Escherichia coli*. The fusion protein contains tobacco etch virus protease cleavage sites on both ends of the receptor and a C-terminal His$_{10}$ tag. Through multiple rounds of error-prone PCR followed by FACS selection, wild-type rNTR1 had acquired nine mutations (H103D$^{2.40}$, H105Y$^{2.42}$, A161V$^{3.44}$, R167L$^{3.50}$, R213L$^{4.69}$, V234L$^{5.35}$, H305R$^{6.32}$, S362A$^{7.46}$, and S417C$^{8.01}$) that enhanced expression in *E. coli* from 500 to 5,000 receptors per cell (2). This enhanced expression level was needed for robust selection, and thus for the experiments reported here.

The results of the full randomization and selection applied here are especially interesting for the D03-specific mutations, because for each position the D03 residue and the wild-type residue are compared side by side, and are in direct competition. Interestingly, Sanger and 454 sequencing results verified the selection of these mutations: R213L$^{4.69}$, V234L$^{5.35}$, and S362A$^{7.46}$ were confirmed as robust shifts, and A161V$^{3.44}$ and H105Y$^{2.42}$ as weak shifts. H305R$^{6.32}$ preferred K after correcting for codon bias, and H103D$^{2.40}$ preferred either D or N. R167L$^{3.50}$ significantly decreased the signaling capability of the receptor, because R167$^{3.50}$ is the key position of the E/DRY motif conserved throughout all G protein-coupled receptors (GPCRs). As discussed in the main text, we found only a mild preference of aliphatic residues over Arg, which allowed us to reconstitute the E/DRY motif and signaling competence in the evolved mutants without compromising stability or functional expression levels. S417 was found to be a promiscuous position.

Library Design and Cloning. To construct the 376 libraries used in this study, the codons encoding NTR1-D03 residues 43–418 were randomized by replacing them, one at a time, with the diversified trinucleotide NNN, where N stands for an equimolar mixture of A, G, C, and T. Depending on the position of the diversified codon relative to the 5′ and 3′ ends of the gene, one of two PCR-based methods was used.

To randomize codons distant from the two ends, two mutagenic oligonucleotides per diversified codon were designed. The first mutagenic oligonucleotide contained the diversified NNN sequence at the targeted codon, but still hybridized with the wild-type D03 sequence with a T$_m$ of 55–65 °C. A second oligonucleotide was designed to hybridize to either the region 5′ of the NNN on the first oligonucleotide, or else form the reverse complement of the first oligonucleotide, including the NNN region (Fig. S1*A*). Each of these oligonucleotides was paired with the appropriate amplification oligonucleotide hybridizing at the end of the gene and encoding one of the two cloning sites (p1fw and p2re; see Table S2 for primer sequences), resulting in PCR amplification of half of the D03 gene. The PCR products of the two separate amplification reactions were then combined and extended to generate the full-length D03 gene containing a single diversified codon. The full-length product was amplified further using 5′-biotinylated primers, which contained recognition sites for the type IIS endonuclease BsmBI, the cleavage with which would generate sticky ends compatible with BamHI and XmaI restriction site overhangs.

The amplification product was digested with BsmBI, any remaining undigested fragment and cleaved overhangs were removed by adding streptavidin and filtering the mixture, and the digested and purified DNA fragments were ligated into the acceptor vector, pRG-del. This acceptor vector is a derivative of pRGD03 (2) in which the BamHI-XmaI fragment encoding D03 has been replaced with a stuffer sequence containing multiple restriction sites (Fig. S1*B*). To clone each amplified full-length library into the acceptor vector, the acceptor vector was first digested with NotI, XmaI, and BamHI, and dephosphorylated with calf intestine alkaline phosphatase (all from New England Biolabs), then purified. The cutting with NotI would destroy intact stuffer fragment and thus prevent it from religation.

To randomize codons 43–50 and 418, all of which are close to one of the two cloning sites, a mutagenic oligonucleotide was designed to contain both the cloning site and an NNN sequence at the targeted site. A second, nonmutagenic oligonucleotide was designed to hybridize to the opposite terminus of the gene. In addition, both oligonucleotides contained the recognition site for BsmBI, the cleavage which would generate sticky ends compatible with overhangs produced by BamHI and XmaI. The two oligonucleotides were used to amplify full-length D03 from wild-type template; the resulting diversified fragment was further amplified using primers that also contained 5′ biotin. As described above, the amplification product was digested with BsmBI, purified, and ligated into pRG-del.

To confirm that library diversity conformed to the design and was free of bias, a representative subset of library variants was sequenced by Sanger sequencing. On average, every second naïve library was further analyzed by sequencing of individual clones (219 libraries). Sequences of a total of 2,170 clones from naïve (unselected) libraries (on average, 13 clones per library) were shown to contain an even distribution of the four bases in all codon positions (Table S1), consistent with full randomization. The observed codon distribution was found to be consistent with unbiased randomization. No bias toward any codon, especially the wild-type codon, was detected. Also after selection, the frequencies of synonymous codons were very similar, with one exception. In the data derived from 454 sequencing, the use of the CCC codon appears low (Fig. S6); in fact, this is an artifact of sequence evaluation, because 12 of 19 prolines in rNTR1 are encoded by CCC. Because the wild-type codon (here CCC) is obscured in our evaluation of the 454 sequencing reaction and, additionally, substitutions to proline are tolerated only at very few positions, proline codons are rarely found, and thus other positions do not contribute additional CCC codons.

For ligation, 33 ng digested and purified DNA library PCR product was mixed with 55 ng digested and purified vector pRG-del and ligated for 16 h at 16 °C in the presence of 0.25 units of T4 DNA Ligase (Invitrogen). The ligation mix was purified using StrataClean Resin (Stratagene), and *E. coli* DH5alpha cells were transformed by electroporation (GenePulser II; Bio-Rad). Cells were recovered in 1 mL super optimal broth with catabolite repression medium for 1 h, centrifuged at 6,000 × *g* for 3 min and cells were plated on four 12- × 12-cm LB agar dishes (LB/1% glucose/100 μg/mL ampicillin). Colonies were scraped off the plates and resuspended in LB medium (1% glucose and 100 μg/mL ampicillin), adjusted to 20% final glycerol, and aliquots were snap-frozen and stored at −80 °C until further use.

Proof-of-Principle Experiment. Position Y347$^{7.31}$ is a crucial residue for ligand binding of rNTR1, and its mutation decreases ligand binding (3). Y347$^{7.31}$ was thus chosen for a proof-of-principle experiment to test selection pressure during FACS selections. The randomized library was generated and subjected to one round of FACS selection as above, and different selection stringencies were analyzed. Clones recovered from selection of the top 1% of receptor-expressing cells exclusively contained the two tyrosine codons, indicating both appropriate selection pressure and selection based on phenotype, i.e., the protein sequence. All libraries were selected accordingly.

Library Expression and Selection. Sixty milliliters 2YT medium (0.2% glucose, 100 μg/mL ampicillin) were inoculated to OD_{600} = 0.05 and grown at 37 °C for ∼2 h to OD_{600} = 0.5. Protein expression is under control of the *lac* promotor, and expression was induced by addition of 250 μM isopropyl-β-D-thiogalactopyranoside and continued for 20 h at 20 °C. Expression cultures were cooled to 4 °C, and an aliquot corresponding to 10^7 cells was centrifuged for 3 min at 6,000 × *g* in a table-top centrifuge, washed in 1 mL TKCl buffer [50 mM Tris (pH 7.4), 150 mM KCl] and resuspended in 1 mL TKCl buffer with 20 nM BODIPY-labeled neurotensin (BP-NT; residues 8–13; neurotensin was obtained from AnaSpec, BODIPY-FL from Molecular Probes, Invitrogen). Samples were kept in the dark at 4 °C for 1–2 h to allow for agonist binding. Cells were washed twice in 1 mL TKCl buffer filtrated through a 50-μm mesh to avoid cell aggregates that could clog the FACS tubings (Becton Dickinson). Washed and singularized cells were applied to selection for high expression using the fluorescence-activated cell sorting approach essentially as described in Sarkar et al. (2). Here, 2,500 cells with expression levels corresponding to the top 1% of a D03 control were collected for each library. Selected cells were plated overnight on LB agar dishes (1% glucose, 100 μg/mL ampicillin; see Fig. S2 for a schematic overview of the selection process).

Selected libraries were analyzed both by Sanger sequencing of the full-length gene and by 454 sequencing of segments carrying the respective mutation.

In preparation for Sanger sequencing, for each library, DNA from 24 single clones was amplified by colony PCR (cPCR) and analyzed by Sanger sequencing of the PCR product, resulting in a sequence covering the whole GPCR. Colonies were scraped off the agar plates, and glycerol stocks were prepared as described for the naïve libraries to conserve the selected libraries.

For 454 sequencing analysis, the glycerol stocks were used for DNA isolation of each of the selected cell pools (see below). The obtained DNA library was prepared for ultradeep sequencing as summarized in Fig. S4 *A* and *B*. How the different PCR products covering different mutant positions were mixed and prepared for 454 sequencing is detailed in Fig. S4*C*.

cPCR and Sequencing. A 20-μL PCR mix containing 1× PCR buffer (Invitrogen 10× PCR buffer without $MgCl_2$), 2 mM $MgCl_2$, 0.8 mM dNTP mix, 100 nM primer NTR1longfw, 100 nM primer NTR1longre (see Table S2 for primer sequences), and 1 μL Taq polymerase was set up per well of a 96-well PCR plate. A single colony was picked gently, using a small pipette tip, and cells were resuspended in the PCR mix by repetitive pipetting. Cells were broken by a 10-min incubation step at 96 °C, following by 35 cycles of 95 °C for 15 s, 54 °C for 15 s, and 72 °C for 100 s, followed by a final elongation step of 5 min. PCR products were purified and sequenced at Synergene Biotech. A total of 6–20 sequencing reactions were performed per library.

Preparation of 454 Sequencing Samples. A glycerol stock of the selected library pools was used to inoculate a 5-mL 2YT culture (1% glucose, 100 μg/mL ampicillin) and grown at 37 °C for 16 h. Cells were centrifuged, and DNA was isolated using a Qiagen BioRobot 8000 with Macherey-Nagel NucleoSpin 96 Plasmid Kits. A 100-ng template DNA was used per amplicon PCR. At the time of the experiment, the read length performance of the 454 technology typically reached 250 bp. Amplicons were thus designed as 250-bp overlapping fragments to fully cover the receptor sequence (Fig. S4*A*; see Table S2 for primer sequences). By design, each library contains only one diversified codon, allowing us to sequence only a small region of the gene (because the Sanger sequencing of the whole gene had already confirmed the absence of other mutations outside of the designed ones). By using the appropriate amplicon primer pair, an amplicon (PCR product) containing the diversified codon was generated (Fig. S4*B*). The PCR products were purified by ultrafiltration (NucleoFast 96 PCR; Macherey-Nagel), and the DNA concentration was quantified using Quant-iT PicoGreen (Invitrogen). The DNA amount corresponding to $2 \cdot 10^{11}$ DNA molecules was used as input material for 454 sequencing (Functional Genomics Center Zürich). DNA from different libraries was mixed in one reaction (Fig. S4*C*), and later assigned to its original library by sequence reference alignment (Fig. S5).

We decided to not sequence all randomized positions falling within one amplicon (typically 50) in the same pool, but rather distribute them over eight aliquots, such that each amplicon is pooled from only 6–8 randomized positions, for the following reason: in a mixed sample of 50 different PCR products coming from the 50 different randomized positions within one amplicon, for each codon 2% of sequence reads would be assignable as mutations within this codon. However, for each position, the remaining 98% of sequence reads would be wild type in a given position and thus align perfectly to D03 in that specific codon, because their diversified codon is at a different position. This high frequency of wild type would make the accurate determination of mutant frequencies rather difficult. From a pilot experiment, we learned that the noise of our 454 sequencing reaction is ∼0.1%, meaning that 0.1% of sequence reads contain one or more mismatches in the nonrandomized gene regions. The observed error frequency here is a sum of base-pair mismatches introduced during amplicon PCRs as well as sequencing errors during the 454 sequencing reaction. The 454 sequencing reaction accuracy is mostly limited by shortcomings in the identification of homopolymer length by pyrosequencing (4, 5), and the noise of 0.1% observed here is mainly observed in homopolymeric sequence stretches. Here, all sequences showing miscalled bases in addition to the randomized codon were excluded from analysis. In the given example, an assignment of 2% of sequences per library and 0.1% noise would have resulted in a signal-to-noise ratio of only 20.

To obtain a higher ratio of assignment percentage over noise percentage (i.e., signal over noise), we have prepared the samples as follows (Fig. S4*C*): 376 different libraries were grouped into eight individual 454 sequencing reactions, run separately on one-eighth of a 454 picotiterplate. Every reaction contains 45–52 individual PCR products in total, of which 6–8 fall within each amplicon (denoted as amplicons 1–7). Alignments of 454 sequence reads to D03 (i.e., assignment of 454 sequence reads to a specific library origin) were later run independently for each 454 sequencing reaction and amplicon. One alignment of 454 sequence reads to D03 thus contains sequence reads of 6–8 individual amplicon PCR products, meaning that sequences from this pool are randomized in one of these 6–8 positions. Thus, 12.5–16% of sequences are assigned per library (i.e., for a given codon that was randomized in this set, this is the percentage of sequences actually carrying an altered codon in this codon), resulting in an improved signal-to-noise ratio of at least 125 (12.5% over 0.1%). Nonetheless, sequence reads of an amplicon fully identical to the D03 reference sequence could not be assigned to which (maintained) position it came from, requiring a statistical correction during data processing (Fig. S5).

Analysis of 454 Sequencing Data. The raw sequences were transferred from the 454 system as large FASTA-format text files. B0.fastA, the pilot experiment, contained 206,405 sequences, all read in forward direction. B1.fastA to B8.fastA contained a total of 638,976 sequences, read in forward or reverse direction. The sequences were imported into EXCEL and processed using a set of custom Visual Basic macros. For our experiment, it was crucial to keep the background of sequencing errors as low as possible. Therefore, we stringently eliminated unreliable parts of the sequence. The 454 sequences tend to acquire insertions and deletions in runs of the same nucleotide, and indeed most sequences eventually went out of frame. Because we did not expect any frameshifts from the original sequences (because functional receptors were selected), we took the occurrence of frame shifts in the sequence as a sign that the sequence quality had deteriorated and truncated the shifted part of the sequence.

To identify the amplicon and the reading frame, we searched for the first 12 base exact match between each sequence and the D03 sequence (forward and reverse), and cut off all nucleotides before the start of the first in-frame codon match. The sequence was aligned without gaps and cut into in-frame nucleotide triplets. Mismatched codons were trimmed back from the end of the sequence until a four-codon (12 base) exact match was found. The truncated sequences were compared with the D03 sequence. Only those sequences that differed by exactly one codon from the D03 sequence were used for further analysis; for B0, these were 86,734 of 206,405 sequences (42%), covering 48 randomized positions. For each randomized position, an average of 1,800 sequences differed from the consensus; for unrandomized positions, 7.6 sequences (noise <0.5%). For B1–B8, 476,322 sequences (69.6%) showed exactly one deviation from D03, covering 331 positions with 771 forward and 737 reverse sequences (total 1,400 sequences) per randomized position, against a background of 4.4 mutations in nonrandomized positions.

Codon and Amino Acid Frequency Distribution. For each amplicon, comprising 3–8 randomized positions, the frequency distribution of the 64 codons was determined for each position in the sequence. The frequency of the wild-type codon in the randomized positions could not be determined directly, because it could not be distinguished from the wild-type codon originating from the sequences randomized in a different position. For amino acids encoded by several synonymous codons, the frequency of the wild-type codon was estimated as the average of the frequencies of synonymous codons. This correction could not be applied for the two amino acids encoded by a unique codon, Met (ATG) and Trp (TGG). The codon frequency distribution was normalized to give a sum over all 64 codons of 100%. The amino acid frequency distribution was derived as the sum of the frequencies of synonymous codons and renormalized to a sum of 100%. For further analysis, we evaluated the effects of selection on the width of the frequency distribution (sequence variability) and the peak of the distribution (sequence consensus). The amino acid distributions demonstrated a high tolerance of the rNTR1-D03 toward randomization: The positional variability of the sequences recovered from the top 1% neurotensin-binding *E. coli* clones isolated by FACS was comparable to that of all class A (rhodopsin-like) GPCR sequences.

Unbiased NNN randomization, combined with the degeneracy of the genetic code, introduces a bias in the amino acid distribution. Ser, Arg, and Leu, encoded by six codons each, are sixfold overrepresented over Trp and Met, encoded by a single codon; other amino acids lie between the two extremes. For most positions, the selection pressure was not strong enough to overcome this intrinsic bias: between amino acids with similar properties, the amino acid sequence consensus frequently went to the one encoded by the highest number of codons, e.g., to Leu in membrane-embedded positions that tolerated Leu, Val, Ile, and Met, or to Arg in positions that required a positively charged amino acid. Usual metrics of sequence variability [e.g., Shannon sequence entropy (6, 7) or Kabat sequence variability (8)] and sequence consensus did not perform well in this context, as they usually assume all amino acids to be equally probable and therefore bias the results in favor of the amino acids overrepresented in the original library, due to the degeneracy of the genetic code. The amino acid consensus sequence therefore differed significantly from the translation of the codon consensus (Fig. S8), and amino acids encoded by a larger number of codons on average appeared more conserved than amino acids encoded by fewer codons.

To avoid this bias, the rmsd of the observed amino acid distribution from the input amino acid distribution generated by unbiased NNN codon randomization was chosen as a measure of the selective pressure shaping the amino acid distribution in a given position. The rmsd for a given position is calculated according to Eq. **S1**:

$$rmsd = 2\sqrt{\frac{\sum_{i=1}^{20}\left(f_{l_i} - f_{s_i}\right)^2}{20}}, \qquad \textbf{[S1]}$$

where f_{1_i} is the frequency of amino acid *i* in the library before selection and f_{s_i} is the frequency after selection. The frequency of amino acid *i* before selection is deduced from the theoretical distribution of a NNN library, which is justified by the Sanger sequencing results of the naïve libraries, which was found to be not biased (Table S1), and the high oversampling of the library diversity throughout all experimental steps. A low rmsd denotes a permissive position, where the selection process had little or no effect on the observed amino acid frequency distribution, whereas a high rmsd is a sign of a restrictive position with a clear amino acid preference. Intermediate rmsds frequently denote positions where the general character of the amino acid is preserved (e.g., aliphatic), but not the exact type (e.g., valine).

This distinction between permissive (low rmsd) and restrictive positions (high rmsd) does not depend on whether the wild-type sequence is conserved or not—a position can be restrictive, but shift away from the wild-type sequence to a new focus, or be permissive and still include and partially preserve the wild-type sequence (e.g., by only selecting for the aliphatic character of the amino acid).

Fig. 1 compares the rmsds obtained from the sequences of the selected members of the rNTR1-D03–based libraries to those derived from more than 20,000 aligned class A GPCR sequences obtained from the GPCR database (GPCRDB; http://www.gpcr.org/7tm/). The ratio of the rNTR1-D03 rmsds to the class A rmsds indicates in which of the two systems a given position is more highly conserved (Fig. 2).

The assessment of sequence conservation and sequence shifts was based on the comparison of the wild-type sequence to the selected consensus, generated from a normalized table of the observed amino acid frequencies divided by the number of synonymous codons for each amino acid. This normalized consensus better reflects the influence of the applied selection pressure on the amino acid distribution than the classical amino acid consensus (Fig. S8). A position was considered robustly conserved if the two consensus sequences (not normalized and normalized by codon frequency) both agreed with the wild-type sequence; it was considered weakly conserved if only the normalized consensus agreed. A robust shift meant that both consensus sequences agreed, but differed from the wild-type sequence. If the normalized consensus differed both from the wild type and the classical consensus, two classifications are possible: the result was classified as not significant if the average codon frequency for the consensus amino acid did not differ significantly from that of the wild-type amino acid, whereas it was termed a weak shift if the wild-type amino acid was clearly underrepresented. A robust shift or conservation can be the result of strong selective pressure, but it is not necessarily so. It can also be a mild selection of an amino acid additionally favored by the codon bias.

Homology Modeling. To visualize the positions and potential interactions of the different mutations, homology models of the ground state and the activated state were built. A structural alignment of all available GPCR structures was used to verify the sequence alignments and identify irregularities in the transmembrane helices in individual structures. Because selection was based on agonist binding, the model of the active state is shown in all figures. Because the first structures of a GPCR in complex with an agonist became available only very recently (9) (PDB ID code 3SN6), the model shown in Figs. 1–3 was based on the structures of opsin in complex with a C-terminal peptide of transducing alpha (PDB ID code: 3DQB; 3.2 Å resolution) and of squid rhodopsin (PDB ID code: 2Z73; 2.5 Å). Due to unresolvable clashes, the N-terminal side of the first helix had to be tilted further away from the axis of the helix bundle, its orientation was taken from a structure of the β2 adrenergic receptor (PDB ID code: 2RH1; 2.4 Å). Various templates and loop generation methods were used to model the loops connecting the helices. Comparison with aligned GPCR structures combined with hydrophobicity and sequence conservation pattern within the NTR lineage (GPCRDB 001_002_015) helped to choose between different loop conformations that satisfied the steric constraints. The conformation of the large third intracellular loop of rNTR1, which in most GPCR structures is replaced by T4 lysozyme, was patterned after the loop conformation observed in squid rhodopsin (PDB ID code: 2Z73), extending transmembrane helices 5 and 6 from PDB ID code 3DQB and connecting them in a helix-turn-helix motif. This loop conformation allowed fitting a Gα or Gαβγ structure to the Gα-derived peptide in template PDB ID code 3DQB without clashing with the loop. Insight II (Accelrys Inc.) and the Rosetta suite of programs (http://www.rosettacommons.org/) (10) were used for modeling; figures were generated using PyMol (Schrödinger, LLC).

Analysis of Single Mutants. Thirty-two shift mutations, based on the D03 background, were selected for further characterization from evaluation of the Sanger sequencing results (Table 1). The Sanger sequencing and the 454 sequencing results are in good agreement. However, the 454 sequencing results allow more robust statistical analysis, because on average 1,428 sequences were analyzed per individual position. Furthermore, statistical corrections were applied (see 454 data analysis above). Thus, for a few positions, the 454 sequencing proposes a different consensus amino acid than the Sanger sequencing results (e.g., position L119$^{2.56}$). We find a L119F$^{2.56}$ shift according to the Sanger sequencing results, whereas according to the 454 sequencing results, L119F$^{2.56}$ is the second preferred residue for L119$^{2.56}$, after Leu, suggesting that it is primarily conserved. The expression levels of L119F$^{2.56}$ show the feasibility of the substitution with respect to expression levels (Table 1), and thus that both residues behave very similarly.

Analysis of Stability in Detergents. Thermostability of selected mutants was analyzed essentially as described in Dodevski and Plückthun (11). A gradient PCR machine was used to incubate receptor aliquots at increasing temperatures (TProfessional Gradient; Biometra). Data were analyzed by a nonlinear fitting using GraphPad Prism 5. The apparent T_m is defined as the temperature at which 50% of receptor molecules retain ligand binding activity after 20 min of incubation.

1. Ballesteros JA, Weinstein H (1995) Integrated methods for the construction of three dimensional models and computational probing of structure function relations in G protein-coupled receptors. *Meth Neurosci* 25:366–428.
2. Sarkar CA, et al. (2008) Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. *Proc Natl Acad Sci USA* 105:14808–14813.
3. Barroso S, et al. (2000) Identification of residues involved in neurotensin binding and modeling of the agonist binding site in neurotensin receptor 1. *J Biol Chem* 275:328–336.
4. Chan EY (2009) Next-generation sequencing methods: Impact of sequencing accuracy on SNP discovery. *Methods Mol Biol* 578:95–111.
5. Ronaghi M, Uhlén M, Nyrén P (1998) A sequencing method based on real-time pyrophosphate. *Science* 281:363–365, 365.
6. Shannon CE (1997) The mathematical theory of communication. 1963. *MD Comput* 14:306–317.
7. Strait BJ, Dewey TG (1996) The Shannon information entropy of protein sequences. *Biophys J* 71:148–155.
8. Kabat EA, Wu TT, Bilofsky H (1977) Unusual distributions of amino acids in complementarity-determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites. *J Biol Chem* 252:6609–6616.
9. Rasmussen SG, et al. (2011) Crystal structure of the β_2 adrenergic receptor-Gs protein complex. *Nature* 477:549–555.
10. Leaver-Fay A, et al. (2011) ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 487: 545–574.
11. Dodevski I, Plückthun A (2011) Evolution of three human GPCRs for higher expression and stability. *J Mol Biol* 408:599–615.

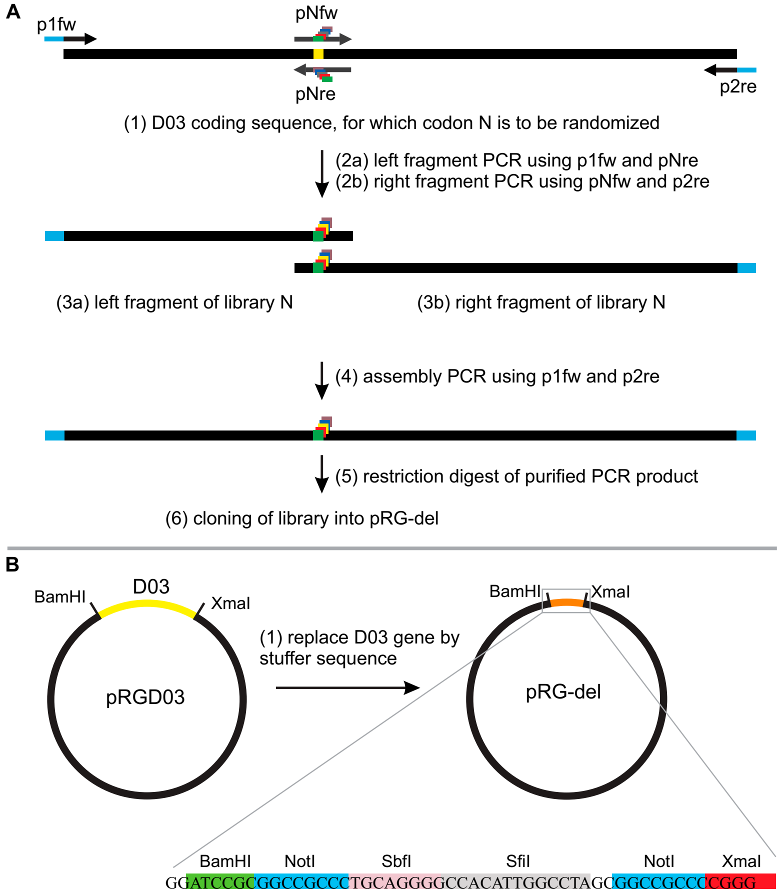
A
p1fw
pNfw
pNre
p2re
(1) D03 coding sequence, for which codon N is to be randomized
(2a) left fragment PCR using p1fw and pNre
(2b) right fragment PCR using pNfw and p2re
(3a) left fragment of library N
(3b) right fragment of library N
(4) assembly PCR using p1fw and p2re
(5) restriction digest of purified PCR product
(6) cloning of library into pRG-del
B
BamHI
D03
XmaI
pRGD03
(1) replace D03 gene by stuffer sequence
BamHI
XmaI
pRG-del
BamHI
NotI
SbfI
SfiI
NotI
XmaI
GGATCCGCGGCCGCCCTGCAGGGGCCACATTGGCCTAGCGGCCGCCCCGGG

Fig. S1. Library generation and cloning. (*A*) Libraries are generated by a two-step PCR assembly strategy. First, two separate PCR reactions are performed with the D03 coding sequence as template: With primers p1fw and pNre (2a), the 5′-end of the library N is generated. Primer pNre and pNfw are NNN randomized in the codon of library position N, thus introducing the desired randomization. Primer p1fw introduces a BsmBI site with BamHI-compatible overhangs, whereas primer p2re introduces a BsmBI site with XmaI-compatible overhangs (blue ends). With primers pNfw and p2re (2b), the 3′-end of library N is generated. The resulting PCR products (3a, 3b) are isolated and purified, and used as template for the subsequent assembly PCR (4). Primers p1fw and p2re are used to generate and amplify the full-length library PCR product from the two fragments. Primers p1fw and p2re introduce BsmBI restriction sites and BamHI (p1fw)- and XmaI (p2re)-compatible overhangs. The full-length library is purified and subsequently cloned into the acceptor plasmid pRG-del. (*B*) The acceptor plasmid pRG-del is generated from pRGD03 by exchange of the D03-coding sequence by a stuffer sequence using BamHI and XmaI.

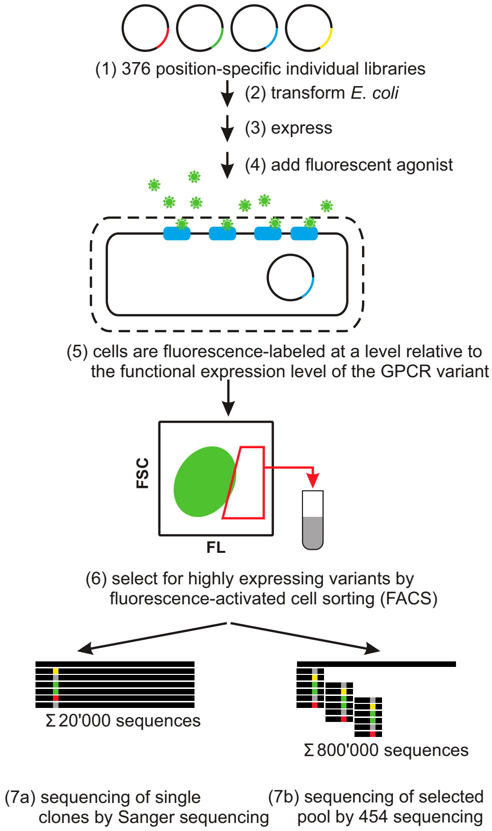
(1) 376 position-specific individual libraries
(2) transform *E. coli*
(3) express
(4) add fluorescent agonist
(5) cells are fluorescence-labeled at a level relative to the functional expression level of the GPCR variant
FSC
FL
(6) select for highly expressing variants by fluorescence-activated cell sorting (FACS)
Σ 20'000 sequences
Σ 800'000 sequences
(7a) sequencing of single clones by Sanger sequencing
(7b) sequencing of selected pool by 454 sequencing

Fig. S2. Selections for high functional expression. A total of 376 individual, position-specific libraries are transformed and expressed in *E. coli* DH5alpha cells (2, 3). The fluorescence-labeled agonist BP-NT (residues 8–13) is used to label the cells according to their functional expression levels (4, 5). Highly expressing cells are identified and isolated by FACS (6). Individual selected cells are grown and then analyzed by Sanger sequencing of the complete gene (7a), and the pool of selected cells is also analyzed by 454 sequencing of a gene segment containing the mutation (7b).

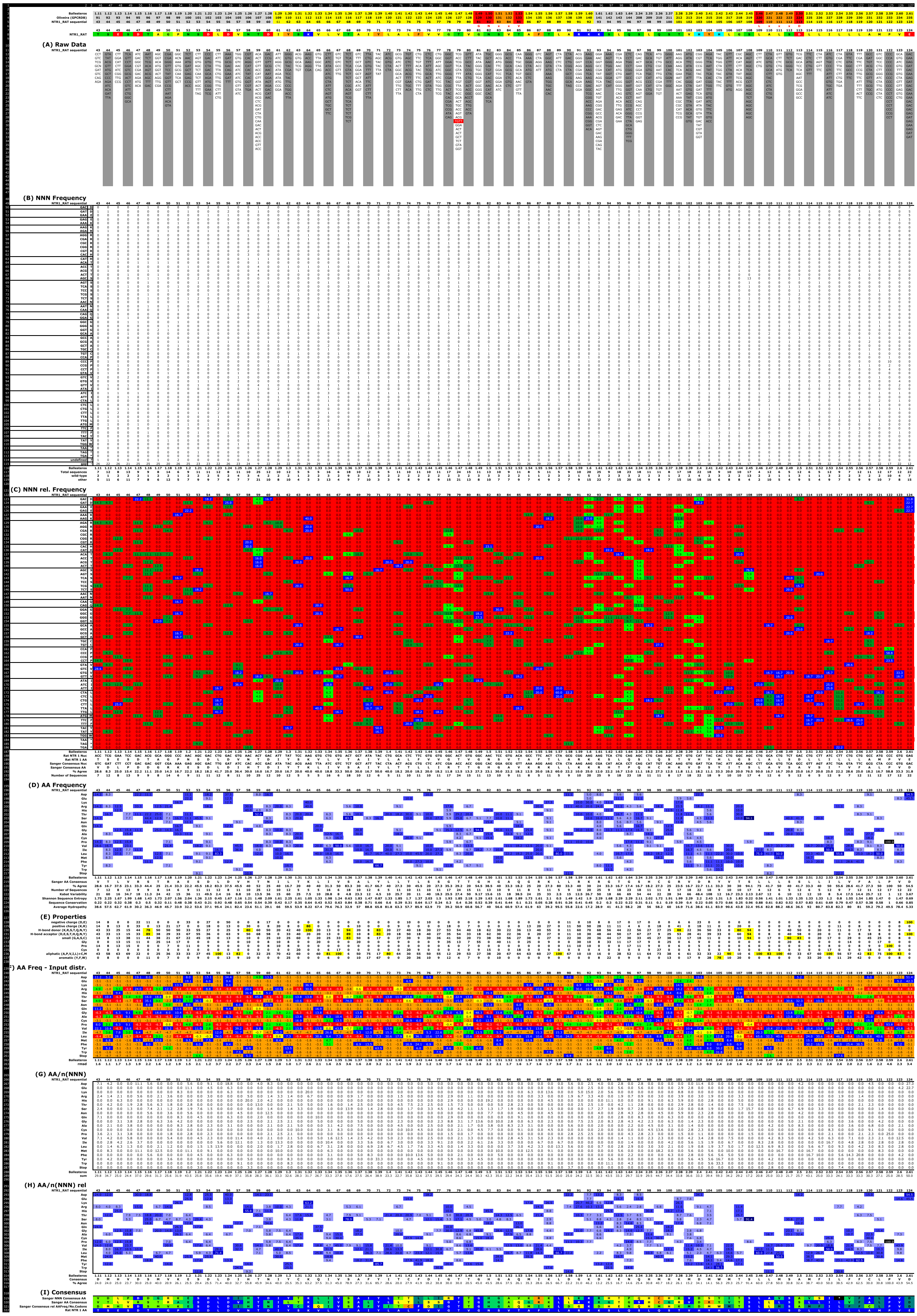

Fig. S3. (Continued)

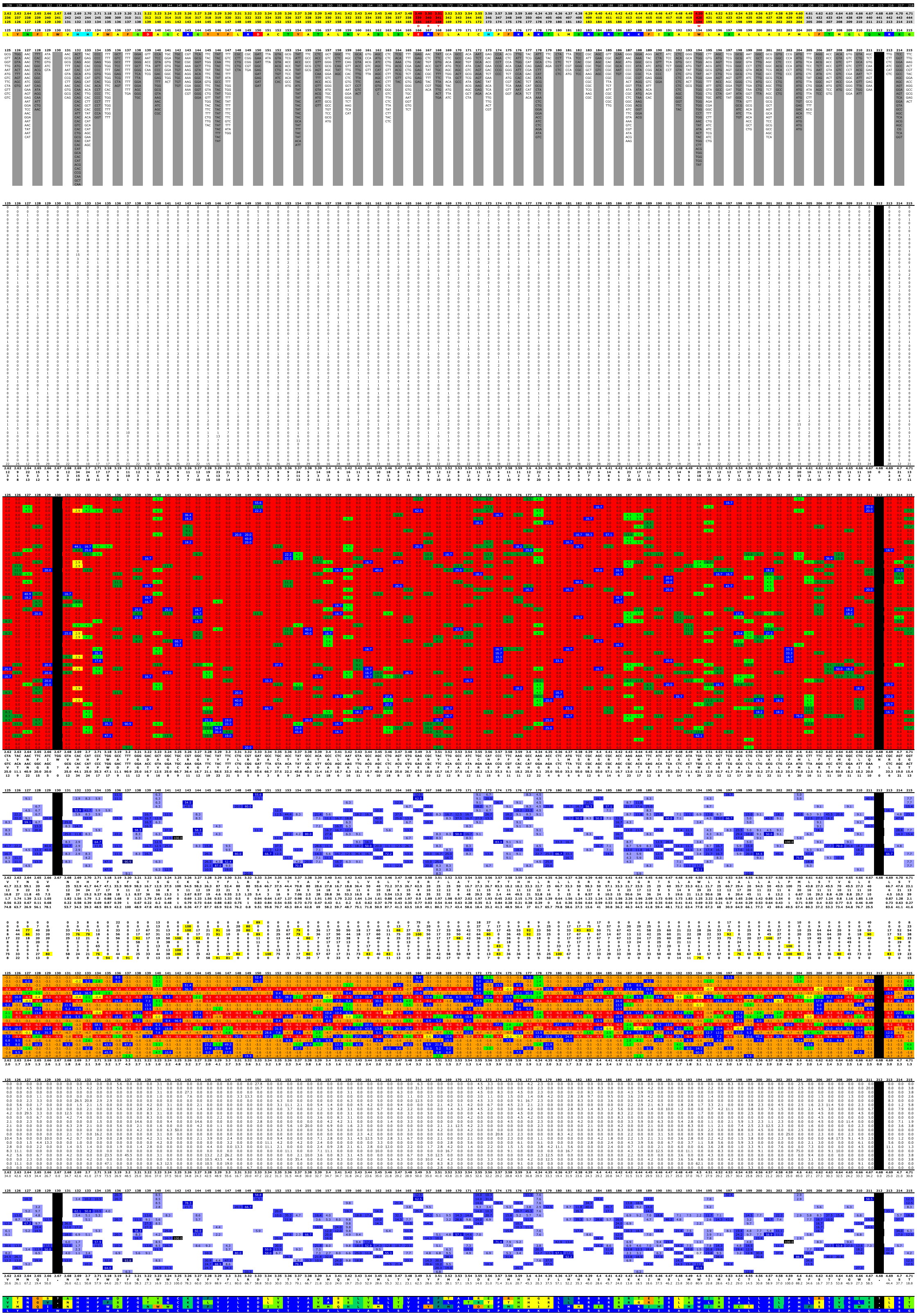

Fig. S3. (Continued)

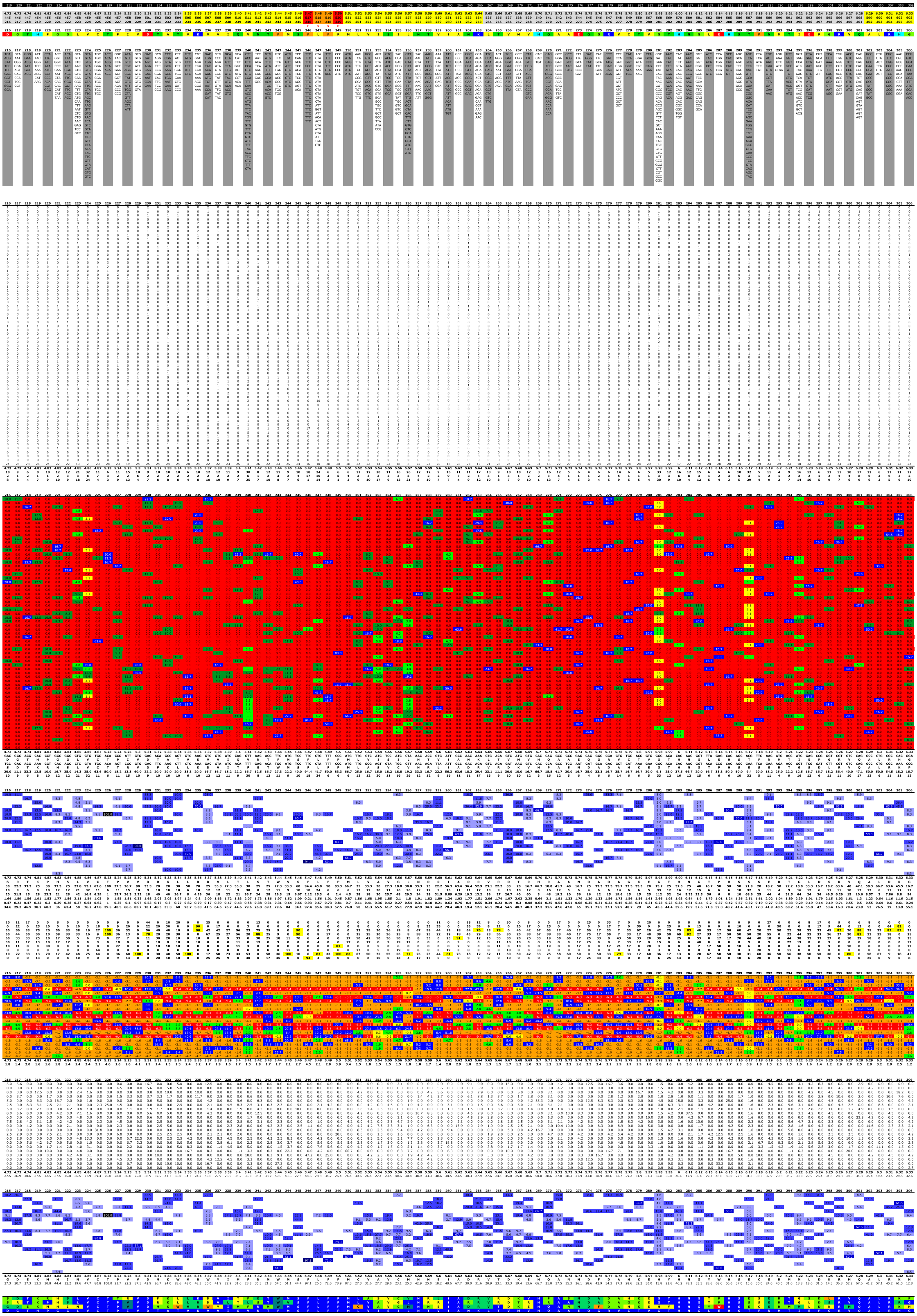

Fig. S3. (Continued)

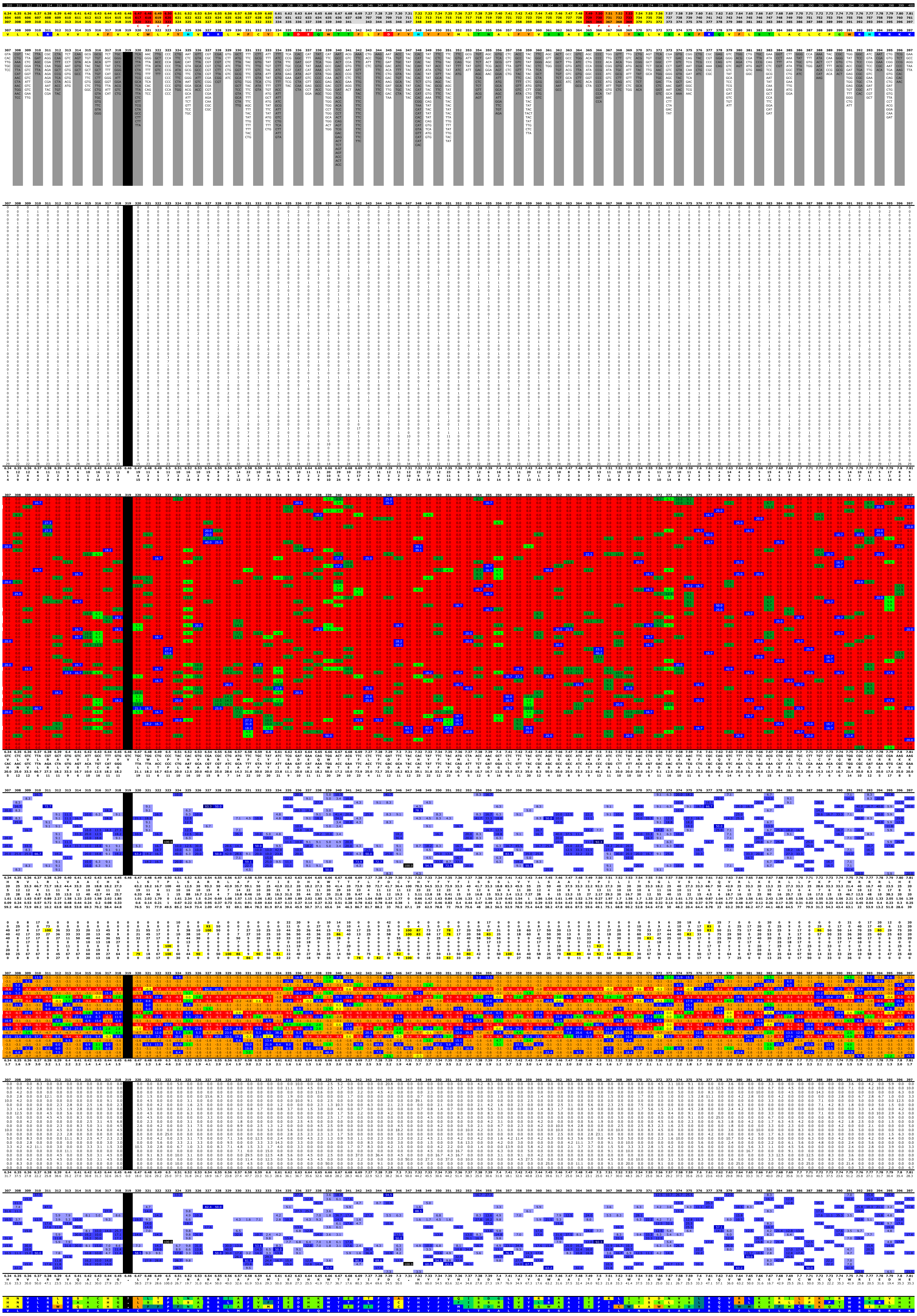

Fig. S3. (Continued)

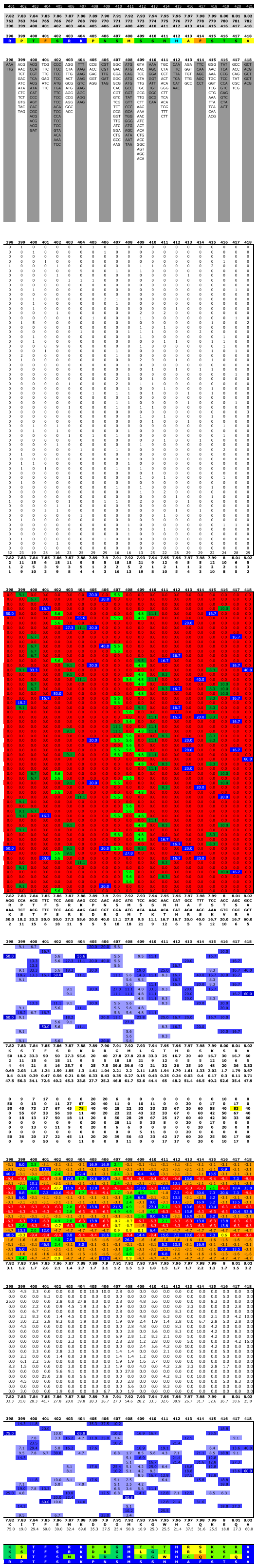

Fig. S3. Sanger sequencing data. (*A*) Raw data enumerates the nucleotide triplets observed in the randomized positions of a total of 4,310 selected clones analyzed by Sanger sequencing, 1,152 of which showed the wild-type codon and 3,158 a different codon. (*B*) NNN frequency lists how often each of the 64 triplets has been observed in a given position. (*C*) NNN relative frequency normalizes this distribution to a sum of 100%. (*D*) AA frequency. The amino acid frequencies were derived from the relative codon frequencies in *C* by summing up the contributions of synonymous codons. Shades of blue indicate the prevalence of the different amino acids, from white for not observed at all to dark blue for highly enriched amino acids. Note that the rNTR1 sequence is given at the top. (*E*) Properties sums up the percentage of amino acids with similar properties, e.g., charged, hydrophilic, small, hydrophobic, aliphatic, and aromatic.

Legend continued on following page

Because an amino acid can belong to more than one group, the sum can be larger than 100%. (*F*) AA Freq-Input distr: the difference between the input amino acid distribution and the distribution after selection is given, which highlights the effect of selection. The number represents the rmsd between the two distributions, which was calculated according to Eq. **S1** (*SI Text*). Amino acids colored blue are strongly enriched; green, mildly enriched; yellow, no significant change; orange, mildly deselected; and red, strongly deselected. (*G*) AA/n(NNN): amino acid frequencies corrected for the bias of the input library by dividing the observed frequencies by the number of synonymous codons. (*H*) AA/n(NNN) rel.: normalizes the corrected distribution to a sum of 100%. The color code is the same as in *D*. (*I*) Consensus compares the consensus sequences derived from the most frequently observed codon, the most frequently observed amino acid, and the most frequently observed amino acid after correcting for the number of synonymous codons to the wild-type sequence of rNTR1. The color code indicates the level of conservation of the substitution in a spectrum from dark blue for sequence identity to red for the least-conservative substitution possible.

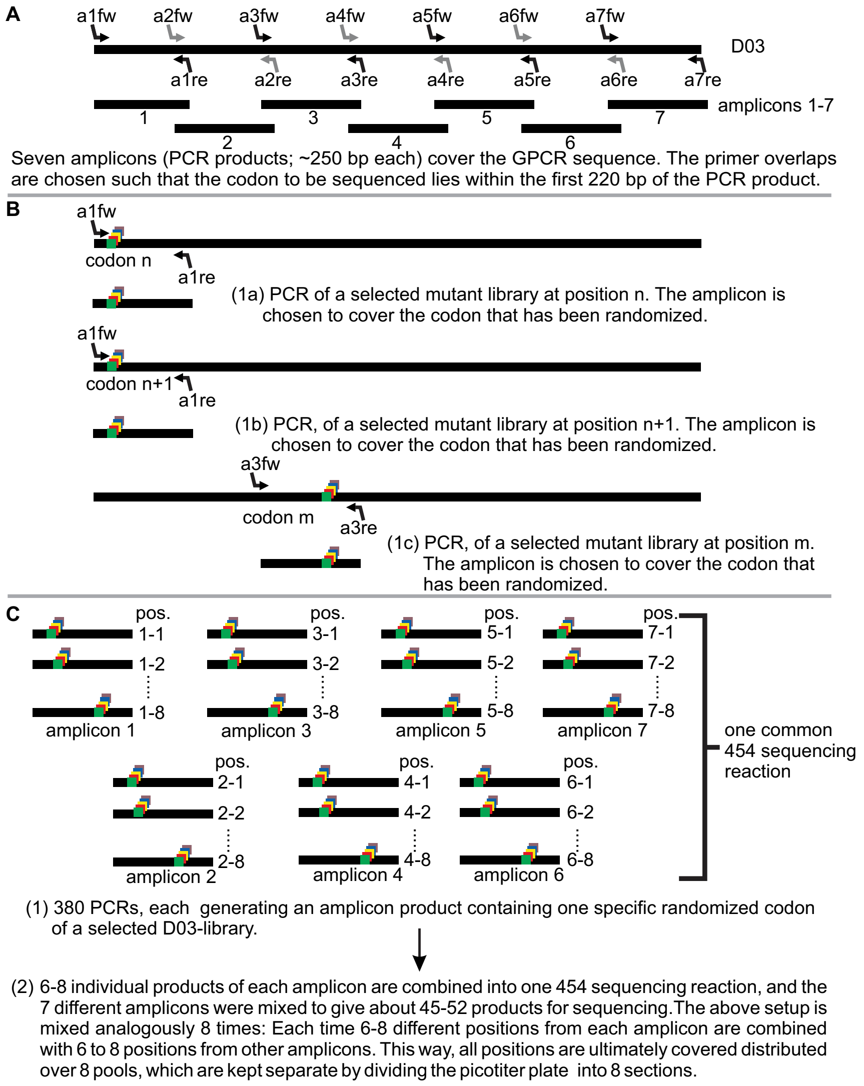
A
a1fw
a2fw
a3fw
a4fw
a5fw
a6fw
a7fw
D03
a1re
a2re
a3re
a4re
a5re
a6re
a7re
amplicons 1-7
1
2
3
4
5
6
7
Seven amplicons (PCR products; ~250 bp each) cover the GPCR sequence. The primer overlaps are chosen such that the codon to be sequenced lies within the first 220 bp of the PCR product.
B
a1fw
codon n
a1re
(1a) PCR of a selected mutant library at position n. The amplicon is chosen to cover the codon that has been randomized.
a1fw
codon n+1
a1re
(1b) PCR, of a selected mutant library at position n+1. The amplicon is chosen to cover the codon that has been randomized.
a3fw
codon m
a3re
(1c) PCR, of a selected mutant library at position m. The amplicon is chosen to cover the codon that has been randomized.
C
pos. 1-1
1-2
1-8
amplicon 1
pos. 3-1
3-2
3-8
amplicon 3
pos. 5-1
5-2
5-8
amplicon 5
pos. 7-1
7-2
7-8
amplicon 7
pos. 2-1
2-2
2-8
amplicon 2
pos. 4-1
4-2
4-8
amplicon 4
pos. 6-1
6-2
6-8
amplicon 6
one common 454 sequencing reaction
(1) 380 PCRs, each generating an amplicon product containing one specific randomized codon of a selected D03-library.
(2) 6-8 individual products of each amplicon are combined into one 454 sequencing reaction, and the 7 different amplicons were mixed to give about 45-52 products for sequencing.The above setup is mixed analogously 8 times: Each time 6-8 different positions from each amplicon are combined with 6 to 8 positions from other amplicons. This way, all positions are ultimately covered distributed over 8 pools, which are kept separate by dividing the picotiter plate into 8 sections.

Fig. S4. The 454 sequencing setup. (*A*) Seven amplicon PCR products (amplicons) of ~250 bp are designed to cover the GPCR gene. (*B*) For each selected library pool, an amplicon PCR containing the diversified codon is generated (1a–1d). (*C*) One 454 sequencing reaction contains 6–8 members of each amplicon. Eight individual 454 sequencing reactions are performed to analyze all 380 amplicon PCR products (2), kept separately on a subdivided picotiter plate.

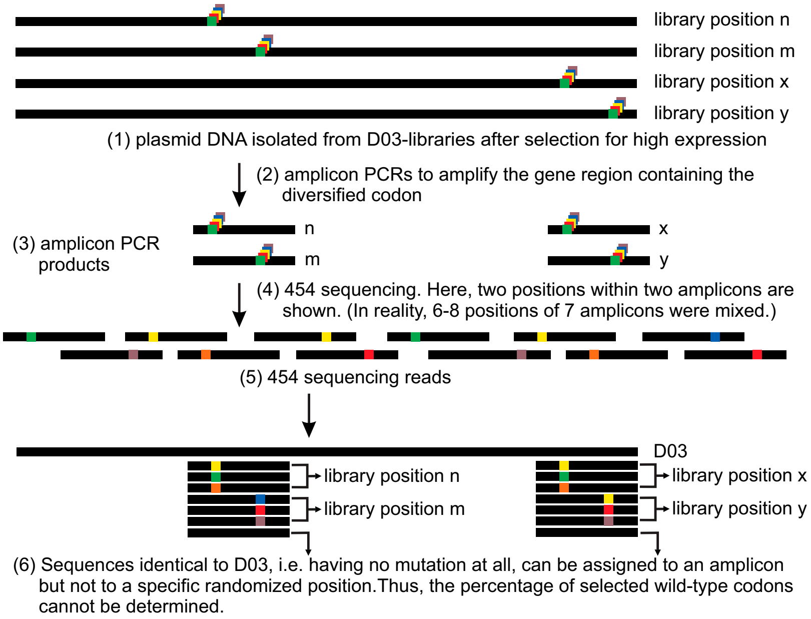
library position n
library position m
library position x
library position y
(1) plasmid DNA isolated from D03-libraries after selection for high expression
(2) amplicon PCRs to amplify the gene region containing the diversified codon
(3) amplicon PCR products
n
m
x
y
(4) 454 sequencing. Here, two positions within two amplicons are shown. (In reality, 6-8 positions of 7 amplicons were mixed.)
(5) 454 sequencing reads
D03
library position n
library position m
library position x
library position y
(6) Sequences identical to D03, i.e. having no mutation at all, can be assigned to an amplicon but not to a specific randomized position.Thus, the percentage of selected wild-type codons cannot be determined.

Fig. S5. The 454 sequencing read assignment. After amplification of the gene region containing the diversified codon (1–3), various amplicon PCR products are analyzed in one 454 sequencing reaction (4). The resulting 454 sequencing reads (5) are assigned by alignment to D03 (6). The diversified codon unambiguously identifies the library origin (6; library positions n, m, x, and y). However, 454 sequence reads with perfect alignment to D03 cannot be assigned to a particular randomized codon and statistical correction is applied for compensation.

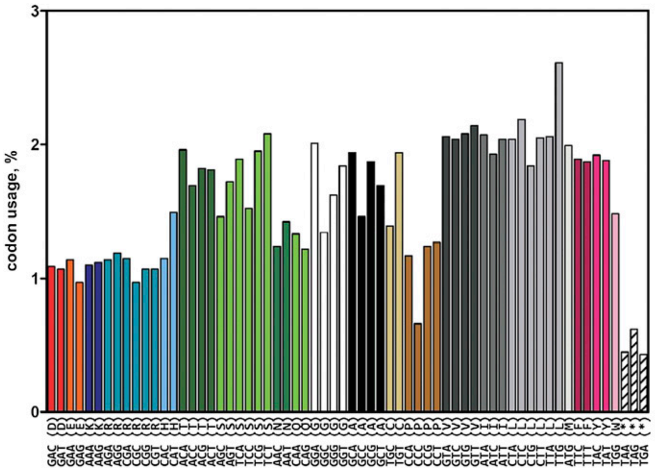

Fig. S6. Codon use at randomized positions was calculated from all 454 sequencing results. Synonymous codons are used at similar frequencies.

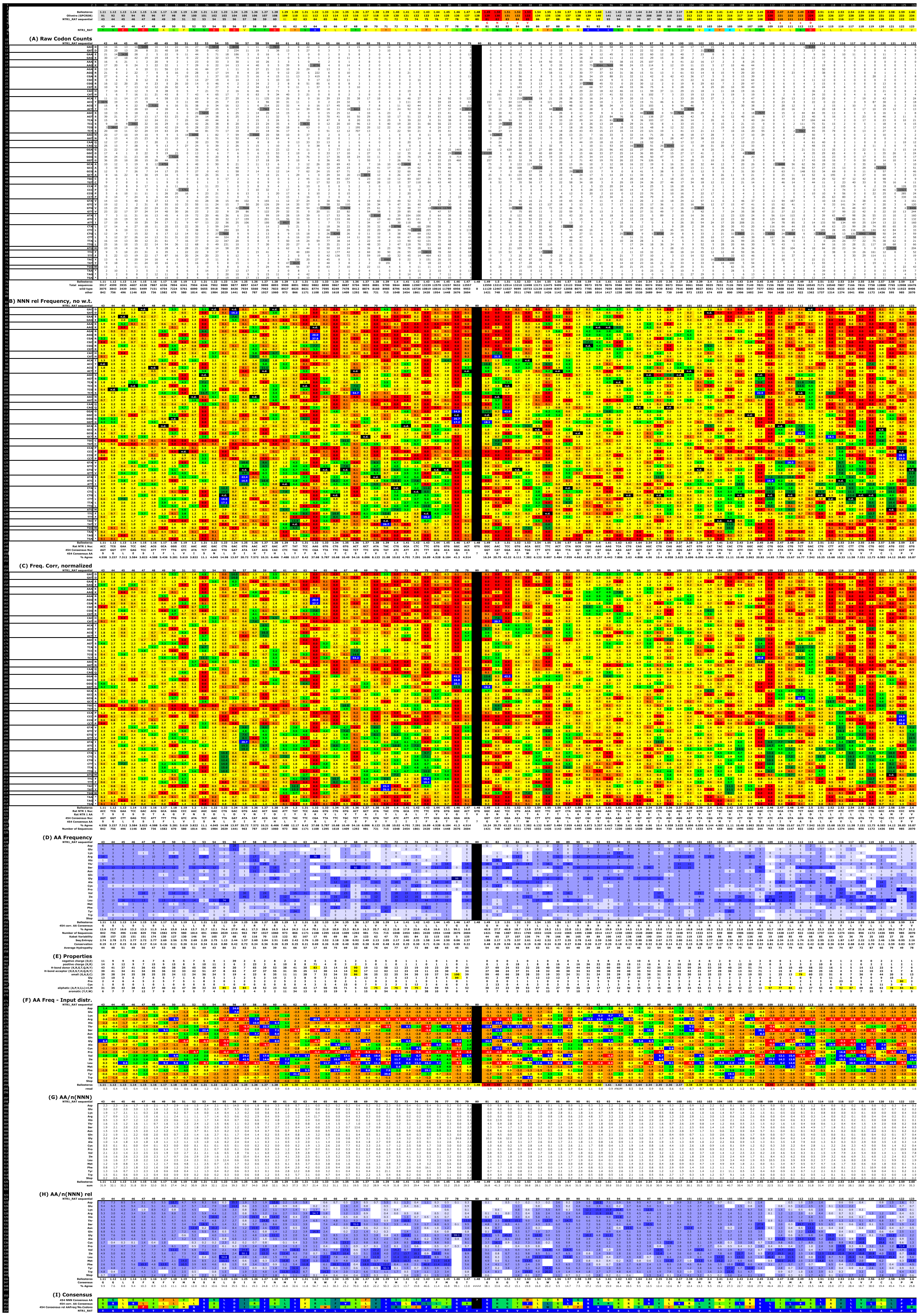

Fig. S7. (Continued)

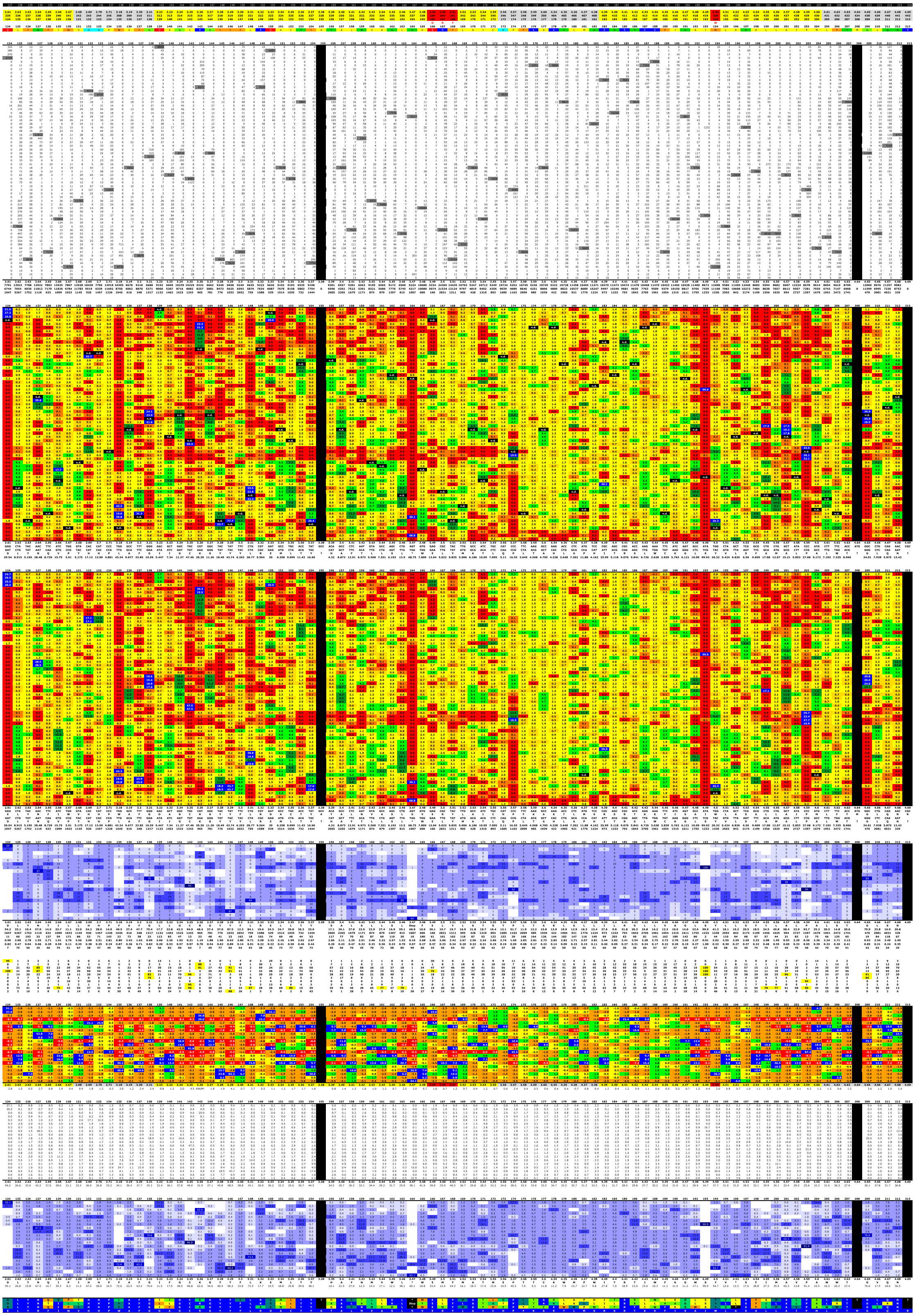

Fig. S7. (Continued)

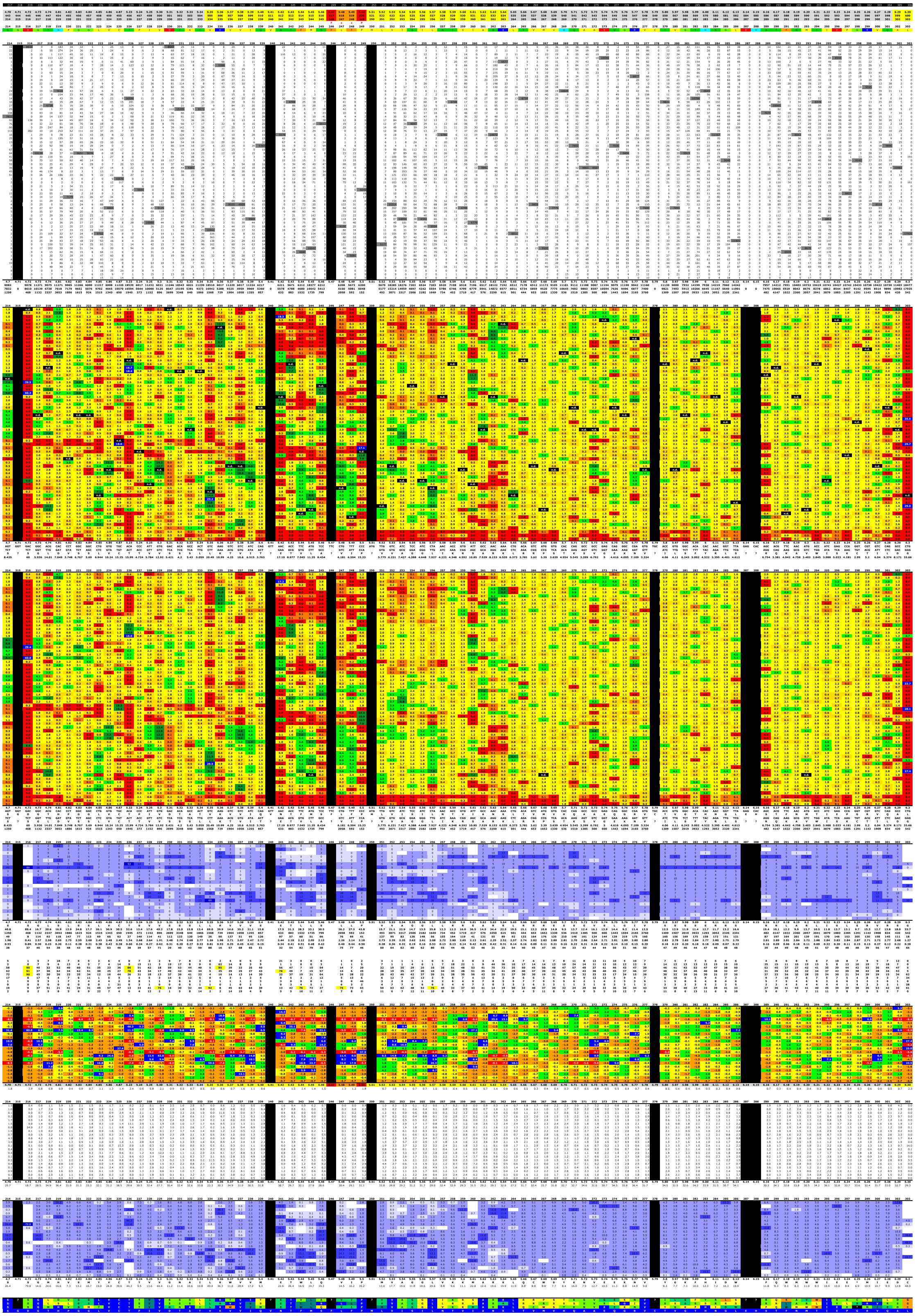

Fig. S7. (Continued)

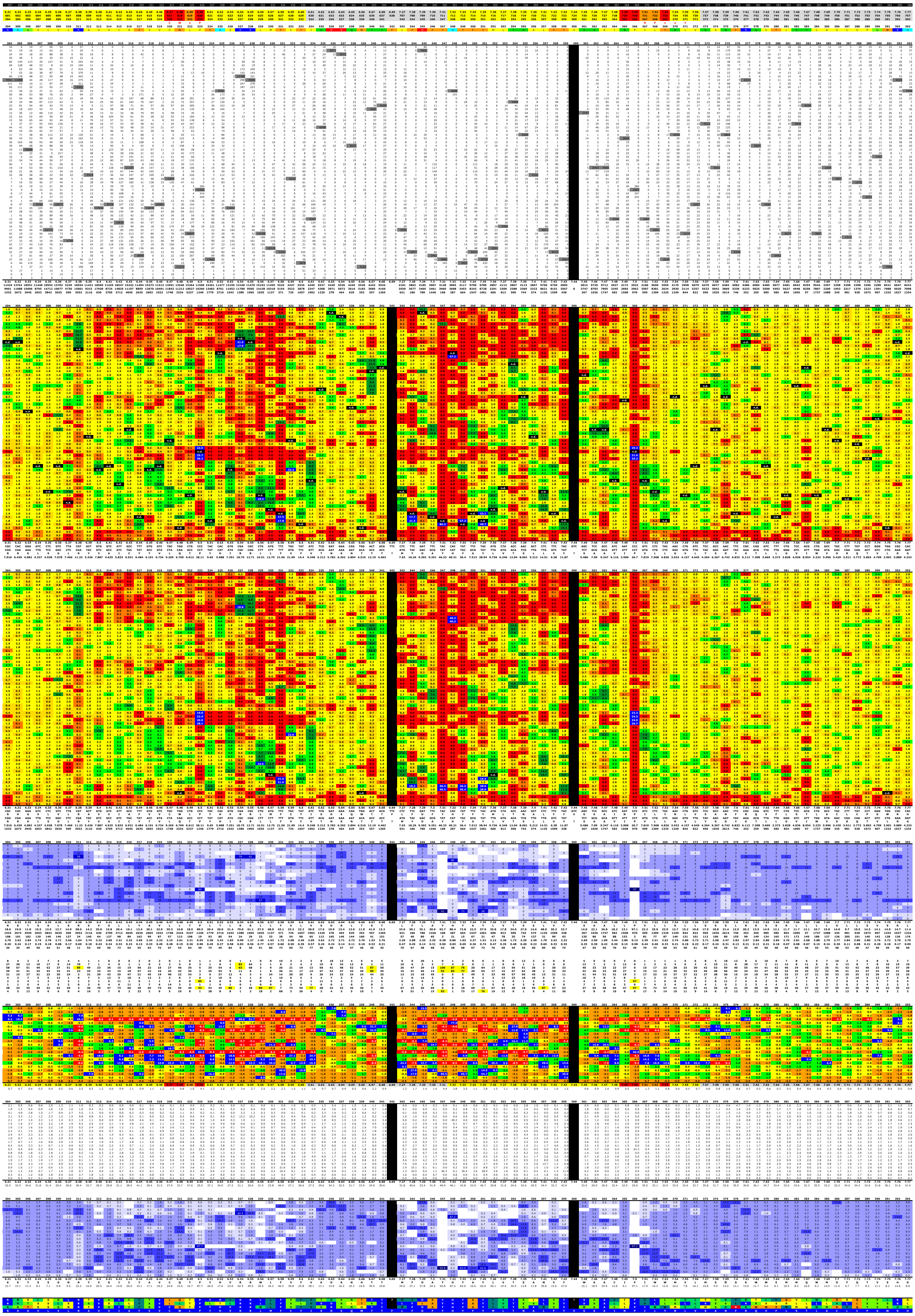

Fig. S7. (Continued)

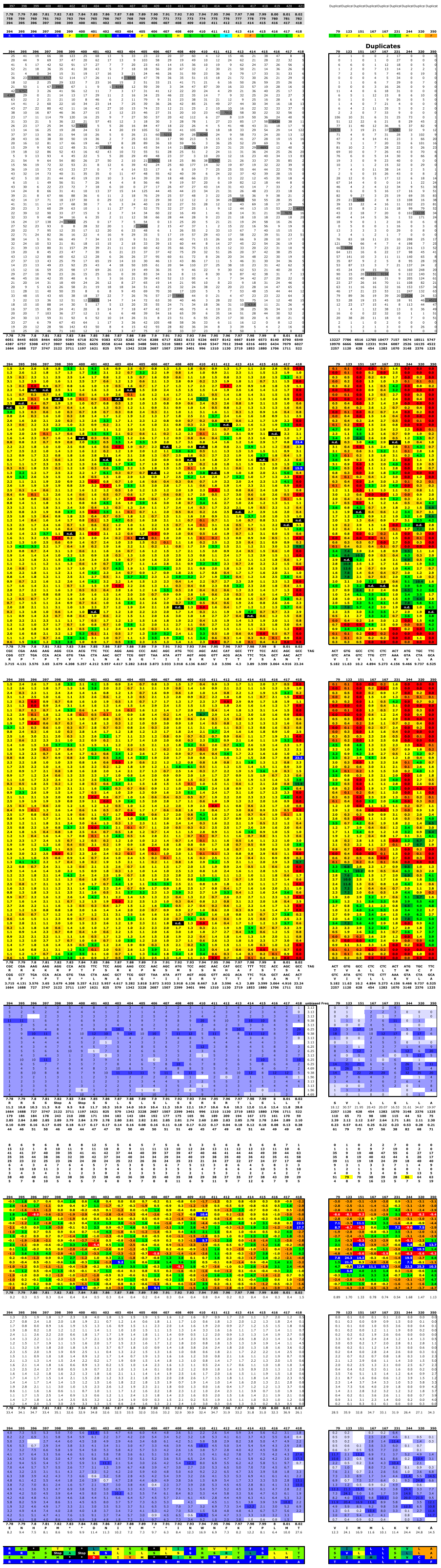

Fig. S7. The 454 sequencing data. (*A*) Raw codon counts enumerate how frequently each codon was encountered in the 454 sequencing results. The high count of the wild-type triplet results from the combination of multiple libraries randomized in different positions in each sequence pool (Fig. S4). (*B*) NNN rel Frequency, no w.t.: the relative frequencies of the different codons, normalized to a sum of 100% and omitting the wild-type codon. (*C*) Freq. Corr, normalized: the frequency of the wild-type codon was estimated as the average of the frequencies of synonymous codons where possible, and the values renormalized to a sum of 100%; *D–I* were derived from these values as described in the legend for Fig. S3.

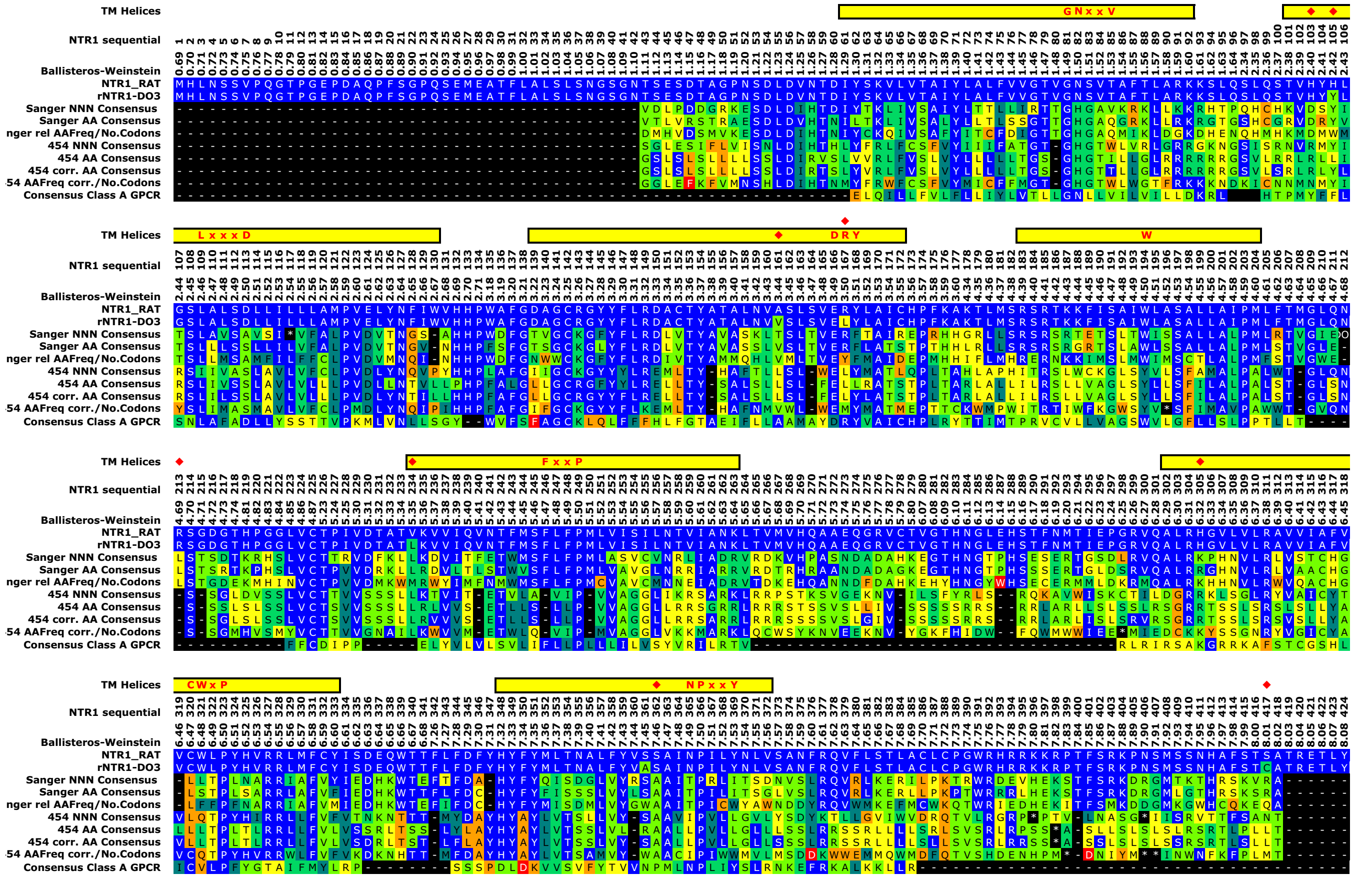

Fig. S8. Alignment of Sanger and 454 consensus sequences with rNTR1, rNTR1-D03, and class A GPCR consensus. Sanger NNN consensus/454 NNN consensus: The NNN consensus shows the amino acid sequence derived from the most frequently observed codon in each position. Sanger AA consensus/454 AA consensus: For the AA consensus, the observed frequencies of synonymous codons were summed up before deriving a consensus. The 454 corr. AA consensus: Because in the 454 experiment the frequency of the wild-type codon cannot be quantitated, its frequency has been deduced as the average of the frequencies of synonymous codons. The 454 AAFreq corr./no. of codons: AA consensus derived from a frequency table that has been corrected for the bias in the initial amino acid library introduced by the degeneracy of the genetic code. Consensus class A GPCR: The classical AA consensus derived from an alignment of all class A GPCR sequences downloaded from the GPCRDB (http://www.gpcr.org/7tm/), omitting the variable loop regions. At this level of sequence divergence, the amino acid frequency distributions show similar influence of codon bias as the 454 results. Figs. S3 and S7 contain the detailed data from which these consensus sequences were derived. In the top line the helices are highlighted, and conserved sequence motifs are indicated. Red diamonds indicate residues predicted to contact the C-terminal helix of Gα, based on transducin peptide contact residues in PDB ID 3DQB.

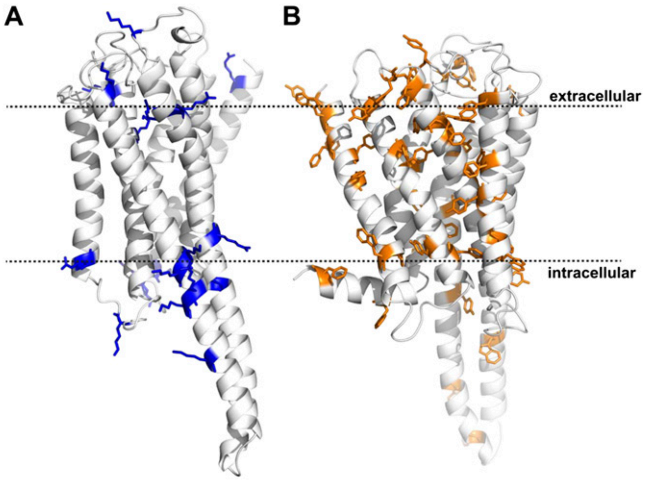

Fig. S9. General amino acid preferences. The figure is based on the consensus sequence shown in Fig. 3. (*A*) Significant selection for positively charged residues (Lys, Arg) is observed at the intracellular surface of the membrane and in the vicinity of the putative ligand binding pocket. (*B*) Membrane-exposed aromatic residues are predominantly located close to the surface of the membrane, whereas toward the center of the membrane, the Leu-Ile-Val-Met pattern predominates, in which these four amino acids predominate in a ratio governed by the codon bias.

Table S1. Diversity of randomized codons

	Codon position			
	1	2	3	Average
A	530	500	526	519
T	567	597	565	576
C	452	523	513	496
G	534	494	526	518
N	88	56	40	61
n = 2,170				

Table S2. Primer sequences

Name	Sequence 5′–3′	Purpose
NTR1longfw	CGCGCAGACTGGATCTAACAACAACAACAATAAC	cPCR
NTR1longre	CAGAACCGCCACCAGAACCGCCACCG	cPCR
p1fw	CTTCCAGTCTCGTCTCGGATCCACCTCGGAATCCGACACGGCAGGGCCCAACAGCGACC	Library generation/amplification
p2re	AATTCGTAGTGAAGACTCGTCTCCCCGGGTAGCGCAGGTGGAAAAGGCATGGTTGCTGGACATGC	Library generation/amplification
a1fw	CCTGTACTTCCAGTCTGGAT	amplicon 1 PCR (T43–Y104)
a1re	CCTGTACTTCCAGTCTGGAT	Amplicon 1 PCR (T43–Y104)
a2fw	AGCCTGCAGAGCACTGTGG	Amplicon 2 PCR (Y105–L163)
a2re	GAAGGGATGGCAGATGGCC	Amplicon 2 PCR (Y105–L163)
a3fw	GCCACAGCCCTCAATGTAGT	Amplicon 3 PCR (S164–G221)
a3re	GAGTGGCAGTGTCCACAATG	Amplicon 3 PCR (S164–G221)
a4fw	AACCTCAGTGGTGACGGCA	Amplicon 4 PCR (G222–V280)
a4re	AACGTGCTGTGCTCTAAACC	Amplicon 4 PCR (G222–V280)
a5fw	GCCAACAAACTGACAGTCATG	Amplicon 5 PCR (G281–M330)
a5re	CGAAGAGGAACGTAGTCCACT	Amplicon 5 PCR (G281–M330)
a6fw	TGGTCTGCTGGCTGCCCTAC	Amplicon 6 PCR (F331–T383)
a6re	CACCCAGGACAAAGGCAGG	Amplicon 6 PCR (F331–T383)
a7fw	AACCTGGTCTCCGCCAACTT	Amplicon 7 PCR (L384–A418)
a7re	TTCAGAACCGCCACCAGAAC	Amplicon 7 PCR (L384–A418)

Table S3. GPCR Sequence consensus motifs

	Consensus	Position	rNTR1
TM 1	GNxxV	$81^{1.49}$–$85^{1.53}$	GNSVT
TM 2	LxxxD	$109^{2.46}$–$113^{2.50}$	LALSD
TM 3	DRY	$166^{3.49}$–$168^{3.51}$	ERY, D03: ELY
TM 4	W	$194^{4.50}$	W
TM 5	FxxP	$246^{5.47}$–$249^{5.50}$	FLFP
TM 6	CWxP	$320^{6.47}$–$323^{6.50}$	CWLP
TM 7	NPxxY	$365^{7.49}$–$369^{7.53}$	NPILY