



OPEN

Directed evolution for high functional production and stability of a challenging G protein-coupled receptor

Yann Waltenspühl, Jeliasko R. Jeliaskov, Lutz Kummer & Andreas Plückthun✉

Membrane proteins such as G protein-coupled receptors (GPCRs) carry out many fundamental biological functions, are involved in a large number of physiological responses, and are thus important drug targets. To allow detailed biophysical and structural studies, most of these important receptors have to be engineered to overcome their poor intrinsic stability and low expression levels. However, those GPCRs with especially poor properties cannot be successfully optimised even with the current technologies. Here, we present an engineering strategy, based on the combination of three previously developed directed evolution methods, to improve the properties of particularly challenging GPCRs. Application of this novel combination approach enabled the successful selection for improved and crystallisable variants of the human oxytocin receptor, a GPCR with particularly low intrinsic production levels. To analyse the selection results and, in particular, compare the mutations enriched in different hosts, we developed a Next-Generation Sequencing (NGS) strategy that combines long reads, covering the whole receptor, with exceptionally low error rates. This study thus gave insight into the evolution pressure on the same membrane protein in prokaryotes and eukaryotes. Our long-read NGS strategy provides a general methodology for the highly accurate analysis of libraries of point mutants during directed evolution.

G protein-coupled receptors (GPCRs) play crucial roles in human physiology^{1,2}. This biological importance is further reflected in their therapeutic relevance. As such, GPCRs constitute the largest class of single drug targets, with an estimated 35% of marketed FDA approved drugs acting through these receptors^{3,4}. In recent decades, substantial efforts have thus been invested to elucidate the structural rationales for aiding the development of therapeutics acting on GPCRs. For a long time, biophysical experiments, including structural studies, have been complicated by most of the receptors' low natural abundance, low functional expression levels in recombinant hosts, and poor stability after removal from the membrane^{5,6}.

To address these difficulties intrinsic to GPCRs, several methodologies have been developed, including advances in recombinant overexpression, purification strategies^{7,8}, crystallization platforms^{9,10}, and protein engineering^{11–15}. This has increased the number of high-resolution structures, providing fundamental insights into the mechanistic understanding of at least some GPCRs^{16,17}. To advance drug discovery programs and to obtain a more comprehensive understanding of GPCR dynamics, complementary approaches to the structural studies are also necessary, including techniques such as surface plasmon resonance (SPR)^{18,19} and nuclear magnetic resonance (NMR)²⁰.

For such challenging experiments, GPCRs need to be sufficiently stable in detergent for a prolonged time period, and they need to be functionally expressible in a recombinant host suitable for isotope labelling. In addition, many receptors of great interest exhibit particularly low yields and/or low stability and have thus remained inaccessible to most biophysical and/or structural studies, despite the above-mentioned technological advances and the large interest in the field.

For the most challenging receptors, however, even the methods developed so far, when individually applied, have not been sufficient. In the present study, we have thus analysed the performance of these methods, and devised analysis tools to help understand why the serial use of different strategies of directed evolution is particularly successful.

Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. ✉email: plueckthun@bioc.uzh.ch

We have previously developed several evolutionary technology platforms to overcome the biophysical limitations of challenging receptors^{21–24}. These methods are all based on the powerful common principles of directed evolution: initially, the wild-type (wt) sequence of a protein is randomised, then, through application of a selection pressure, variants with increased functional expression^{21,23,24} or—directly—mutants with increased stability in detergent²² are enriched from a library. In these directed evolution experiments, the selection pressure is controlled by probing ligand binding with a fluorescently labelled ligand, thereby simultaneously ensuring functional folding of the receptor. The power of these directed evolution methods is underlined by the so enabled structural and functional studies of several reported GPCRs with improved the biophysical characteristics^{21,22,24–28}.

The initially developed evolution methods require expression in *Escherichia coli* (*E. coli*). *E. coli*-based evolution²¹ for improving functional receptor surface expression is not applicable to all GPCRs—some GPCRs cannot be functionally expressed at any meaningful level, possibly due to the missing secretory quality control machinery of the eukaryotic cell and/or differences in membrane lipid composition. A second *E. coli*-based method, aimed at improving protein stability in detergent by cellular high-throughput encapsulation solubilization and screening (CHESS)²², i.e., by encapsulating *E. coli* cells and converting them into nanoscale dialysis tubes, is only accessible to receptors that can be functionally expressed in *E. coli*. These limitations were overcome with the development of a third selection method based on a eukaryotic host²⁴. *Saccharomyces cerevisiae*-based receptor evolution (SaBRE) has opened the door to selection for challenging receptors that were not previously amenable to selection and evolution technologies in prokaryotic hosts. Nonetheless, SaBRE alone has also not been able to bring all receptors to the required level of stability.

We were interested in determining the structure of the human oxytocin receptor (OTR) and discovered that it is one of the most challenging targets for selection, as it expresses very poorly in almost all hosts—even in insect cells²⁹. To solve this problem, we decided to combine our methods, and we hypothesized that the implementation of SaBRE makes previously inaccessible receptors subsequently amenable to CHESS.

Due to its biologically conserved role in organizing both sexual reproduction and social behavior³⁰ the OTR is clinically targeted not only for preventing spontaneous premature labour, but also for the treatment of mental health disorders including Asperger's syndrome and schizophrenia^{31–34}. However, novel drug discovery programs targeting the OTR were so far complicated by the receptor's challenging in vitro behaviour, and this would be greatly facilitated by identification of stable, well-expressing OTR mutants which could also be crystallised.

Here we present a strategy that successfully combines SaBRE and CHESS by bridging the two distinct evolution strategies with an additional series of selection steps to produce stable and highly expressing human OTR variants. First, we applied the SaBRE method, selecting for well-expressing OTR variants in a eukaryotic host (*S. cerevisiae*) from a randomised library of the wild-type OTR (wtOTR) gene. Then, we switched to *E. coli* as the selection host. We pursued several rounds of selection for functional expression and stability in *E. coli* prior to applying the CHESS approach, where variants are selected for stability in detergent. Crucially, CHESS would not have been possible with a direct transition from SaBRE-selection in *E. coli* for functional expression is an essential intermediate step.

To investigate exactly why these additional steps are necessary, we wished to trace the directed evolution across the eukaryotic and prokaryotic host. We wanted to use next-generation sequencing (NGS), but to do so we had to overcome the challenge of requiring extremely long reads (the whole ORF of the GPCR) with extremely high accuracy. For this purpose, we developed a high-throughput sequencing pipeline based on single-molecule real-time (SMRT) sequencing with careful intrinsic control of technical sequencing read errors. The long reads generated by this particular NGS approach allowed sequencing of full-length OTR variants, and therefore information of mutational linkage was maintained. Overall, we generated more than 55,000 unique sequences, which allowed the identification of critical mutations in the OTR gene important for the overall selection success.

Results

Evolution of well-expressing OTR variants. It is not uncommon for GPCRs to have poor biophysical properties. One such GPCR is the human OTR, which has very low intrinsic stability and especially low functional expression levels^{29,35}. When wild-type OTR (wtOTR) is overexpressed in *E. coli* or *S. cerevisiae*, no surface expression is detected in flow cytometry experiments (data not shown). In fact, wtOTR even appears to be toxic to *E. coli* cells, even when expressed under optimised conditions, i.e., low-copy number plasmid, low temperatures, and with N- and C-terminal fusion proteins to help translocation of the overexpressed receptor to the inner membrane. The toxicity of OTR overexpression to *E. coli* might be related to the lack of cholesterol in *E. coli* membrane, which is equally important for OTR stability and integrity^{36,37}. The poor biophysical behaviour of the wtOTR as a starting structure and its biological importance make it an ideal target to employ previously developed engineering strategies for increasing expression and stability, such as CHESS²².

However, to perform CHESS with the OTR, it has first to be expressible in *E. coli* at least at some modest level, which is not the case. Thus, to obtain receptor variants that can be expressed in *E. coli*, we chose a successive selection strategy starting with SaBRE, using yeast as a eukaryotic expression host, before switching to a prokaryotic expression system. For selection with SaBRE, we randomised the wtOTR gene with an error-prone polymerase chain reaction (epPCR). *S. cerevisiae* was subsequently transformed with the created DNA library (SaBRE 1.0). The library was expressed, and surface expression was probed with the fluorescently labelled peptide antagonist PVA³⁸ (HiLyte Fluor 647-Lys⁸ PVA). However, no specific signal was observed for the naïve library, even at several different ligand concentrations tested, well above the dissociation constant of the ligand (Fig. 1a, Supplementary Fig. S1). On the contrary, we observed an increasing nonspecific signal at increasing fluorescent ligand concentrations, indicating that high concentrations of labelled ligand complicate the selection process.

Therefore, the highest concentration of fluorescent ligand where single specific events were still observed in flow cytometry was used to probe receptor expression levels during selection. In the first round of SaBRE

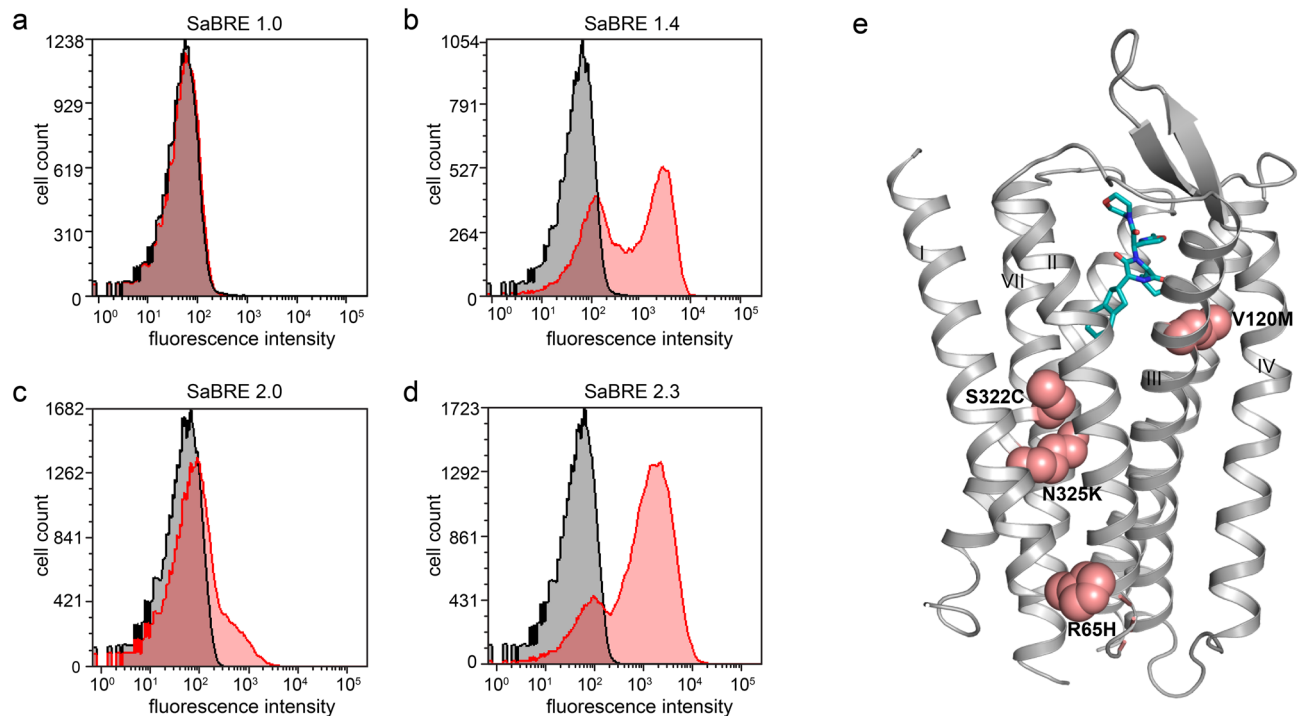


Figure 1. Identification of well-expressing OTR variants with SaBRE. **(a,b)** Histogram plots of fluorescent ligand binding data measured by flow cytometry. The total signal (red curves) and nonspecific signal (in the presence of unlabelled ligand, black curves) are shown for the starting pool SaBRE 1.0 **(a)** and the final pool SaBRE 1.4 **(b)** of the first round of SaBRE. **(c,d)** Histogram plots of fluorescent ligand binding flow cytometry data with total signal (red curves) and nonspecific signal (black curves) are shown for the starting pool SaBRE 2.0 **(c)** and the final pool SaBRE 2.3 **(d)** of the second round of SaBRE. **(e)** Mutations (pink spheres) of the single clone, OT-y01, identified in pool SaBRE 1.4, mapped on the OTR structure (grey, PDB ID: 6TPK) in complex with retosiban (turquoise).

(Supplementary Fig. S2a), a very stringent gating strategy was pursued to maximise selection of specific binding events. After three consecutive sorting steps comprising re-cultivating sorted yeast cells, re-expression of OTR variants, and selection with fluorescence-activated cell sorting (FACS), a pool was enriched (SaBRE 1.4) that showed significantly increased surface expression (Fig. 1b). From this pool, plasmid DNA was isolated and multiple clones were sequenced, but only one OTR variant (termed OT-y01) was identified multiple times. This single enriched clone contained five amino acid point mutations (A19^{N-termT}, R65^{1.58H}, V120^{3.33M}, S322^{7.46C} and N325^{7.49K}; Ballesteros and Weinstein numbering³⁹ denoted as superscript). With the exception of A19T, all mutations are located within transmembrane (TM) helix interfaces and most likely contribute to improved helix packing (Fig. 1e).

To generate a more diverse starting point for selection in *E. coli* or CHES, an additional round of SaBRE was conducted (Supplementary Fig. S2). OT-y01 was randomised once again by epPCR and *S. cerevisiae* was transformed with the created library (SaBRE 2.0). This library already showed an increased expression level when probed in flow cytometry compared to the initial library SaBRE 1.0 (Fig. 1a,c). In total, four subsequent steps of enrichment were carried out after the second randomization (Fig. 1d, Supplementary Fig. S1a). The plasmid DNA from the last two rounds of SaBRE (SaBRE 2.3 and SaBRE 2.4) were isolated, and 48 clones from each pool were sequenced and analysed. Sequencing determined that the SaBRE 2.3 pool has greater diversity (43 individual variants, with the most enriched clone appearing twice) compared to the SaBRE 2.4 pool (30 individual clones, with the most enriched clone appearing nine times). The more diverse pool, SaBRE 2.3, was chosen as a starting point for switching the expression host from yeast to bacteria to allow the broadest possible sampling of variants.

For this purpose, yeast-evolved OTR variants enriched in pool SaBRE 2.3 were isolated and subsequently *E. coli* was transformed with these constructs. However, only a poorly specific binding signal could be detected in *E. coli* for variants of the SaBRE 2.3 pool (Fig. 2a,b). While now functional expression was observed, the switch from *S. cerevisiae* to *E. coli* proved more difficult than expected. Unsurprisingly, no specific binding signal was consequently detected for variants enriched from the SaBRE 2.3 pool when probed in CHES. In this method, the *E. coli* cells of the library are first encapsulated by a polymer layer, such that each cell is wrapped individually and subsequently solubilised in detergent (Fig. 2e,f). Thereby, each cell is converted into a nanoscopic dialysis bag.

E. coli. OTR variants isolated from the SaBRE 2.3 pool were randomised a third time by epPCR and *E. coli* was transformed with this library (termed *E. coli* 3.0). The *E. coli*, *2.E. coli*, a notable increase in surface expression was observed (*E. coli*, *2d*). This pool also showed a clearly delineated binding signal shift when used in CHES^[22] (Fig. 2, *E. coli*, *E. coli*).

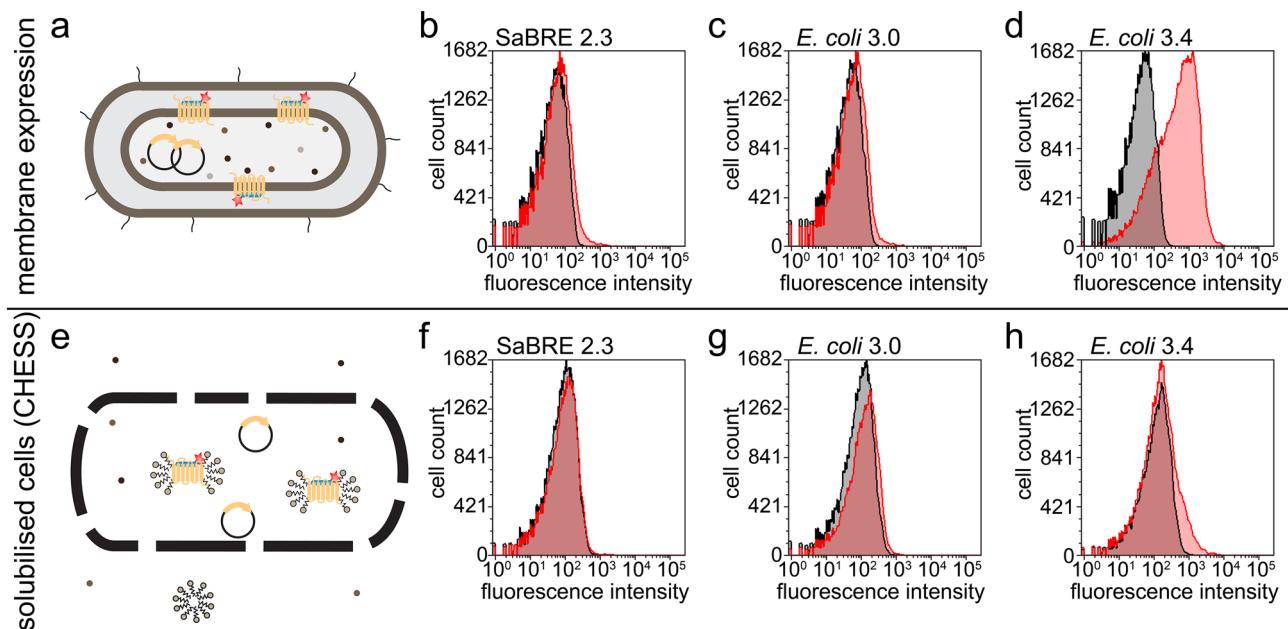


Figure 2. Selection for high functional expression and detergent stability in *E. coli*. (a) Schematic representation of the probing with fluorescently labelled (red star) ligand (blue rod) for well-expressing OTR variants (yellow rods) in the membrane of intact *E. coli* cells. (b,d) Histogram plots of flow cytometry detecting binding of fluorescent ligand on intact *E. coli* cells. Total signal (red curves) and nonspecific signal (black curves) are shown for the pools SaBRE 2.3 (b) *E. coli* 3.0 (c) and *E. coli* 3.4 (d), expressed in *E. coli*. (e) Schematic representation of a polymer-encapsulated (black lines) *E. coli* cell, whose internal space is subsequently solubilised with detergent, as performed in CHES, converting the cells essentially to nanoscopic dialysis bags. (f,h) Histogram plots of flow cytometry detecting binding of fluorescent ligand within polymer capsules. Total signal (red curves) and nonspecific signal (black curves) are shown for the pools SaBRE 2.3 (f) *E. coli* 3.0 (g) and *E. coli* 3.4 (h), after *E. coli* expression, encapsulation and solubilisation.

Next-generation sequencing of selection pools. To analyse the mutations enriched in single OTR variants during selection, we initially determined the sequence of 48 clones (by Sanger sequencing) from pools SaBRE 2.3, SaBRE 2.4, *E. coli* 3.4 and *E. coli* 3.5. Surprisingly, all of the sequenced pools proved to be diverse. For example, the SaBRE 2.3 pool had a total of 43 unique sequences identified (38 singly occurring clones and 5 doubly occurring clones). High diversity precludes comprehensive analysis by Sanger sequencing, due to its small sampling size. To facilitate sequence analysis of the diverse pools enriched by evolution, we developed an NGS sequencing strategy for this purpose.

We required a sequencing strategy that fulfilled two main criteria: (1) the sequencing method must provide very accurate single reads to allow the identification of point mutations in a single variant, and (2) the sequencing platform must generate reads long enough to ensure sequencing of a single complete OTR gene in one stretch (read lengths > 1000 base pairs (bp)), thus allowing not only the identification of mutations but also capturing their co-occurrences. The method best matching these criteria was single-molecule real-time (SMRT) sequencing, offered by PacBio. In a SMRT sequencing experiment, single circularised gene variants are sequenced in individual wells, constituting zero mode waveguides (ZMW). A rolling circle amplification approach is used in which the complementary strand is synthesised by an immobilised polymerase, with fluorophore release upon nucleotide incorporation, achieving read lengths of up to 100,000 bp⁴⁰. A single circularised OTR variant (1056 bp) present in a ZMW should result in a read with multiple passes (subreads) because of its comparably small length compared to the maximum read length of the rolling circle amplification approach. These subreads should later allow generation of accurate circular consensus sequencing (CCS) reads of the OTR variants present in the according ZMW.

The individual pools were simultaneously sequenced on a single SMRT cell using a PacBio Sequel instrument. A total of 624,145 ZMW reads were generated with a mean read length of 23,661 bp and a median subread length of 1,112 bp, resulting in a mean subread number of 20.6 (Fig. 3b,c). During data analysis we found, analogously to previous reports⁴¹, that individual subreads contained a high number of insertions and deletions (indels). These indels are not an expected by-product of our evolution strategy (and none have been found by Sanger sequencing from the same pool and other pools from other evolution projects), but rather a direct consequence of the sequencing approach. As incorrectly identified indels critically impact the quality of the sequence analysis by shifting the reading frame of the sequenced gene, we developed a novel analysis approach.

To overcome the low quality of the subreads and generate consensus sequences from single ZMW outputs with high accuracy, we implemented a processing strategy that included five steps (Fig. 3a). First, only reads with at least 20 subreads were considered. Second, the single subreads were aligned to the wtOTR gene sequence. The aligned subreads were then used to generate consensus sequences. Consensus sequences with incorrect read

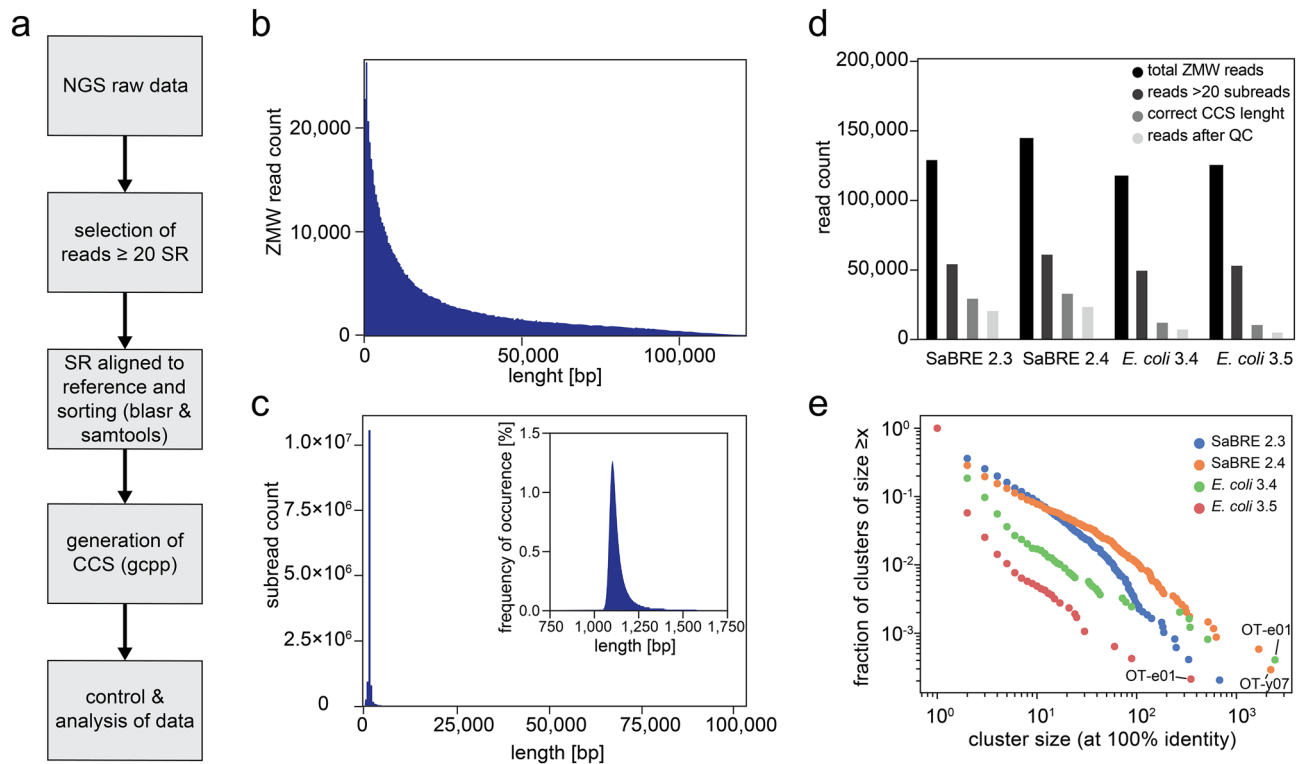


Figure 3. Single-molecule real-time sequencing of OTR selection pools. **(a)** NGS data processing workflow. **(b)** Histogram representation of all reads and the respective read lengths generated during SMRT sequencing. **(c)** Histogram representation of all subreads and the respective subread lengths generated during SMRT sequencing. Blow-up of the main peak is depicted in the top right corner. **(d)** Diagram of read counts retained after the indicated processing steps (top right corner) for all four pools sequenced. **(e)** Scatter plot of the cluster sizes and the fraction of clusters observed within the indicated (top right corner) sequence pools. The cluster size represents the number of times a particular unique sequence is observed. The fraction of clusters is obtained by counting all clusters of size x divided by the total number of clusters.

lengths were excluded from analysis, as we did not expect our evolution strategy to produce indels. In a final control step, consensus sequences were translated and the top 1% of sequences with the highest mutation rates, typically caused by a frameshift in the open reading frame, were excluded. Applying this stringent protocol on the raw sequencing data resulted in a retention rate of around 10% (Fig. 3d).

To test the validity and accuracy of the obtained sequences generated by this approach, we compared these NGS sequences with the sequences obtained from Sanger sequencing. For all four pools sequenced, we found almost identical mutation rates per gene and similar occurrences of single point mutations and amino acid changes on the protein level (Supplementary Fig. S3) when comparing the two sequencing methods. Furthermore, we found in the datasets of NGS sequences and Sanger sequences the same clones to be most enriched for both final pools (SaBRE 3.4 and *E. coli* 3.4) (Table S3). We thus concluded that the NGS sequences so obtained and processed had the same quality as the sequences obtained from Sanger sequencing.

Additional variations to the data processing pipeline, such as different software for alignment and consensus calling as well as the implementation of more stringent cut-offs, were tested without substantially improving the retention rate or accuracy (compared to Sanger sequencing).

Despite only 10% of all ZMW reads resulting in translatable consensus sequences of high accuracy, we still obtained 8000–20,000 sequences for each individual pool and thus a total of 58,050 sequences from one NGS run.

In-depth analysis of enriched selection pools. Our NGS approach dissected a large number of variants suitable for an in-depth analysis of the selection outcome. We found a median mutational rate of 9 amino acid changes per OTR variant after SaBRE, and a median mutational rate of 12 amino acid changes per OTR variant in the *E. coli*-derived selection pools. Certain positions were enriched for mutations across the entire pool, independent of the specific OTR variant or of other mutations. Preponderance of single amino acid mutations at a particular position is likely due to their beneficial effect on the biophysical properties of the receptor. These mutations and positions were identified by calculating the frequency of the most dominant non-wt ($f_{\text{non-wt}}$) amino acid occurring at every position on the OTR gene (Fig. 4, Supplementary Fig. S3).

Five amino acid positions are observed to all have single non-wt amino acid in 100% of analysed sequences. These mutations (A19^{N-term}T, R65^{1.58}H, V120^{3.33}M, S322^{7.46}C, and N325^{7.49}K) stem from the single clone, OT-y01, used as the source for the second round of SaBRE and, while their occurrence is not surprising, their complete uniformity confirms the fidelity of NGS processing pipeline. In the selection outcome of SaBRE, the most

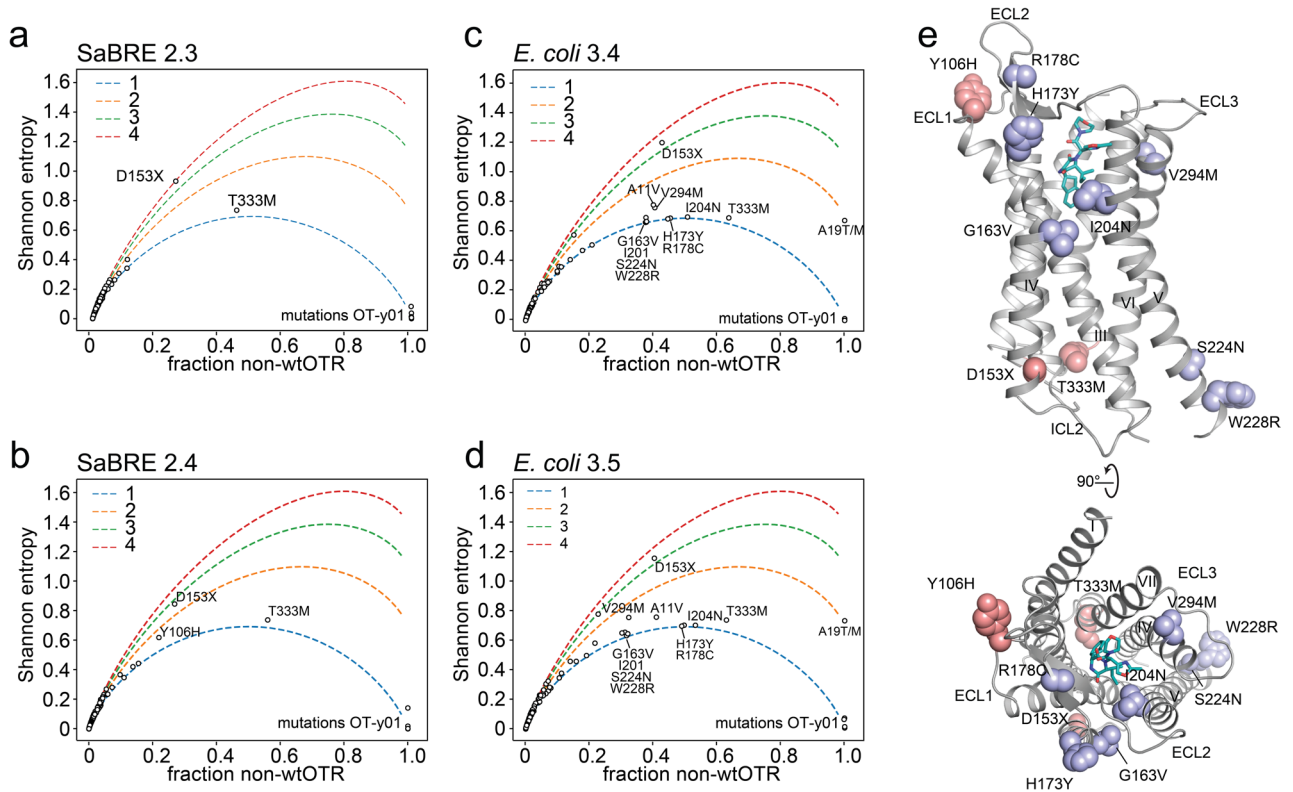


Figure 4. Observed mutational frequencies. (a–d) Shannon entropy calculated with Eq. (1) for every individual amino acid position (black circles) on the OTR, plotted against fraction of non-wtOTR amino acid occurrence. To help visualise the theoretical Shannon entropy for $n = 1$ (blue), $n = 2$ (orange), $n = 3$ (green) and $n = 4$ (red), non-wt amino acids observed are depicted as dashed lines. Individual positions being mutated with a higher frequency than the most enriched clone of a pool are indicated with a label, as they must have been independently selected several times. (e) Amino acids positions with a high non-wtOTR frequency are mapped on the OTR (grey, PDB ID: 6TPK) bound to retosiban (turquoise). Mutations occurring at a high frequency after SaBRE are coloured pink, positions mutated at a high frequency after *E. coli*-based selection are coloured light purple.

frequently observed non-wt mutations are D153^{4.42}N and T333^{8.47}M, occurring at a frequency of 21% and 56%, respectively (Fig. 4a,b). Mutations in both these positions, D153^{4.42} and especially T333^{8.47}, occur more frequently than the top three single clones (accounting for 19.5% of pool diversity in SaBRE 2.4, Supplementary Table S3), indicating their phenotypic selection in different sequence backgrounds. Both positions are located at the ends of transmembrane helices (TM 4 and TM 7, respectively) and both point into the receptor core (Fig. 4e). We hypothesise that these mutations might optimise helix–helix packing interactions. Our hypothesis is further strengthened by the observation these interactions are independent (i.e., the frequency of each observed amino acid pair is almost equal to the product of the individual frequencies, Supplementary Fig. S5) and this independence is not system-dependent—it is observed throughout the evolution process in *E. coli* and in *S. cerevisiae*.

In the final pool of the selection in *E. coli* (*E. coli* 3.4), position H173^{4.62}Y (45%), R178^{EL2}C (45%), and I204^{5.42}N (51%) all occur at a higher frequency (Fig. 4c) than the most enriched clones of the corresponding pools (top three clones account for 41.1%, Supplementary Table S3), again underlining that they have been phenotypically selected in different backgrounds. I204 is pointing into the binding pocket of the receptor (Fig. 4e) and has previously been shown to be crucial for functional integrity of the binding pocket⁴². Mutations in this position could possibly be caused by an accompanying affinity maturation for the fluorescent ligand used for selection. Interestingly, mutations in position H173 and R178 both occur linked together, yet never in combination with a mutation in position D153, which otherwise occurs frequently, possibly indicating that they together would be unfavourable. H173 is located at the end of TM helix 4, and R178 is located in the β -hairpin formed by extracellular loop 2 of the receptor. All three positions (H173, R178 and I204) were not mutated at a high frequency in SaBRE, indicating that the positions might be crucial especially for increased production in *E. coli*.

While these observed mutational frequencies indicate how frequent the most popular non-wt amino acid is at a given position of the OTR, the data do not hold any information on the degeneracy of non-wt amino acids at that position (i.e., they do not describe whether or not there are multiple new residue types observed). This metric is captured by the Shannon entropy, which can be calculated from the sequencing data. We calculated the Shannon entropy for each amino acid position along the OTR gene (Fig. 4a–d). To provide context for the less-than-intuitive entropy values, we additionally plotted hypothetical curves for the Shannon entropy assuming a uniform distribution for a given number of non-native amino acids. This formula (Eq. (1)) depends only

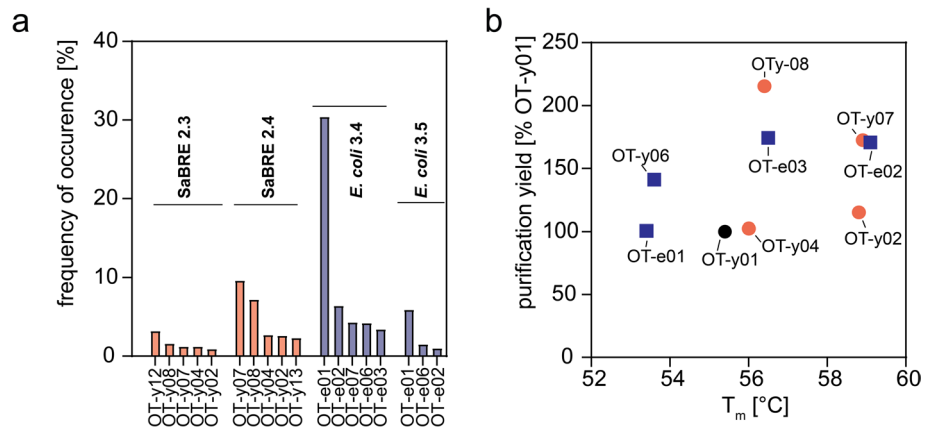


Figure 5. Characterization of most enriched single clones. **(a)** Occurrence of the most enriched variants of all four pools sequenced as determined by NGS. Variants identified in SaBRE pools (OT-yXX) are coloured in salmon, and variants identified in pools enriched during *E. coli*-based selection (OT-eXX) are coloured blue. Sequences were numbered according to their occurrence in initial Sanger sequencing results. **(b)** The purification yield of OTR purified from HEK293 cells, normalised to OT-y01, is compared to the apparent T_m derived from CPM-based thermostability measurements.

on the frequency of the wt amino acid (f_{wt}) in the respective position (i) and the total number (n) of non-wt amino acids observed.

Adapted Shannon entropy (S):

$$S = -f_{wt} \times \ln(f_{wt}) - \sum_1^n \frac{1-f_{wt}}{n} \times \ln\left(\frac{1-f_{wt}}{n}\right). \quad (1)$$

By comparing the calculated values to the theoretical curves, we can determine, at a glance, roughly how many different mutations are observed at each position. For most mutated positions only a single non-wt amino acid is observed (because of the low mutation rate) and thus the observed Shannon entropies fall along the theoretical line for $n = 1$ (Fig. 4). This outcome is expected, based on the underlying statistical nature of library creation by epPCR resulting in few mutations, and the accompanying limitation posed by codon bias, limiting the type of codon changes.

Nonetheless, exceptions were observed. For example, one of the few exceptions with a high entropy, i.e., a position with multiple different amino acids occurring, was position D153 (Fig. 4). Position D153 was found to be mutated to either tyrosine, asparagine, histidine, valine and even glycine. Therefore, in position D153 five out of the seven amino acids which are accessible with just one base pair change to the respective codon are observed. This high entropy emphasizes the detrimental effect the aspartate in position D153 apparently exerts on receptor integrity, as it seems that any other amino acid is preferred at this position.

To capture the diversity of each pool, we calculated the cumulative density function of cluster size (Fig. 3e). Cluster size here is defined as the number of observations of each unique sequence. Unsurprisingly, we observed a negative correlation between cluster size and frequency of occurrence, i.e., there are many observations of unique single sequences and only few observations of clusters of many identical sequences (Fig. 3e). For all NGS pools, there is a single, dominant cluster with ~ 100 – 1000 observations of one sequence, clearly indicating enrichment by the selection methods.

The SaBRE 2.4 pool is more enriched than the SaBRE 2.3 pool, as demonstrated by a higher fraction of large clusters and simultaneously a drop in the frequency of small clusters (Fig. 3e). Interestingly, after one sorting round with CHES, the enriched pool *E. coli* 3.5 appears to be *more* diverse than its parental pool *E. coli* 3.4. This might be an effect caused by the new selection pressure applied during sorting for stability with CHES and the subsequent need to amplify the receptor gene by PCR (introducing more mutations), as re-cultivation of course is impossible from the permeabilised bacterial cells.

To finally assess purification yield and stability of the most enriched single OTR variants, they were expressed in HEK293T cells, and subsequently solubilised and purified. We probed purification yield and assessed thermostability by monitoring protein unfolding as a function of temperature using the previously described 7-diethylamino-3-(4-maleimidophenyl)-4-methylcoumarin (CPM) assay⁴³ (Fig. 5, Supplementary Table S3). Because wtOTR could not be purified in sufficient yield, we could not determine its melting temperature (T_m). Therefore, we selected the parental variant OT-y01 as reference for all tested OTR clones. As described previously, we found a trend that higher purification yield correlates with higher stability, even though this correlation is not very strong. Most tested variants show higher thermostability and expression yield compared to OT-y01, which indicates a clear selection for these properties.

Interestingly, we also observed that purification yields from mammalian cells does not seem to correlate with enrichment after SaBRE or selection in *E. coli*. The most stable clone tested was OT-e02 with a melting

temperature of 59.2 °C and OT-y02 with a melting temperature of 58.8 °C. Noteworthy, OT-y02 is also the clone that served as the basis for the construct later successfully crystallised in LCP⁴².

Discussion

GPCRs commonly require protein engineering to enable detailed biophysical studies and structure determination. While several receptors were sufficiently stabilized by using a particular directed evolution strategy, the most challenging receptors could not. The otherwise intractable human OTR serves as an example.

Among the available engineering strategies, CHES, the selection method for stability in detergent, has proven to be an especially powerful technique for the identification of suitable receptor mutants. While mutants, stabilized through mutagenesis, can display conformational restraints compared to wt receptors¹⁴, their great potential is underlined by CHES-improved GPCR mutants utilised in challenging biophysical and ligand-binding studies, including SPR^{18,19}, NMR^{44–46} and X-ray crystallography⁴⁷. However, until now, the limiting factor for applying CHES has been that it requires some level of functional protein expression in *E. coli*.

Here we presented a strategy to make the otherwise intractable human OTR accessible to CHES by a strategic combination of two directed evolution approaches: starting with selection for improved functional expression in yeast (SaBRE), followed by a subsequent selection for functional expression in *E. coli*, followed by CHES. After transferring the SaBRE-selected clones to *E. coli*, we observed that the variants were still not expressing well in *E. coli*, precluding the direct application of CHES.

In contrast, higher functional expression was retained when switching in the opposite direction, bringing mutants selected in *E. coli* or *S. cerevisiae* to a higher eukaryotic host, such as insect cells or mammalian cells^{21,24}. A possible explanation for this observation might be the difference of protein quality-control machinery. Thus, a protein that is only able to fold in the presence of a complex machinery in higher eukaryotic cells would not fold in cells lacking that machinery, while the converse is not true, and a biophysically improved mutant may still show improvements in the presence of a more complex machinery.

With the goal of ultimately applying CHES, which takes place in *E. coli*, the yeast-evolved variants provided an excellent starting point for subsequent selection for high expression in *E. coli*, while the wtOTR did not. We are confident that this strategy can be further adapted not only for other GPCRs, but also other membrane proteins that are initially toxic to *E. coli*. Such a stepwise strategy may be needed when membrane proteins need to be expressible in *E. coli*, e.g., for isotope labelling, or when they are to be examined in experiments which require long-time stability in a given detergent.

As we sought to determine the sequence features that led to enhanced stability in our evolution experiments, we turned to SMRT sequencing, an NGS approach that allowed us to sequence the entire 1056 bp target gene, and thus without losing the connections between mutations distant in the sequence. For this purpose, we developed a sequencing pipeline for the analysis of SMRT-sequencing-derived NGS libraries with very high accuracy, such that the very few true point mutations were clearly distinguishable from technical sequencing errors. With this approach we were able to acquire over 55,000 unique high-accuracy sequences from four different pools. By comparing sequences obtained from NGS with sequences obtained by Sanger sequencing, we were able to validate their correctness.

The limiting factor for the overall number of sequences obtained by our NGS approach was the low subread sequence quality, caused by the high occurrence of indels from SMRT sequencing. Consequently, we found the key to correctly assembling circular consensus sequencing reads was to first align each subread to the wt gene before generating the consensus sequence. Multiple strategies, programs, and parameters were empirically screened to find the best working solution. We restricted our approach to the use of pre-existing software, as the challenge of correctly identifying variants from indel-heavy data is an area of active research⁴⁸ and beyond the scope of this manuscript.

To our knowledge, this is the first report of a long-read sequencing technology implemented to analyse DNA pools from directed evolution, where every single DNA molecule contains individual information. In future experiments, our NGS approach could simplify and accelerate the analysis of any diversified DNA pool. With the ability to simultaneously generate and analyse high sequence numbers, very quantitative and complete analyses of diverse pools become possible. Furthermore, this sequencing approach could provide a suitable platform to fully exploit selections based on rationally designed synthetic libraries, e.g., by removing the need for DNA barcodes for sequence identification in deep mutational scanning studies⁴⁹.

Our analyses of the NGS-derived sequences allowed us to identify several co-enriched positions during the distinct evolution rounds. We observed D153X (where X = N, Y, V, G, & H, ordered according their observed frequency) and T333M to be present in several individual OTR variants enriched in yeast. Because of their high frequency of appearance, we reasoned that these mutations are particularly critical to the *S. cerevisiae*-based selection outcome. Conversely, in the sequences of *E. coli*-based OTR variants, we frequently observed either the co-occurrence of mutations H173Y and R178C or the mutation I204N. Interestingly, none of these three mutations are facing towards the helix bundle. Simultaneously, none of these three mutations were frequently observed in yeast-derived OTR variants. We thus speculate that these three mutations contribute to increased functional expression of the receptor in the *E. coli*, even though we cannot pinpoint yet the exact step where the optimized OTR performs better in the more demanding expression environment in *E. coli*.

With few exceptions, we found that most positions had only mutated to one other type of amino acid. This is most likely due to a fundamental limitation in exploring sequence space by epPCR, which will typically explore only explore codon changes accessible by a one-base mutation. As previously reported by Schlinkmann et al.^{26,50}, comprehensive synthetic libraries containing all desired codons at given positions are needed to optimise the directed evolution outcome. With key positions identified in our selections and the subsequent elucidation of the OTR structure⁴², the rationalization of such binary libraries for the OTR as well as the closely related vasopressin

receptors⁵¹ can now be successfully implemented. The transferability of these critical mutations from the OTR to vasopressin receptors could be directly assessed by grafting the corresponding mutations into the vasopressin receptors, similarly to the previously reported successful transfer of the SaBRE-derived mutations from the κ -opioid receptor²⁴ to the δ -opioid receptor⁵².

In conclusion, while GPCRs with especially poor properties cannot be successfully optimised even with individual directed evolution technologies, we could show that an engineering strategy, based on the subsequent use of functional expression in yeast (SaBRE), followed by *E. coli*, and then the use of the especially powerful CHES was able to solve this problem. The successful selection for improved and crystallisable variants of the human oxytocin receptor, a GPCR with particularly low intrinsic production levels, serves as an example. Our study has also shown, by the different nature of enriched mutations, that the evolution pressure on the same membrane protein in prokaryotes and eukaryotes is different. The very accurate long-read NGS strategy that we have developed for this purpose provides a general methodology for the highly accurate analysis of libraries of point mutants during directed evolution.

Methods

DNA library construction and transformation. Human OTR (2–389) was randomised using the GeneMorph II random mutagenesis kit (Agilent), aiming for a mean of two amino acid changes per gene. Two reactions with 200 ng template DNA were either amplified for 20 or 25 cycles, respectively. The created libraries from both reactions were column-purified (QIAquick PCR purification kit, Qiagen) and combined at a 1:1 molar ratio.

S. cerevisiae strain BY4741⁵³ (EUROSCARF) was transformed as previously described²⁴. In brief, sites homologous to the linearised yeast expression vector pMS003het²⁴ (*NheI/BamHI* digested) were added at each end of the OTR gene during epPCR by the primers, to allow homologous recombination in yeast. To obtain sufficient DNA yields for transformation, the created OTR libraries were further amplified by standard PCR. BY4741 cells were transformed by square wave electroporation on a GenePulser Xcell electroporator (BioRad) analogously to a previously published method⁵⁴. After electroporation, cells were allowed to recover in YPD medium without shaking at 30 °C before pelleting and resuspending in 500 ml SDD-Leu-medium (Synthetic Defined Medium without Leu; 6.9 g L⁻¹ yeast nitrogen base without amino acids (Formedium), 690 mg L⁻¹ complete supplement mixture without leucine (Formedium), 20 g L⁻¹ glucose, 35 mM sodium citrate tribasic, 35 mM citric acid) for selective growth at 30 °C for 24 h, and finally stored as glycerol stocks at – 80 °C until use. Libraries had a diversity of 5×10^7 (SaBRE 1.0) and 7×10^7 (SaBRE 2.0).

For efficient library creation in *E. coli*, restriction sites (*BamHI* and *SpeI*) were introduced by the primers during amplification to both ends of the created library. To obtain the plasmid pEC01, pRG/III⁵⁵ was modified by replacing the C-terminal thioredoxin and the polyhistidine-tag by an SGSGGGSG linker, followed by superfolder Green Fluorescent Protein (sfGFP)⁵⁶ followed by a C-terminal Avi-tag. The amplified and digested library (*BamHI* and *SpeI*) was ligated into linearised pEC01 overnight at 16 °C. Ligated vector was purified (QIAquick PCR purification kit, Qiagen) and eluted in distilled H₂O. Electrocompetent DH5 α cells (Thermo Fisher) were transformed with clean and concentrated ligation product by electroporation. Directly after electroporation, cells were allowed to recover in 4 mL SOC medium under shaking at 37 °C for 1 h. Thereafter, ampicillin (100 ng μ L⁻¹) was added, and cells were allowed to grow for another 2 h. To remove non-transformed cells, which fail to divide under ampicillin selection and grow to aggregating filaments instead, the recovered cells were passed through a 5- μ m filter (Millipore). Cells collected from the flow-through were further incubated in 40 mL fresh 2YT containing 7% sucrose (w/v), 1% glucose (w/v) and 100 ng μ L⁻¹ ampicillin for 17 h and finally stored as glycerol stocks at – 80 °C. The total library size (*E. coli* 3.0) was 4×10^6 as estimated from determining colonies in dilution series.

Fluorescent ligand binding with yeast cells. A pre-culture was inoculated to OD₆₀₀=0.2 in Synthetic Defined Medium without Leu (SDD-Leu⁻) and cultivated for 16–18 h at 30 °C. For expression, yeast cells from the pre-culture were collected and resuspended in SDG-Leu⁻ medium (identical to SDD-LEU⁻ with galactose instead of glucose) to an OD₆₀₀=1.0 and incubated for 24 h at 20 °C. After expression, 5×10^8 cells were collected and washed once in TELi (50 mM Tris-HCl pH 9.0 (at 4 °C), 100 mM lithium acetate, 1 mM EDTA), before resuspending in TELi supplemented with 50 mM DTT. To allow efficient permeabilization of the cell wall, cells were incubated at 20 °C for 30 min and finally washed twice in ice cold TELi. 1×10^8 washed and permeabilised yeast cells were collected and resuspended in 2 mL TELi supplemented with 50 nM HiLyte Fluor 647-Lys⁸ PVA (Eurogentec) for 2 h on ice without exposure to light. To assess unspecific binding, corresponding samples were additionally supplemented with 2.5 μ M unlabelled PVA as competitor. Before measuring flow cytometry, cells were collected by centrifugation and resuspended in 1 mL TELi.

Fluorescent ligand binding with bacterial cells. A pre-culture was inoculated to an OD₆₀₀=0.05 in LB medium containing ampicillin (100 ng μ L⁻¹) and 1% glucose (w/v) and cultivated for 2–3 h at 37 °C. The pre-culture was subsequently used to inoculate 5 mL 2YT containing 0.2% glucose and 100 ng μ L⁻¹ ampicillin to an OD₆₀₀=0.05. When the expression culture had grown to OD₆₀₀=0.5–0.6, expression was induced by addition of 250 μ M IPTG and was allowed to proceed for 20 h at 20 °C. Then, 10^9 cells were collected and washed twice with 1 mL ice-cold TKCl (50 mM Tris-HCl, 150 mM KCl, pH 7.4). 10^7 washed cells were incubated in 250 μ L TKCl containing 100 nM HiLyte Fluor 647-Lys⁸-PVA for 2 h on ice without exposure to light. To assess unspecific binding, corresponding samples were additionally supplemented with 2.5 μ M unlabelled PVA for competition. Before measuring flow cytometry, cells were collected by centrifugation and resuspended in 200 μ L TKCl.

Fluorescent ligand binding with polymer-encapsulated solubilised bacterial cells. 1×10^{10} receptor-expressing *E. coli* cells were collected and encapsulated with one layer of chitosan and one layer of alginate, strictly following the procedure published previously²². Encapsulated cells were solubilised in PBS-ED (137 mM NaCl, 2.7 mM KCl, 8.1 mM Na₂HPO₄, 1.8 mM KH₂PO₄, pH 7.4, 1 mM EDTA, cComplete protease inhibitor EDTA-free tablets (Roche), 1% *n*-dodecyl- β -D-maltopyranoside (DDM, Anatrace) (w/v), 0.1% cholesterol hemisuccinate (CHS, Sigma Aldrich) (w/v)) containing 50 nM HiLyte Fluor 647-Lys⁸-PVA (binding buffer) or 50 nM HiLyte Fluor 647-Lys⁸-PVA and 2.5 μ M unlabelled PVA (competition buffer) at 4 °C for 1 h. Solubilised cells, i.e., polymer capsules, were washed once with either binding or competition buffer. Washed capsules were incubated for another 3 h in the respective buffer. Before carrying out flow cytometry measurements, capsules were collected by centrifugation and resuspended in 500 μ L PBS-ED.

Flow cytometry and FACS. Flow cytometry was performed on a BD FACSCanto II cytometer (BD Biosciences) and fluorescence-activated cell sorting was performed on a BD FACSAria III sorter (BD Biosciences). During FACS, the top 0.5% of the fluorescent events were retained. Cells were directly sorted into growth medium and subsequently cultivated to stationary phase at 30 °C (yeast) or 37 °C (bacteria), respectively. Capsules were sorted into distilled H₂O and the DNA encoding OTR was amplified by PCR. Amplified DNA was re-cloned as described above. Flow cytometry data were analysed with FCS Express 5 Flow (De Novo Software).

Thermostability measurements. HEK293T/17 cells (ATCC) were cultivated in Dulbecco's modified medium (Sigma Aldrich) supplemented with 100 units/mL penicillin, 100 μ g/mL streptomycin (Sigma Aldrich) and 10% (v/v) fetal calf serum (BioConcept). Cells were maintained at 37 °C in a humidified atmosphere of 5% CO₂, 95% air. Transient transfections of OTR containing an N-terminal fused FLAG-tag were performed with TransIT-293 (Mirus Bio) according to the manufacturer's protocol.

HEK293T cells were seeded and transfected in 6-well plates (Falcon) at a cell density of 1.5×10^6 cells per well. 48 h after transfection cells were harvested, washed once in PBS (pH 7.4) and stored at -20 °C. Purification of expressed OTR constructs was performed as published in detail (Schöppe et al., manuscript submitted). In brief, cells were lysed in hypotonic solution by resuspending in low salt buffer (10 mM HEPES pH 7.5, 20 mM KCl, 10 mM MgCl₂, 50 μ g mL⁻¹ Pefabloc SC (Carl Roth), 1 μ g mL⁻¹ Pepstatin A (Carl Roth)) and subsequently solubilised in solubilization buffer (low salt buffer supplemented with 500 mM NaCl, 0.5% DDM (w/v), 0.1% CHS (w/v), 200 μ M PVA). Insoluble cell debris was removed by centrifugation and the cleared supernatant was incubated with Anti-FLAG M2 magnetic beads (Sigma Aldrich) at 4 °C for 2.5 h. The receptor-bound resin was washed with 25 column volumes (CV) of Wash Buffer I (50 mM HEPES pH 7.5, 500 mM NaCl, 10 mM MgCl₂, 10% (v/v) glycerol, 1.0% (w/v) DDM, 0.2% (w/v) CHS, 8 mM ATP, 50 μ M PVA) followed by 35 CV of Wash Buffer II (50 mM HEPES pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 50 μ M PVA). PVA-bound OTR was eluted in one CV of Elution Buffer (50 mM HEPES pH 7.5, 500 mM NaCl, 3 mg mL⁻¹ 3 \times FLAG Peptide (Sigma Aldrich), 10% (v/v) glycerol, 0.05% (w/v) DDM, 0.01% (w/v) CHS, 50 μ M PVA). CPM dye was added to eluted pure PVA-bound OTR and thermostability was assessed on a Mx3005P QPCR System (Agilent) qPCR machine. Data were analysed GraphPad Prism software (version 8.1.1, GraphPad Prism).

SMRT sequencing. The OTR genes (2–353) enriched in pools SaBRE 2.3, SaBRE 2.4, *E. coli* 3.4 and *E. coli* 3.5 were amplified by PCR and individualised and simultaneously circularised by ligating a unique barcode adapter using the SMRTbell Express Template Prep Kit 2.0 (PacBio) and Barcoded Overhang Adapters (PacBio) (Supplementary Fig. S2b). Amino acids 354–389 of the OTR C-terminus were excluded from analysis to increase overall CCS read accuracy by shortening the subread length and thus consequently increasing the mean number of subreads achieved from a single read. Sequencing was performed on a PacBio Sequel instrument (Pacific Bioinformatics) using one 1 M SMRT cell (Pacific Bioinformatics) as service by the Functional Genomics Centre Zurich (FGCZ).

Received: 20 January 2021; Accepted: 5 April 2021

Published online: 21 April 2021

References

- Lagerström, M. C. & Schiöth, H. B. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat. Rev. Drug Discov.* **7**, 339–357 (2008).
- Katritch, V., Cherezov, V. & Stevens, R. C. Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* **53**, 531–556 (2013).
- Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B. & Gloriam, D. E. Trends in GPCR drug discovery: New agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842 (2017).
- Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5**, 993–996 (2006).
- Maeda, S. & Schertler, G. F. X. Production of GPCR and GPCR complexes for structure determination. *Curr. Opin. Struct. Biol.* **23**, 381–392 (2013).
- Serrano-Vega, M. J., Magnani, F., Shibata, Y. & Tate, C. G. Conformational thermostabilization of the β 1-adrenergic receptor in a detergent-resistant form. *PNAS* **105**, 877–882 (2008).
- Gupta, K. et al. MultiBac: Baculovirus-mediated multigene DNA cargo delivery in insect and mammalian cells. *Viruses* **11**, 198 (2019).
- Errey, J. C. & Fiez-Vandal, C. Production of membrane proteins in industry: The example of GPCRs. *Protein Express. Purif.* **169**, 105569 (2020).
- Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nat. Protoc.* **4**, 706–731 (2009).

10. Parker, J. L. & Newstead, S. Current trends in α -helical membrane protein crystallization: An update. *Protein Sci.* **21**, 1358–1365 (2012).
11. Thal, D. M. *et al.* Recent advances in the determination of G protein-coupled receptor structures. *Curr. Opin. Struc. Biol.* **51**, 28–34 (2018).
12. Chun, E. *et al.* Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. *Structure* **20**, 967–976 (2012).
13. Magnani, F., Shibata, Y., Serrano-Vega, M. J. & Tate, C. G. Co-evolving stability and conformational homogeneity of the human adenosine A2A receptor. *PNAS* **105**, 10744–10749 (2008).
14. Scott, D. J., Kummer, L., Tremmel, D. & Plückthun, A. Stabilizing membrane proteins through protein engineering. *Curr. Opin. Chem. Biol.* **17**, 427–435 (2013).
15. Bill, R. M. *et al.* Overcoming barriers to membrane protein structure determination. *Nat. Biotechnol.* **29**, 335–340 (2011).
16. Grisshammer, R. New approaches towards the understanding of integral membrane proteins: A structural perspective on G protein-coupled receptors. *Protein Sci.* **26**, 1493–1504 (2017).
17. Kühlbrandt, W. The resolution revolution. *Science* **343**, 1443 (2014).
18. Huber, S. *et al.* SPR-based fragment screening with neurotensin receptor 1 generates novel small molecule ligands. *PLoS ONE* **12**, e0175842 (2017).
19. Ranganathan, A. *et al.* Ligand discovery for a peptide-binding GPCR by structure-based Screening of fragment- and lead-like chemical libraries. *ACS Chem. Biol.* **12**, 735–745 (2017).
20. Igonet, S. *et al.* Enabling STD-NMR fragment screening using stabilized native GPCR: A case study of adenosine receptor. *Sci. Rep.* **8**, 8142 (2018).
21. Sarkar, C. A. *et al.* Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 14808–14813 (2008).
22. Scott, D. J. & Plückthun, A. Direct molecular evolution of detergent-stable G protein-coupled receptors using polymer encapsulated cells. *J. Mol. Biol.* **425**, 662–677 (2013).
23. Klenk, C., Ehrenmann, J., Schütz, M. & Plückthun, A. A generic selection system for improved expression and thermostability of G protein-coupled receptors by directed evolution. *Sci. Rep.* **6**, 21294 (2016).
24. Schütz, M. *et al.* Directed evolution of G protein-coupled receptors in yeast for higher functional production in eukaryotic expression hosts. *Sci. Rep.* **6**, 21508 (2016).
25. Dodevski, I. & Plückthun, A. Evolution of three human GPCRs for higher expression and stability. *J. Mol. Biol.* **408**, 599–615 (2011).
26. Schlinkmann, K. M. *et al.* Maximizing detergent stability and functional expression of a GPCR by exhaustive recombination and evolution. *J. Mol. Biol.* **422**, 414–428 (2012).
27. Scott, D. J., Kummer, L., Egloff, P., Bathgate, R. A. D. & Plückthun, A. Improving the apo-state detergent stability of NTS1 with CHESS for pharmacological and structural studies. *Biochim. Biophys. Acta Biomembr.* **1838**, 2817–2824 (2014).
28. Ehrenmann, J. *et al.* High-resolution crystal structure of parathyroid hormone 1 receptor in complex with a peptide agonist. *Nat. Struct. Mol. Biol.* **25**, 1086–1092 (2018).
29. Gimpl, G., Klein, U., Reiländer, H. & Fahrenholz, F. Expression of the human oxytocin receptor in baculovirus-infected insect cells: High-affinity binding is induced by a cholesterol-cyclodextrin complex. *Biochemistry* **34**, 13794–13801 (1995).
30. Donaldson, Z. R. & Young, L. J. Oxytocin, vasopressin, and the neurogenetics of sociality. *Science* **322**, 900–904 (2008).
31. Modahl, C., Fein, D., Waterhouse, L. & Newton, N. Does oxytocin deficiency mediate social deficits in autism? *J. Autism Dev. Disord.* **22**, 449–451 (1992).
32. Hollander, E. *et al.* Oxytocin infusion reduces repetitive behaviors in adults with autistic and Asperger's disorders. *Neuropsychopharmacology* **28**, 193–198 (2003).
33. Pedersen, C. A. *et al.* Intranasal oxytocin reduces psychotic symptoms and improves Theory of Mind and social perception in schizophrenia. *Schizophr. Res.* **132**, 50–53 (2011).
34. Andersen, L. F., Lyndrup, J., Åkerlund, M. & Melin, P. Oxytocin receptor blockade: A new principle in the treatment of preterm labor? *Am. J. Perinatol.* **6**, 196–199 (1989).
35. Mouillac, B., Sen, T., Durroux, T., Gaibelet, G. & Barberis, C. Chapter 13. Expression of human vasopressin and oxytocin receptors in *Escherichia coli*. In *Progress in Brain Research* Vol. 139 (eds Waxman, S. *et al.*) 163–177 (Elsevier, 2002).
36. Klein, U., Gimpl, G. & Fahrenholz, F. Alteration of the myometrial plasma membrane cholesterol content with beta-cyclodextrin modulates the binding affinity of the oxytocin receptor. *Biochemistry* **34**, 13784–13793 (1995).
37. Gimpl, G. & Fahrenholz, F. Cholesterol as stabilizer of the oxytocin receptor. *Biochim. Biophys. Acta Biomembr.* **1564**, 384–392 (2002).
38. Mouillac, B., Manning, M. & Durroux, T. Fluorescent agonists and antagonists for vasopressin/oxytocin G protein-coupled receptors: Usefulness in ligand screening assays and receptor studies. *Mini Rev. Med. Chem.* **8**, 996–1005 (2008).
39. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In *Methods in Neurosciences* Vol. 25 (ed. Sealfon, S. C.) 366–428 (Academic Press, 1995).
40. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genomics Proteomics Bioinform.* **13**, 278–289 (2015).
41. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
42. Waltenspühl, Y., Schöppe, J., Ehrenmann, J., Kummer, L. & Plückthun, A. Crystal structure of the human oxytocin receptor. *Sci. Adv.* **6**, 5419 (2020).
43. Alexandrov, A. I., Mileni, M., Chien, E. Y., Hanson, M. A. & Stevens, R. C. Microscale fluorescent thermal stability assay for membrane proteins. *Structure* **16**, 351–359 (2008).
44. Nasr, M. L. *et al.* Covalently circularized nanodiscs for studying membrane proteins and viral entry. *Nat. Methods* **14**, 49–52 (2017).
45. Goricanec, D. *et al.* Conformational dynamics of a G-protein α subunit is tightly regulated by nucleotide binding. *PNAS* **113**, E3629–E3638 (2016).
46. Schuster, M. *et al.* Optimizing the α 1B-adrenergic receptor for solution NMR studies. *Biochim. Biophys. Acta Biomembr.* **1862**, 183354 (2020).
47. Egloff, P. *et al.* Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in *Escherichia coli*. *PNAS* **111**, E655–E662 (2014).
48. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *BioRxiv* **2**, 563866 (2020).
49. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).
50. Schlinkmann, K. M. *et al.* Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. *PNAS* **109**, 9810–9815 (2012).
51. Barberis, C., Mouillac, B. & Durroux, T. Structural bases of vasopressin/oxytocin receptor function. *J. Endocrinol.* **156**, 223–229 (1998).
52. Claff, T. *et al.* Elucidating the active δ -opioid receptor crystal structure with peptide and small-molecule agonists. *Sci. Adv.* **5**, 9115 (2019).
53. Baker Brachmann, C. *et al.* Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**, 115–132 (1998).

54. Van Deventer, J. A. & Wittrup, K. D. Yeast surface display for antibody isolation: Library construction, library screening, and affinity maturation. In *Monoclonal Antibodies: Methods and Protocols* (eds Ossipow, V. & Fischer, N.) 151–181 (Humana Press, 2014).
55. Grisshammer, R., Duckworth, R. & Henderson, R. Expression of a rat neurotensin receptor in *Escherichia coli*. *Biochem. J.* **295**, 571–576 (1993).
56. Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24**, 79–88 (2006).

Acknowledgements

We thank the staff of the Cytometry Facility at the University of Zurich for support during FACS as well as Anna Bratus of the Functional Genomics Center at the University of Zurich for facilitating PacBio sequencing. We further thank Weihong Qi of the Functional Genomics Center at the University of Zurich, J. Ehrenmann and Y. Stark for helpful discussions. This work was supported by Schweizerischer Nationalfonds Grant 31003A_182334 to A.P.

Author contributions

Y.W. and L.K. created libraries and performed directed evolution in yeast. Y.W. created libraries and performed directed evolution in *E. coli*. Y.W. expressed, purified and measured thermostability of the proteins. J.J. set up the next generation sequencing data processing. J.J. and Y.W. analysed the next generation sequencing results. Project management was carried out by Y.W. and A.P. The manuscript was prepared by Y.W., J.J. and A.P. All authors contributed to the final editing and approval of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-87793-9>.

Correspondence and requests for materials should be addressed to A.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021