Title
A Massively Parallel Assay for Understanding Receptor-Ligand Relationships

Author
Jones, Eric

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Massively Parallel Assay for Understanding Receptor-Ligand Relationships

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy in Molecular Biology

by

Eric Jones

2018

ABSTRACT OF THE DISSERTATION

A Massively Parallel Assay for Understanding Receptor-Ligand Relationships

by

Eric Jones

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2018

Professor Sriram Kosuri, Chair

In this dissertation, I describe the development and application of a multiplexed method for high-throughput screening of receptor-ligand interactions. Such interactions underpin our cells' ability to sense and respond to their environment and represent a primary venue for therapeutic intervention. By leveraging advancements in DNA synthesis, genome editing, and next-generation sequencing, we have built a platform to measure the activity of a mixed population of receptors through RNA-seq of barcoded genetic reporters. We demonstrate the utility of the method for large-scale identification of chemical-receptor interactions and biochemical characterization of receptor function.

First, small molecules can interact with many biological targets in an organism, and uncovering these relationships is critical for modulating their function. Mammalian olfactory receptors (ORs), a large family of G protein-coupled receptors (GPCRs), mediate the sense of smell through activation by odorant small molecules. Each OR can respond to many odorants, and vice versa, making exploring this space one interaction at a time difficult. We used the platform to screen chemicals against a multiplexed library of ORs. We screened three concentrations of 181

odorants, where in each well we record the activity of 39 ORs simultaneously, and identified 79 novel associations, including ligands for 15 orphan receptors.

Second, GPCRs are ubiquitous throughout mammalian biology. They are conformationally dynamic which is essential to their function, but makes them recalcitrant to many techniques of structural determination. Here, we mutagenize and characterize all 7,828 possible missense variants of the beta-2-adrenergic receptor. On a broad scale, we find positions that respond similarly to mutation share certain properties of their environment and functional role within the protein. We recapitulate the importance of known critical residues and motifs and identify new residues important for function. Additionally, we describe an unreported, conserved extracellular motif maintained in both the inactive and active conformation of the protein that is essential for function.

As a whole, multiplexed screening enables the investigation of many outstanding questions in receptor biology. It is applicable to the disparate biological niches and systems that receptors occupy. As demonstrated in this dissertation, it has the potential to be a powerful tool for mapping receptor-ligand interactions and understanding receptor biochemistry.

The dissertation of Eric Jones is approved.

Elissa A Hallem

Hiroaki Matsunami

Todd O Yeates

Sriram Kosuri, Committee Chair

University of California, Los Angeles

2018

Dedicated to my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Professor Sriram Kosuri, for his guidance, mentorship, and help during my tenure as a graduate student. The work presented in this dissertation would not be possible without his continuous support. The training, resources, and motivation he has provided was invaluable and has made me a better scientist.

Furthermore, I would like to thank Professors Elissa Hallem, Hiroaki Matsunami, and Todd Yeates for serving as my committee members. Their continual guidance, comments, and suggestions are greatly appreciated.

Other current and past members of the Kosuri Lab have also played an instrumental role in my development as a scientist. I would like to thank Dr. Rocky Cheung, a postdoc, for his expertise in synthetic biology as well as his numerous helpful scientific contributions, suggestions, and discussions. Also, Dr. Rishi Jajoo, my close collaborator on the Olfaction work, contributed greatly to the success of the project. I've had the pleasure of working with Nathan Lubock for the entirety of my graduate career and we've become close friends in the process while working on the Deep Mutational Scanning work. My two former undergraduate assistants (and research associate): Daniel Cancilla and Megan Satyadi, it has been an honor watching you grow as scientists. I would also like to thank the current and former postdocs, students, and members of the Kosuri Lab lab, Dr. Hwangbeom Kim, Dr. Calin Plesa, Jessica Davis, Guillaume Urtecho, Kimberly Insigne, Clifford Boldridge, Angus C. Sidore, Christina Burghard, Suraj Alva, Michael Corrado, David Yao, Taylor Ward, Jeff Wang, Arielle Tripp, Marcia Brink, Joyce Samson, and Johnny Lee.

In addition, I've had the pleasure of collaborating with the Matsunami lab at Duke University; Professor Hiro Matsunami was always extremely knowledgeable and helpful regarding Olfaction and Olfactory Receptors.

This work would not be possible without the exceptional dedication and competence of the UCLA Broad Stem Cell Research Center, particularly Suhua Feng. Additionally the UCLA Technology Center for Genomics & Bioinformatics was very helpful as well.

I would like to acknowledge the Chemistry-Biology Interface Training Grant program at UCLA which is supported by the USPHS National Research Service Award 5T32GM008496.

I would like to thank Professor Martin Poenie at the University of Texas at Austin who gave me my first opportunity to do research. His outlook and integrity in science and life continue to inspire me on a daily basis. I would also like to thank Dr. Schonna Manning, a postdoc and staff scientist, who mentored me directly throughout my undergraduate research career.

I would like to thank my grandparents, aunts, uncles, cousins, and parents for their support. I especially wish to acknowledge my parents who have always supported me and pushed me to achieve my goals.

Finally, I would like to thank the friends and colleagues that I have met at UCLA. Everyone that I have met has taught me something and impacted my life positively. In particular, my good friends, N.L., C.K., C.R., G.H., W.T., A.C., A.S., A.C., C.H. D.C., and J.O.

VITA

Education

B.S. Biochemistry, University of Texas at Austin

May 2013

Research and Work Experience

Graduate researcher, Kosuri Lab, UCLA

Sept 2013 – Oct 2018

Consultant, Octant

Sept 2017 – Oct 2018

Consultant, Bioscentric

Apr 2017 – July 2017

Undergraduate Research Assistant, Poenie/Brand Lab, UT Austin

Sept. 2010 - May 2013

Honors and Awards

Outstanding Poster Presentation UCLA-MBI retreat

Apr 2018

USPHS National Research Service Award

Jul 2014 – Jun 2017

Graduate Dean's Scholar Award, UCLA

Apr 2013

Patents

Eric Jones and Sriram Kosuri. Mutliplexed Receptor-Ligand Interaction Screens, UCLA, *Patent Application Submitted*

PUBLICATIONS

A Scalable, Multiplexed Assay for Decoding Receptor-Ligand Interaction. **Jones EM**, Jajoo R, Cancilla D, Lubock N, Satyadi M, de March C, Cheung R, Bloom JS, Matsunami H, Kosuri S. *In Review*.

Deep Mutational Scanning of the Beta-2 Adrenergic Receptor. **Jones EM**, Lubock N, Cancilla D, Satyadi M, Davis J, Kosuri S. In Preparation.

Many rare genetic variants have unrecognized large-effect disruptions to exon recognition. Cheung R, Insigne KD, Yao D, Burghard CP, **Jones EM**, Goodman DB, Kosuri S. *Accepted Molecular Cell.*

Structure-based design of functional amyloid materials. Li D, **Jones EM**, Sawaya MR, Furukawa H, Luo F, Ivanova M, Sievers SA, Wang W, Yaghi OM, Liu C, Eisenberg DS. J Am Chem Soc. 2014 Dec 31;136(52):18044-51. doi: 10.1021/ja509648u. Epub 2014 Dec 19.

CHAPTER ONE

Introduction

Background

Cell surface receptors play a fundamental and unique role in biology. Broadly, they allow a cell to sense its external environment and modulate internal signaling pathways, evoking a response. Their essentiality is highlighted by their ubiquitous presence across taxonomy, and are found all the way from viruses to humans. Unsurprisingly, their highly specialized role as the mediators between the internal and external has thrust them into many diverse niches, a few examples of which I will describe. First, T-cell receptors (TCRs) discriminate between the self and non-self by binding peptides attached to the major histocompatibility complex. Upon recognition of peptides of foreign origin, TCRs activate the immune response[1]. Second, immediately after sustaining a wound, platelets will arrive to clot the blood vessel. Platelets release growth factors and cytokines that bind to receptors in skin cells and white blood cells to clear the debris of dead and damaged cells and promote the growth of new cells[2]. Third, the process of neurotransmission to and from the brain is mediated by receptors. Ligand-activated ion channels and metabotropic receptors bind various neurotransmitters released from the axon of a neighboring, synapsed neuron to elicit an excitatory or inhibitory response in the neuron[3]. Furthermore, the involvement of cell-surface receptors in diverse functional roles and accessibility at the membrane makes them well-suited for therapeutic intervention. In fact, 34% of Food and Drug Administration (FDA)-approved drugs target the ~400-member receptor superfamily, G protein-coupled receptors (GPCRs), alone[4]. Given the preeminence of cell surface receptors throughout human biology, a priority of biomedical researcher is to understand the relationship between a receptor and the ligand(s) it binds.

A New Era of Biology

Molecular biology is in the midst of a revolution. Experimental tools are being developed at an unprecedented rate enabling the undertaking of ambitious research projects not possible 10

years ago. The transformation is founded on advancements in three particular areas: (i) DNA synthesis, (ii) genome engineering, and (iii) next-generation sequencing (NGS).

Until very recently, the construction of synthetic genes and sequences was limited by three main, interwoven factors: cost, gene length, and number of constructs[5]. Common academic and commercial gene synthesis methods are comprised of oligonucleotides synthesized by phosphoramidite chemistry that are stitched together using various techniques. Traditionally, these techniques are ligation and polymerase chain assembly[6–9]. Recently, isothermal recombination- and assembly-based approaches have enabled the construction of synthetic sequences that are tens to hundreds of kilobases long[10,11]. A parallel technology, oligonucleotide microarray synthesis, has blossomed at the same time. Analogous to inkjet printing, a library of oligos is built immobilized on a microchip one nucleotide at a time. Microarray synthesis has unlocked the ability to build pools of tens to hundreds of thousands of short (<230 nt) sequences in a pooled format at a cost orders of magnitude cheaper than column-synthesized oligonucleotides[12–14] . High-throughput, scalable methods for assembling microarray-derived oligos into genes have enabled the construction of hundreds to thousands of genes of intermediate length (~1 kb) from a microarray pool[15,16]. Improvements in length and scale have broadened the problems researchers can tackle. Experiments requiring the construction of dozens to hundreds of multi-kb long genes are now economically feasible. Additionally, scientists can utilize microarray pool directly to probe the effect of thousands of short sequences synthesized *en masse*.

Genome engineering was once a long and laborious process requiring significant effort both for programming a nucleus to target a defined locus and achieving efficient disruption of a gene through non-homologous end joining or insertion with homology directed repair[17]. The advent of CRISPR/Cas9 and similar methods have alleviated both of these limitations for mammalian

genome engineering[18]. The CRISPR/Cas9 system requires a short programmable 20 nt 'guide' RNA sequence with very malleable constraints to target any specific region of the genome. Development of a chassis cell line for a specific experimental model often requires the addition or deletion of multiple transgenes and this can be achieved in multiplex with CRISPR/Cas9[19]. Additionally, construction of these models has been reduced to the scale of months or weeks to accomplish[20]. Alternative genome engineering technologies have been developed such as transposon- and recombinase-based systems. Well-suited for stable overexpression, transposon integration systems allow for the addition of a single or several genes at a variable copy number into semi-random regions of the genome[21,22]. For applications requiring controlled copy numbers, recombinase integration systems are more apt. Site-specific introduction of a recombinase recognition site enables the recombination of a construct containing a paired recognition site into the defined locus in the presence of the recombinase[23,24]. This technology is especially useful when constructs need to be limited to one copy per cell[25].

Perhaps the most fundamental development has been the invention of NGS and the following advancements that have drastically reduced the cost of sequencing per base[26,27]. In existence for more than a decade, initial applications centered around sequencing whole genomes[28,29]. Short-read NGS (sequencing by synthesis) enables reading thousands to millions of short DNA sequences at once. This enables the identification of genes but also the quantification of genes relative to one another -- NGS has become a measurement tool. Since its inception, numerous diverse applications have come online measuring: the transcriptome, chromatin accessibility, chromosome architecture, the translatome, and much more[30–33]. Separately, single molecule long-read sequencing enables access to repetitive and complex sequence regions and enhanced haplotyping of genomes[34–38].

Multiplexed Functional Assays

Rapid innovation in the aforementioned areas has led to the emergence of an entirely new style of research. Multiplexed Functional Assays, or MFAs, characterize a functional biological output of thousands of different sequences *in vitro* entirely at the same time in a single experiment[39–42]. These experimental platforms utilize DNA synthesis to generate large sequence libraries (up to tens of thousands), genome engineering to prepare host cell lines to compute a defined function, and NGS to interpret the phenotypic differences in that defined function within the DNA library in multiplex.

Such assays are quite versatile and have been applied towards a diverse set of problems. Initial iterations profiled the effects of DNA regulatory elements on gene expression in bacterial and human cell models[43–47]. More recently, many reports have comprehensively detailed the fitness effects of mutations within genes. High-throughput protein mutagenesis paired with a generalizable or gene-specific functional selection has classified the fitness associated with human variation and revealed hidden insight into the biochemical mechanism of many genes[48–51]. Moreover, as these tools have become more sophisticated, MFA's have investigated more nuanced topics such as RNA splicing[52]. The advent of MFA's enables the broader scientific community to tackle a plethora of biological questions that were both labor and cost prohibitive prior.

Focus of Thesis Projects

We hypothesized MFAs could be developed that would be informative to answer two major questions surrounding receptor-ligand relationships:

1) What natural and synthetic ligands bind which receptors for a given class?

Exhaustively mapping ligands to their cognate receptors often requires exploring a large potential interaction space. Therefore, technology that enables screening ligands against cohorts of receptors in multiplex drastically reduces the complexity of the search space.

2) What is the contribution of each individual residue for signal transduction of a receptor-ligand binding event? Mutagenesis coupled with an assay for protein activity has long been a pillar of the biochemical toolbox for understanding a protein's structure-function relationship. By comprehensively profiling thousands of single mutations to a receptor, we can probe its conformational stability, dynamics, signaling, and ligand binding.

We chose to pilot this platform on a class of receptors known as G protein-coupled receptors (GPCRs) because of their prominence in human disease, therapeutic targeting, and applicability to the aforementioned questions. GPCRs are a superfamily comprised of ~800 members involved in many biological processes: hormone sensing, neurotransmission, smell, taste, vision, immunity, sleep, and many more[53,54]. They are targeted by 34% of FDA-approved drugs and implicated in diseases such as asthma, obesity, anxiety, depression, diabetes, Alzheimer's disease, Parkinson's Disease, cancer, and AIDS[4,54]. The remainder of this chapter will introduce and motivate the two projects that comprise my thesis.

Mapping Receptor-Ligand Interactions in Mammalian Olfactory Receptors

Ligands often interact with cohorts of receptors instead of a single receptor. Such polypharmacology enables the biological system to encode redundancy and achieve ligand and cell-type specificity of a magnitude unable to be reinforced by a single receptor[55,56]. For example, the bone morphogenetic pathway (BMP) comprises a promiscuous receptor group that binds multiple ligands. Individual cell types express different receptor repertoires, performing distinct computations that vary when stimulated with different compositions of their cognate ligands[57]. Such polypharmacology is a powerful biological mechanism for encoding a large set of responses to variable inputs. Pointedly, this phenomenon is exceedingly prominent in mammalian olfaction[58–60].

First, mammalian olfaction -- the sense of smell -- is a complex neurobiological process that translates chemical inputs (odorants) into odor perception. Initially, odorants bind olfactory receptors (ORs) expressed monoallelically by olfactory sensory neurons (OSNs) in the nasal epithelium[59,61–63]. OSN axons converge in the forebrain's olfactory bulb to form glomeruli, spherical axon clusters of the same receptor type hypothesized to be arranged spatially according to the chemical structure of the odor they respond to[60,64–66]. Mitral cells synapse with these axons and transmit information to various regions of the cerebral cortex associated with determining chemical structures of odorants, categorizing odors and assessing similarities between them, conscious odor perception, and automatic and emotional responses to odor[67–73].

These receptors bind odorant molecules in a combinatorial fashion, each receptor can bind many chemicals and each chemical can bind many receptors. Our mapping of odorant-receptor interactions is very limited. About 86% of human and mouse olfactory receptors remain orphan[74]. This is largely because of two problems: olfactory receptors (OR) are very difficult to express in heterologous systems and the combinatorial complexity of the screening space is too difficult to screen each receptor-ligand pair individually[75,76]. Our sparse mapping of odorants to their respective ORs is inherently limiting to our understanding of the chemical basis of odor recognition and perception. Furthermore, this is a bottleneck at the primary layer of olfaction that inhibits our ability to probe the downstream neurobiological processes governing odor perception. Therefore, advanced screening technology for mapping receptor-ligand interactions would be greatly enabling for researchers. In Chapter 2, we present a method to screen ligands in high-throughput against a cohort of olfactory receptors in multiplex

Deep Mutational Scanning of the Beta-2 Adrenergic Receptor

GPCRs are structurally dynamic and, upon ligand stimulation, undergo a conformational shift to prompt signal transduction[77–79]. Their dynamic nature makes them recalcitrant to standard methods for structure determination such as x-ray crystallography and cryo-electron microscopy (cryo-EM)[80,81]. However, over the past decade monumental efforts have led to structures for more than 50 receptors[82]. Although, each method has its own restrictions. X-ray crystallography provides only a single snapshot of a conformational state and often requires truncation, non-native addition, or artificial stabilization of certain protein segments[83]. Inactive state structures make up the vast majority of crystal structures, with only ~18 active state structures existing for the >200 total GPCR structures. Additionally, monomeric GPCRs are at the lower bound of the molecular weight requirement for cryo-EM and only a single complete structure has been solved with spectroscopy[84,85].

An alternative to structural biology, mutagenesis, coupled with a screen for function, has long been a foundation of biochemistry for understanding the structure-function relationship of a protein and GPCRs specifically[86]. Historically, technical constraints restricted the number of mutations that could be generated and characterized. Recent, aforementioned advancements in DNA synthesis, genome editing, and NGS have led to the creation of Deep Mutational Scanning (DMS), an experimental technique to functionally assay all possible missense variants of a given protein[87]. DMS has been powerful for interpreting genetic variation in clinically relevant human proteins[48–50]. It can also serve as a powerful aid for investigating the functional significance of individual regions and residues of a protein, especially when augmented with a crystal structure. Such an approach has been applied to GPCRs, however the phenotypic screens applied were limited to cell-surface expression and ligand binding rather than a direct output of GPCR function[88,89].

We chose to perform DMS on the beta-2 adrenergic receptor (β2AR), one of the most exhaustively studied proteins in history[90–96]. It has been comprehensively structurally and functionally characterized and presents a great opportunity to pilot and validate functional DMS of GPCRs. In Chapter 3, we present a technology capable of profiling the relative activation of G protein signaling for every missense variant of the beta-2 adrenergic receptor in multiplex to multiple agonist conditions.

Conclusions

Taken together, the following two chapters entail a new approach for studying the relationship between receptors and the ligands they bind. From screening small-to-intermediate cohorts of receptors against large chemical panels to massive libraries of receptor variants against a few agonist conditions, we display the versatility and robustness of this platform. I hope to impart to you by the end of this thesis the transformative ability this technology will have on receptor-ligand biology.

References

1. Janeway, C. A., Jr, Travers, P., Walport, M. & Shlomchik, M. J. *The major histocompatibility complex and its functions*. (Garland Science, 2001).

2. Barrientos, S., Stojadinovic, O., Golinko, M. S., Brem, H. & Tomic-Canic, M. Growth factors and cytokines in wound healing. *Wound Repair Regen.* **16,** 585–601 (2008).

3. Lodish, H. *et al. Neurotransmitters, Synapses, and Impulse Transmission*. (W. H. Freeman, 2000).

4. 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).

5. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11,** 499–507 (2014).

6. Elsevier. Synthetic Biology - 1st Edition. Available at: https://www.elsevier.com/books/synthetic-biology/zhao/978-0-12-394430-6. (Accessed: 10th September 2018)

7. Carr, P. A. & Church, G. M. Genome engineering. *Nat. Biotechnol.* **27,** 1151–1162 (2009).

8. Czar, M. J., Anderson, J. C., Bader, J. S. & Peccoud, J. Gene synthesis demystified. *Trends Biotechnol.* **27,** 63–72 (2009).

9. Xiong, A.-S. *et al.* Chemical gene synthesis: strategies, softwares, error corrections, and applications. *FEMS Microbiol. Rev.* **32,** 522–540 (2008).

10. Annaluru, N. *et al.* Total synthesis of a functional designer eukaryotic chromosome. *Science* **344,** 55–58 (2014).

11. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).

12. Blanchard, A. P., Kaiser, R. J. & Hood, L. E. High-density oligonucleotide arrays. *Biosensors and Bioelectronics* **11,** 687–690 (1996).

13. Hughes, T. R. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19,** 342–347 (2001).

14. Saaem, I., Ma, K.-S., Marchi, A. N., LaBean, T. H. & Tian, J. In situ synthesis of DNA microarray on functionalized cyclic olefin copolymer substrate. *ACS Appl. Mater. Interfaces* **2,** 491–497 (2010).

15. Plesa, C., Sidore, A. M., Lubock, N. B., Zhang, D. & Kosuri, S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* **359,** 343–347 (2018).

16. Kosuri, S. *et al.* Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.* **28,** 1295–1299 (2010).

17. Carroll, D. Genome engineering with targetable nucleases. *Annu. Rev. Biochem.* **83,** 409–439 (2014).

18. Xiong, X., Chen, M., Lim, W. A., Zhao, D. & Qi, L. S. CRISPR/Cas9 for Human Genome Engineering and Disease Research. *Annu. Rev. Genomics Hum. Genet.* **17,** 131–154 (2016).

19. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).

20. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8,** 2281–2308 (2013).

21. Li, X. *et al.* piggyBac transposase tools for genome engineering. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E2279–87 (2013).

22. Vink, C. A. *et al.* Sleeping beauty transposition from nonintegrating lentivirus. *Mol. Ther.* **17,** 1197–1204 (2009).

23. Duportet, X. *et al.* A platform for rapid prototyping of synthetic gene networks in

mammalian cells. *Nucleic Acids Res.* **42,** 13440–13451 (2014).

24. Zhu, F. *et al.* DICE, an efficient system for iterative genomic editing in human pluripotent stem cells. *Nucleic Acids Res.* **42,** e34 (2014).

25. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* **45,** e102 (2017).

26. Wetterstrand, K. A. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). (2013).

27. Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **6,** 287–303 (2013).

28. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17,** 333–351 (2016).

29. Kircher, M. & Kelso, J. High-throughput DNA sequencing--concepts and limitations. *Bioessays* **32,** 524–536 (2010).

30. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

31. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10,** 669–680 (2009).

32. Brar, G. A. & Weissman, J. S. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* **16,** 651–664 (2015).

33. Eagen, K. P. Principles of Chromosome Architecture Revealed by Hi-C. *Trends Biochem. Sci.* **43,** 469–478 (2018).

34. Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* **16,** 344–358 (2015).

35. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using

single-molecule sequencing. *Nature* **517,** 608–611 (2015).

36. Schatz, M. C., Delcher, A. L. & Salzberg, S. L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20,** 1165–1173 (2010).

37. English, A. C. *et al.* Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genomics* **16,** 286 (2015).

38. Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res.* **25,** 1750–1756 (2015).

39. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11,** 1782–1787 (2016).

40. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101,** 315–325 (2017).

41. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106,** 159–164 (2015).

42. Santiago-Algarra, D., Dao, L. T. M., Pradel, L., España, A. & Spicuglia, S. Recent advances in high-throughput approaches to dissect enhancer function. *F1000Res.* **6,** 939 (2017).

43. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 14024–14029 (2013).

44. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27,** 1173–1175 (2009).

45. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* **45,** 1021–1028 (2013).

46. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23,** 800–811 (2013).

47. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30,** 271–277 (2012).

48. Majithia, A. R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* (2016). doi:10.1038/ng.3700

49. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50,** 874–882 (2018).

50. Kotler, E. *et al.* A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Mol. Cell* **71,** 178–190.e8 (2018).

51. Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200,** 413–422 (2015).

52. Cheung, R. *et al.* Many rare genetic variants have unrecognized large-effect disruptions to exon recognition. *bioRxiv* 199927 (2018). doi:10.1101/199927

53. Rosenbaum, D. M., Rasmussen, S. G. F. & Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **459,** 356–363 (2009).

54. Katritch, V., Cherezov, V. & Stevens, R. C. Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmacol. Sci.* **33,** 17–27 (2012).

55. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3,** 353–359 (2004).

56. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14,** 111–129 (2015).

57. Antebi, Y. E. *et al.* Combinatorial Signal Perception in the BMP Pathway. *Cell* **170,** 1184–1196.e24 (2017).

58. Saito, H., Chi, Q., Zhuang, H., Matsunami, H. & Mainland, J. D. Odor coding by a Mammalian receptor repertoire. *Sci. Signal.* **2,** ra9 (2009).

59. Malnic, B., Hirono, J., Sato, T. & Buck, L. B. Combinatorial receptor codes for odors. *Cell* **96,** 713–723 (1999).

60. Ressler, K. J., Sullivan, S. L. & Buck, L. B. A zonal organization of odorant receptor gene expression in the olfactory epithelium. *Cell* **73,** 597–609 (1993).

61. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65,** 175–187 (1991).

62. Buck, L. B. Smell and taste: the chemical senses. *Principles of neural science* **4,** 625–647 (2000).

63. Niimura, Y. & Nei, M. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 6039–6044 (2005).

64. Ressler, K. J., Sullivan, S. L. & Buck, L. B. Information coding in the olfactory system: evidence for a stereotyped and highly organized epitope map in the olfactory bulb. *Cell* **79,** 1245–1255 (1994).

65. Mombaerts, P. *et al.* Visualizing an olfactory sensory map. *Cell* **87,** 675–686 (1996).

66. Uchida, N., Takahashi, Y. K., Tanifuji, M. & Mori, K. Odor maps in the mammalian olfactory bulb: domain organization and odorant structural features. *Nat. Neurosci.* **3,** 1035–1043 (2000).

67. Fantana, A. L., Soucy, E. R. & Meister, M. Rat olfactory bulb mitral cells receive sparse glomerular inputs. *Neuron* **59,** 802–814 (2008).

68. Kadohisa, M. & Wilson, D. A. Separate encoding of identity and similarity of complex familiar odors in piriform cortex. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 15206–15211 (2006).

69. Barnes, D. C., Hofacer, R. D., Zaman, A. R., Rennaker, R. L. & Wilson, D. A.

Olfactory perceptual stability and discrimination. *Nat. Neurosci.* **11,** 1378–1380 (2008).

70. Haberly, L. B. Parallel-distributed processing in olfactory cortex: new insights from morphological and physiological analysis of neuronal circuitry. *Chem. Senses* **26,** 551–576 (2001).

71. de Araujo, I. E., Rolls, E. T., Velazco, M. I., Margot, C. & Cayeux, I. Cognitive modulation of olfactory processing. *Neuron* **46,** 671–679 (2005).

72. Kobayakawa, K. *et al.* Innate versus learned odour processing in the mouse olfactory bulb. *Nature* **450,** 503–508 (2007).

73. Li, W., Howard, J. D., Parrish, T. B. & Gottfried, J. A. Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. *Science* **319,** 1842–1845 (2008).

74. *Springer Handbook of Odor*. (Springer International Publishing, 2017).

75. Peterlin, Z., Firestein, S. & Rogers, M. E. The state of the art of odorant receptor deorphanization: a report from the orphanage. *J. Gen. Physiol.* **143,** 527–542 (2014).

76. Lu, M., Echeverri, F. & Moyer, B. D. Endoplasmic reticulum retention, degradation, and aggregation of olfactory G-protein coupled receptors. *Traffic* **4,** 416–433 (2003).

77. Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. GPCR Dynamics: Structures in Motion. *Chem. Rev.* **117,** 139–155 (2017).

78. Hilger, D., Masureel, M. & Kobilka, B. K. Structure and dynamics of GPCR signaling complexes. *Nat. Struct. Mol. Biol.* **25,** 4–12 (2018).

79. Weis, W. I. & Kobilka, B. K. The Molecular Basis of G Protein-Coupled Receptor Activation. *Annu. Rev. Biochem.* **87,** 897–919 (2018).

80. Bill, R. M. *et al.* Overcoming barriers to membrane protein structure determination. *Nat. Biotechnol.* **29,** 335–340 (2011).

81. Xiang, J. *et al.* Successful Strategies to Determine High-Resolution Structures of

GPCRs. *Trends Pharmacol. Sci.* **37,** 1055–1069 (2016).

82. Tesmer, J. J. G. Hitchhiking on the heptahelical highway: structure and function of 7TM receptor complexes. *Nat. Rev. Mol. Cell Biol.* **17,** 439–450 (2016).

83. Isberg, V. *et al.* GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **44,** D356–64 (2016).

84. Safdari, H. A., Pandey, S., Shukla, A. K. & Dutta, S. Illuminating GPCR Signaling by Cryo-EM. *Trends Cell Biol.* **28,** 591–594 (2018).

85. Park, S. H. *et al.* Structure of the chemokine receptor CXCR1 in phospholipid bilayers. *Nature* **491,** 779–783 (2012).

86. Kristiansen, K. Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacol. Ther.* **103,** 21–80 (2004).

87. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11,** 801–807 (2014).

88. Schlinkmann, K. M. *et al.* Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 9810–9815 (2012).

89. Heredia, J. D. *et al.* Mapping Interaction Sites on Human Chemokine Receptors by Deep Mutational Scanning. *J. Immunol.* **200,** 3825–3839 (2018).

90. Rasmussen, S. G. F. *et al.* Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **450,** 383–387 (2007).

91. Johnson, M. Molecular mechanisms of beta(2)-adrenergic receptor function, response, and regulation. *J. Allergy Clin. Immunol.* **117,** 18–24; quiz 25 (2006).

92. Bang, I. & Choi, H.-J. Structural features of β2 adrenergic receptor: crystal structures and beyond. *Mol. Cells* **38,** 105–111 (2015).

93. Kurose, H. β2-Adrenergic receptors: Structure, regulation and signaling by partial and full agonists. *Allergol. Int.* **53,** 321–330 (2004).

94. Dror, R. O. *et al.* Activation mechanism of the β2-adrenergic receptor. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 18684–18689 (2011).

95. Kolb, P. *et al.* Structure-based discovery of beta2-adrenergic receptor ligands. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 6843–6848 (2009).

96. Ostrowski, J., Kjelsberg, M. A., Caron, M. G. & Lefkowitz, R. J. MUTAGENESIS OF THE f32-ADRENERGIC RECEPTOR: HOW STRUCTURE ELUCIDATES FUNCTION. *Annu. Rev. Pharmacal. Taxical.* **32,** 167–163 (1992).

CHAPTER TWO

A Scalable, Multiplexed Assay for Decoding Receptor-Ligand Interactions

Title: A Scalable, Multiplexed Assay for Decoding Receptor-Ligand Interactions

Authors: Eric M. Jones[1]†, Rishi Jajoo[1]†, Daniel Cancilla[1], Nathan B. Lubock[1], Jeff Wang[1], Megan Satyadi[1], Rocky Cheung[1], Claire de March[2], Joshua S. Bloom[3], Hiroaki Matsunami[2], Sriram Kosuri[1]*

Affiliations:
[1]Department of Chemistry and Biochemistry, UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, Quantitative and Computational Biology Institute, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, and Jonsson Comprehensive Cancer Center, UCLA, Los Angeles, CA 90095, USA
[2]Department of Molecular Genetics and Microbiology, and Department of Neurobiology and Duke Institute for Brain Sciences, Duke University Medical Center, Research Drive, Durham, NC 27710, USA
[3]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA

Abstract:

Small molecules may interact with many biological targets, and uncovering these relationships is critical for modulating their function. We developed a chemical screening platform for multiplexed receptor libraries using next-generation sequencing of barcoded genetic reporters and screened three concentrations of 181 odorants against 39 olfactory receptors simultaneously, identifying 79 novel associations including ligands for 15 orphan receptors. This platform allows the cost-effective mapping of large chemical libraries to receptor repertoires at scale.

Introduction

Interactions between small molecules and receptors underpin an organism's ability to sense and respond to its internal state and the environment. For many drugs and natural products, the ability to modulate many biological targets at once is crucial for their efficacy[1–3]. Thus, to understand the effect of many small molecules, we need to comprehensively characterize their functional interactions with biological targets. This many-on-many problem is laborious to study one interaction at a time, and is especially salient in the mammalian sense of smell[4,5].

Olfaction is mediated by a class of G protein-coupled receptors (GPCRs) known as olfactory receptors (ORs)[6]. GPCRs are a central player in small molecule signaling and are currently targeted by 34% of US Food and Drug Administration (FDA) approved drugs[7]. ORs are a large family of class A GPCRs with approximately 396, 1130, and 1948 intact receptors in humans, mice, and elephants respectively[8]. Each OR can potentially interact with many odorants, and inversely, each odorant with many ORs. The majority of ORs remain orphan-- i.e., have no known ligand-- because of this vast combinatorial space, further compounded by the fact that recapitulating mammalian GPCR function *in vitro* is challenging[9,10]. In addition, no experimentally determined structure for any mammalian OR is available, hindering computational efforts to predict which odorants can activate each OR[11].

Most GPCR and OR assays test chemicals against each receptor individually[12,13]. Multiplexed assays, where the activities of multiple receptors-- often referred to as a library of receptors-- are measured in the same well, would increase the throughput but have remained technically challenging. In such an assay, each cell expresses a single type of receptor and, upon activation, transcribes a short barcode sequence that identifies the particular receptor expressed in that cell. The enrichment of barcoded transcripts corresponding to each receptor's activation are then measured by microarrays or next-generation sequencing. Such multiplexed GPCR activity assays have previously been attempted by transient transfection of individual

receptors and subsequent pooled screening[14,15]. However, these assays are difficult to perform, especially in olfaction, for several reasons. First, ORs, like many GPCRs, are difficult to express in their non-native contexts and often require specialized accessory factors and signaling proteins to function heterologously[16]. Second, transient transfection must be performed for tens to hundreds of individual cell lines each time an assay is performed. Thus, experimental protocols for such multiplexed screens are expensive, labor intensive, and often carried out in a low-throughput manner. Using stable lines would alleviate these burdens, but building stable OR reporter lines is challenging and has only worked in two reported cases[17,18].

Results

Here we report a new high-throughput screen to characterize small molecule libraries against mammalian OR libraries in multiplex. To do this, we developed both a stable cell line capable of functional OR expression (ScL21) and a multiplexed reporter for OR activity (Fig. 2.1a, 2.3). Activation of each OR leads to the expression of a reporter transcript with a unique 15-nucleotide barcode sequence. Each barcode identifies the OR expressed in that cell; this enables OR activation to be measured by quantifying differential barcode expression with RNA-seq. This technology enables the simultaneous profiling of a single chemical's activity against a library of receptors in a single well.

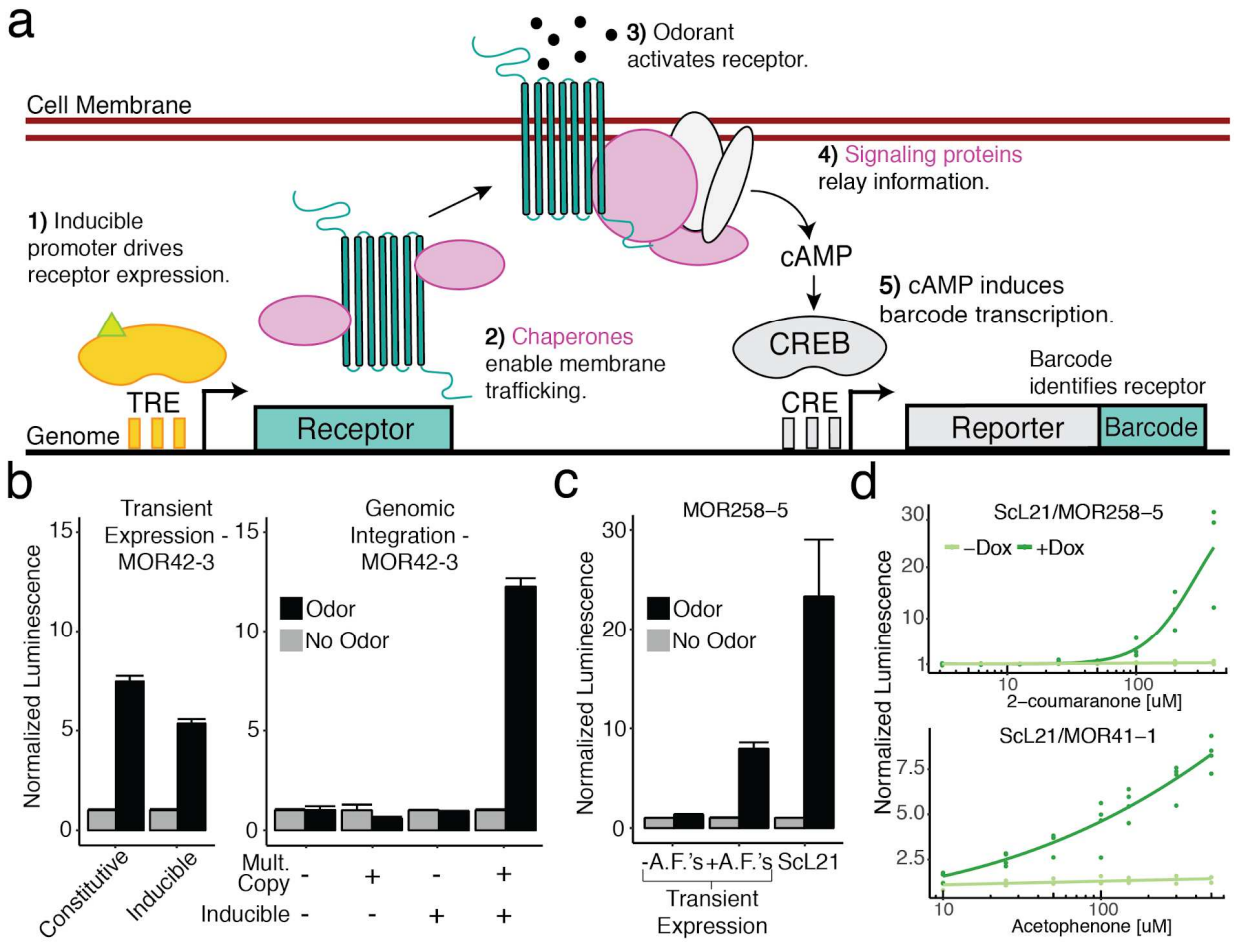


Fig. 2.1. A Genomically Integrated Synthetic Circuit Allows Screening for Mammalian Olfactory Receptor Activation. (**a)** Schematic of the synthetic circuit for stable OR expression and function in an engineered HEK293T cell line (ScL21). Heterologous accessory factors expressed include (pink): RTP1S, RTP2, $G_{\alpha olf}$, and Ric8b. (**b**) MOR42-3 reporter activation expressing the receptor transiently (left) or genomically integrated (right) at varying copy number, under constitutive or inducible expression in HEK293T cells. (**c**) MOR258-5 reporter activation with/without accessory factors (A.F.s), RTP1S and RTP2, transiently coexpressed in HEK293T cells compared to stable receptor expression in ScL21. (**d**) Reporter activation response curves for MOR258-5 and MOR41-1 genomically integrated in ScL21.

We first engineered a stable cell line, ScL21, capable of functionally expressing ORs and responding to odorant stimuli by transcribing an RNA barcode. First, we found that multi-copy integration and inducible receptor expression are both essential for reporter activation, but individually neither of these features is sufficient to generate a response (Fig. 2.1b, Fig. 2.4). Then, to allow larger OR repertoires to be assayed, we added features known to improve OR

function[16,19,20]. We stably integrated a pool of 4 accessory factors at multi-copy under inducible expression: $G_{\alpha olf}$ and Ric8b for signal transduction, and RTP1S and RTP2 to promote surface expression (Fig. 2.1c, Fig. 2.3 and 2.4). To select a single line for further use, we isolated clones and screened for robust activation of two ORs known to require accessory factors to function heterologously (Fig. 2.4). In addition, we then incorporated protein trafficking tags previously shown to increase surface expression[21,22], included DNA insulator sequences to reduce background reporter activation, optimized the cAMP response element (CRE) to improve reporter signal, and combined these improvements into a single transposable vector to speed cell line development (Fig. 2.5). We validated our system on two murine ORs with known ligands, and observed induction- and dose-dependent activation (Fig. 2.1d).

To pilot the platform, we chose 42 phylogenetically divergent murine ORs with both known and unknown chemical specificities and created a library of OR-expressing cell lines (Fig. 2.6). To create the individual cell lines, we first cloned and mapped the ORs to their corresponding barcodes and transposed the plasmids individually into the genomes of HEK293T cells[23]. After selection we pooled the cell lines together, generating assay-ready libraries for repeated testing (Fig. 2.2a). Unlike a luciferase reporter assay, each well contains the entire OR library and a single chemical's activity is measured against the entire library of ORs in a single well. We plated the cell library in 6-well culture dishes and screened odorants known to activate ORs in our library (Fig. 2.7); all but 3 ORs passed quality filtering to obtain reliable estimates of activation (See Methods). Analysis of the sequencing readout recapitulated previously identified odorant-receptor pairs[12], and chemical mixtures appropriately activated multiple ORs (Fig. 2.7). We found the assay was robust to chemicals such as the adenylate cyclase stimulator, forskolin, which non-specifically induces barcode transcription independent of the OR each cell expresses. This is likely because our library-based approach measures the relative activation of ORs to each other, normalizing any global effects due to off-target reporter activation.

Next, we adapted the platform for high-throughput screening in 96-well format. To decrease reagent cost and assay time, we optimized an in-lysate reverse transcription protocol and used dual indexing to uniquely link barcode reads to the correct well once samples were mixed for sequencing (see Methods). With these improvements, the assay is able to recapitulate dose-response curves for known odorant-receptor pairs (Fig. 2.7). We observed reproducible results between identically treated but biologically independent wells (Fig 2.8 and 2.9).

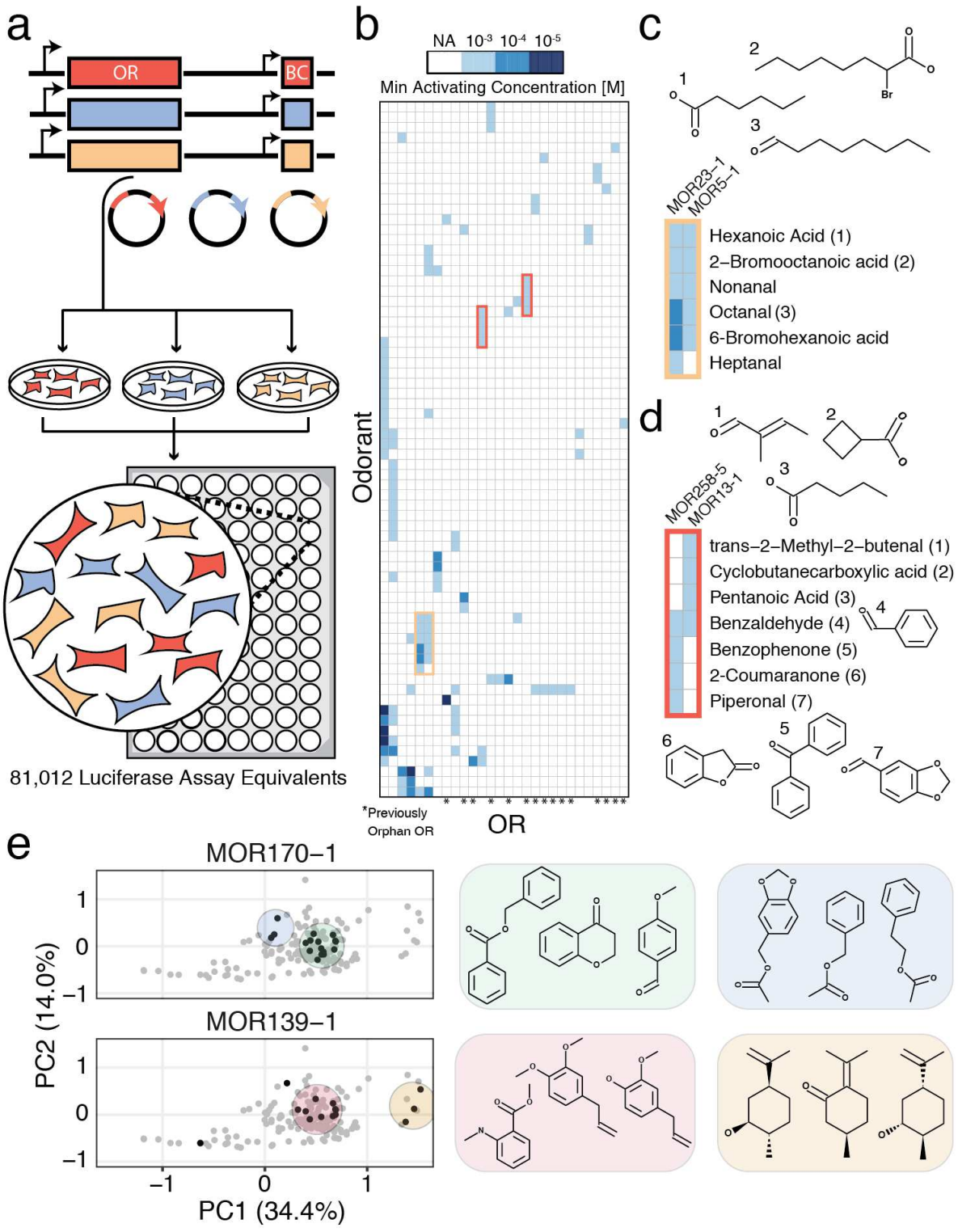


Fig. 2.2 Large-Scale, Multiplexed Screening of Olfactory Receptor-Odorant Interactions. (**a**) Experimental workflow for OR library generation and high-throughput screening. To perform assay, we cloned OR genes and barcodes into plasmids, engineered cell lines via individual transposition of plasmids, pooled cell lines and performed screen in 96 well plates. We assayed

the equivalent of 81,012 wells of a screen where interactions are tested individually. (**b)** Heatmap of interactions from the screen clustered by odorant and receptor responses, and shaded by the minimum activating odorant concentration that triggered reporter activity. Only ORs and chemicals that registered at least one interaction are shown. (**c**) Chemical names and structures for odorants that activate MOR23-1 and MOR5-1. (**d**) Chemical names and structures for odorants that activate MOR258-5 and MOR13-1. (**e**) Chemical hits identified for MOR170-1 and MOR139-1 (black) mapped onto a PCA projection of the chemical space of our odorant panel (grey). Shaded areas highlight hits that cluster together in chemical space.

We subsequently screened 181 odorants with both known and unknown receptor specificity at three concentrations in triplicate against the 39-member OR cell library, or 81,012 wells if each combination had been tested individually including controls (Fig. 2.2a and Tables 2.1 and 2.2). Each 96-well plate in the assay contained independent positive control odorants and solvent (DMSO) for normalization (Fig. 2.9). We used a generalized linear model to determine OR-odorant interactions (see Methods)[24]. We found 112 significant interactions (out of >7,000 combinations), of which 79 are novel, and 24 that target 15 orphan receptors (Benjamini-Hochberg corrected FDR = 1%; Fig. 2.2b, Fig. 10, and Table 2.3)[25]. Overall, 28 of 39 receptors were activated by at least one odorant, and 67 of 181 odorants activated at least one OR (Table 2.3).

We compared results to a previous study and analyzed individual interactions in a different context. First, we chose 36 interactions with at least 1.2-fold induction to retest individually in a previously developed transient OR activation system[26] (Fig. 2.11). Of the 27 significant interactions at an FDR of 1%, 20 of them replicated in this orthogonal system (Fig. 2.10). Notably, some of the seven interactions which did not replicate in this orthogonal system, look like true hits. For instance, our assay registered two hits for MOR19-1 with high chemical similarity (methyl salicylate and benzyl salicylate), suggesting they are likely not false positives (Fig. 2.11). Additionally, three of nine interactions not passing the 1% FDR threshold showed activation in the orthogonal assay, indicating that a conservative FDR threshold likely generated

some false negatives. A previous large-scale OR deorphanization study used some of the same receptors and chemicals, and we found that 9/12 of their reported interactions with EC_{50} below 100μM were also detected in our platform, though we did not identify most of the previous low affinity interactions[12] (Fig. 2.12). Conversely, we also detected 14 positive interactions absent from the previous study. Finally, our assay replicated the vast majority of non-interacting odorant-OR pairs (493/507).

Using the data generated by this high throughput assay, we found that chemicals with similar features activate the same ORs, including those receptors we deorphanize in this study (Fig. 2.2c). For example, the previously orphan MOR19-1 has clear affinity for the salicylate functional group, while MOR13-1 is activated by four chemicals with hydrogen bond accepting groups attached, and in three cases, to stiff non-rotatable scaffolds. We also detect ORs with partial overlap in chemical specificity; MOR13-1 detects compounds with terminal carbonyls while MOR258-5 detects cyclic conjugated molecules (Fig. 2.2d). Benzaldehyde, an intermediate size carbonyl, activates both ORs.

To more systematically understand how chemical similarity relates to receptor activation, we used a recently developed molecular autoencoder[27] to computationally map each tested chemical onto a ~292-dimensional continuous representation of chemical space and visualized the results with Principal Components Analysis (Fig. 2.13). Chemicals for 11/17 multi-hit receptors cluster together across the first two principal components, which explains 48.44% of the variance (Fig. 2.14). For instance, of 13 aliphatic aldehydes or carboxylic acids with >5 carbons in our chemical panel, 10 activate MOR5-1 (Fig. 2.2c). Interestingly, this analysis also highlights the instances where ORs are sensitive to several distinct sets of chemicals (Fig. 2.2e). For example, MOR139-1 is activated by compounds that belong to two distinct clusters: one with benzene rings and the other with cyclohexane rings, hinting at the selective features of

these odorants. Similarly, MOR170-1 exhibits a broad activation pattern: this receptor responds to ~50% of all odorants in our panel that contains both a benzene ring, and either a carbonyl or ether group. Most of these odorants form a single cluster with the exception of the acetate compounds that form a separate cluster. Understanding the global chemical space that activates each OR establishes the groundwork for the prediction of novel odorant-OR interactions.

Conclusions

We anticipate that this platform can be scaled to test the 396-member human OR repertoire and comprehensively define OR response to any odorant of interest. The approximate cost per well is on par with existing assays, but per receptor-ligand interaction interrogated, multiplexing dramatically reduces cost and labor. Our incomplete understanding for how ligands[28], drugs[1], hormones, natural products [29] and odors[12] interact with potential cellular targets limits our ability to rationally develop new molecules to modulate receptor activity. Multiplex methods like this platform offer a scalable solution to generate large-scale datasets that will help guide both empirical and algorithmic efforts to better dissect the complex interactions between small molecules and biological targets[30].

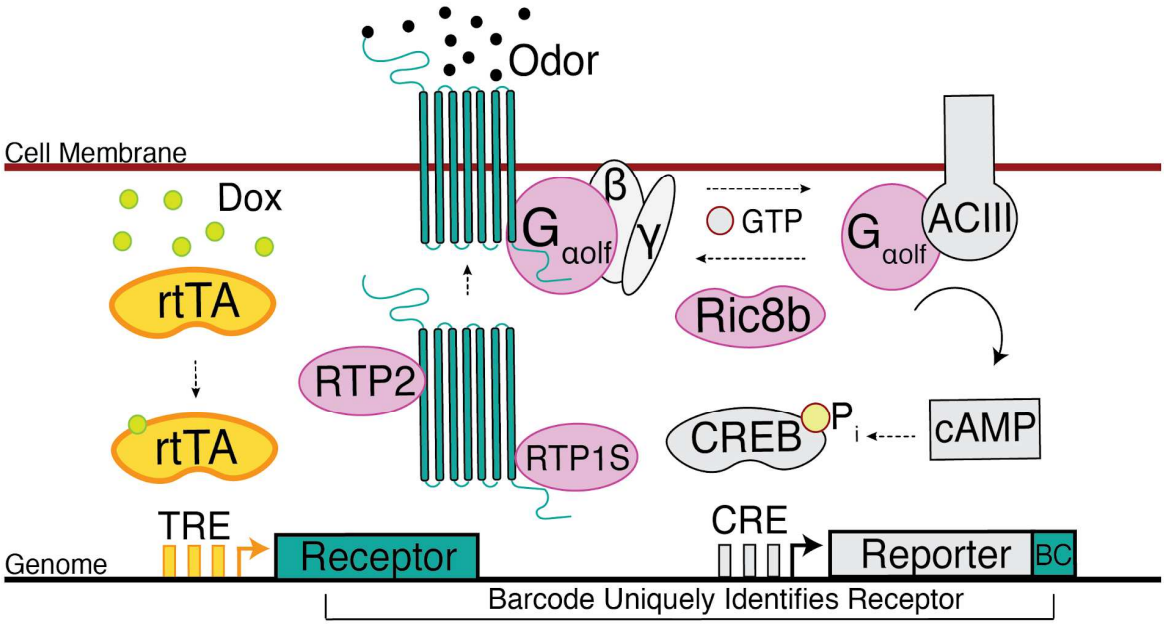


Figure 2.3. Schematic of the Synthetic Olfactory Activation Circuit in the Engineered Cell Line. Full graphical representation of the expressed components for expression/signaling of the ORs and the barcoded reporter system as shown in Fig. 1 of the main text. Receptor expression is controlled by the Tet-On system (Orange). After doxycycline induction, the OR is expressed on the cell surface with assistance from two exogenously expressed chaperones, RTP1S and RTP2 (pink). Upon odorant activation, G protein signaling triggers cAMP production. Signaling is augmented by transgenic expression of the native OR G alpha subunit, $G_{\square olf}$, and its corresponding GEF, Ric8b (pink). cAMP leads to activation of the kinase PKA that phosphorylates the transcription factor CREB leading to expression of the barcoded reporter.

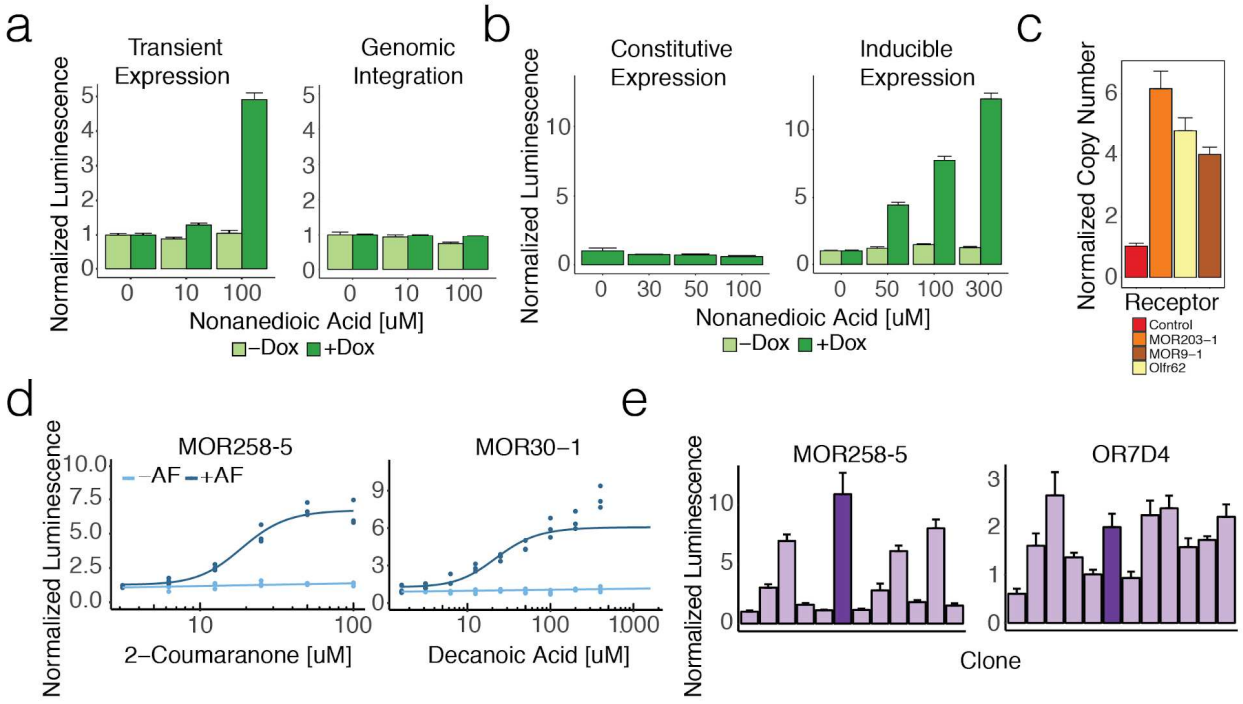


Fig.2.4. Engineering HEK293T Cells for Stable, Functional OR Expression. (**a**) Comparison of MOR42-3 activation under inducible receptor expression either transiently transfected (left) or integrated at single copy into the H11 genomic locus (right). (**b**) Comparison of MOR42-3 reporter activation integrated at multiple copies in the genome with the PiggyBac Transposon System under constitutive or inducible receptor expression. (**c**) Relative receptor/reporter copy number determined with qPCR for three transposed ORs relative to a single copy integrant. (**d**) Comparison of MOR258-5 and MOR30-1 reporter activation (stimulated with 2-coumaranone and Decanoic Acid respectively) co-transfected with or without Accessory Factors (AF) $G_{\alpha olf}$, Ric8b, RTP1S, and RTP2. (**e**) Cell line generation for stable accessory factor expression. After transfection, clones were isolated and screened for activation of ORs, MOR258-5 and OR7D4, that require accessory factors for functional expression. The dark purple bar represents the clone (ScL21) selected for further experiments.

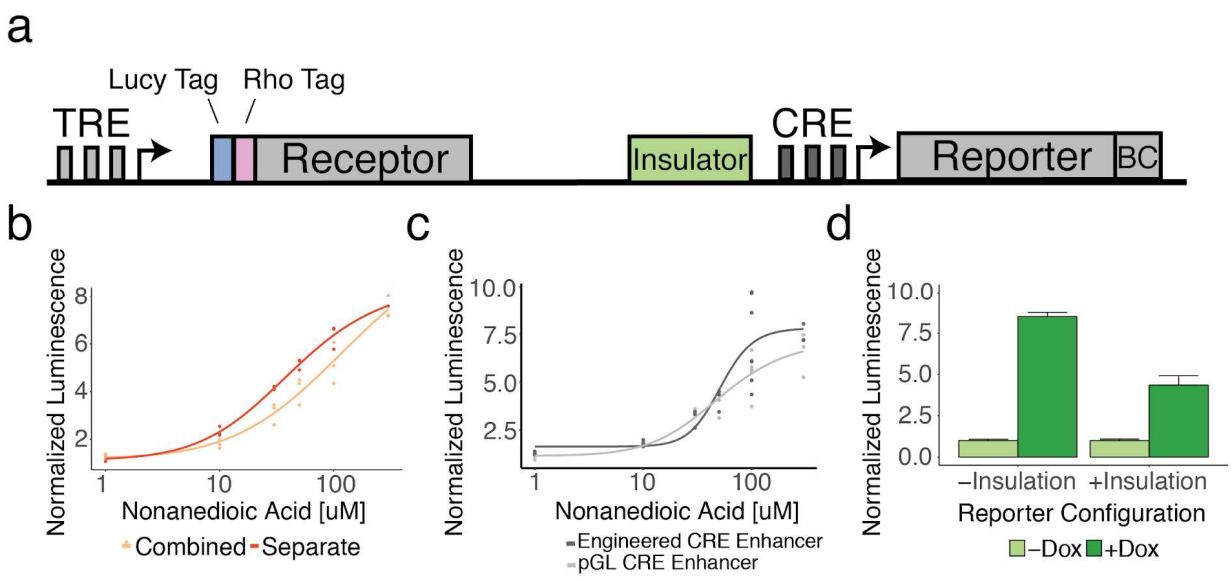


Figure 2.5. Design of a Multiplexed Genetic Reporter for OR Activation. (**a**) Annotated Vector map for plasmid containing the OR expression cassette and genetic reporter for integration. (**b**) MOR42-3 reporter activation in cells transiently co-expressing the receptor and genetic reporter on separate plasmids or together. (**c**) Fold activation of MOR42-3 driven by an engineered CRE enhancer (7 CREB binding sites) compared to Promega's pGL4.19 CRE enhancer. (**d**) Genetic reporter basal activation upon inducible expression of MOR42-3 with or without a DNA insulator upstream of the CRE enhancer.

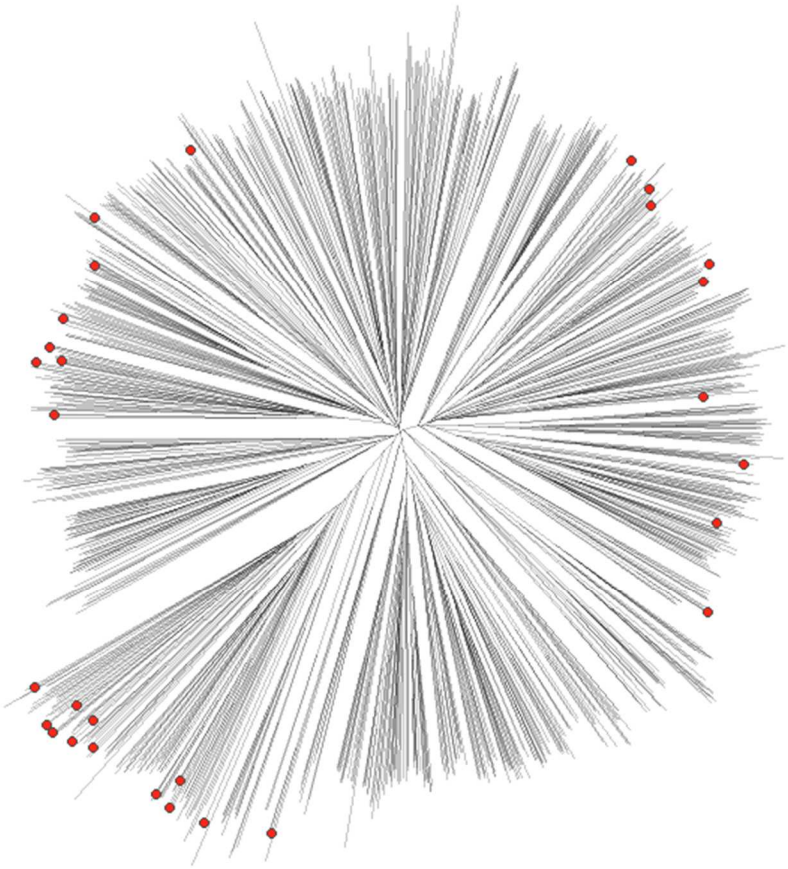


Figure 2.6. Evolutionary Tree of Mouse ORs. Phylogenetic tree inferred from amino acid sequence of functional murine ORs. The length of lines indicates degree of divergence between ORs. Red dots indicate ORs that were selected for inclusion in this study.

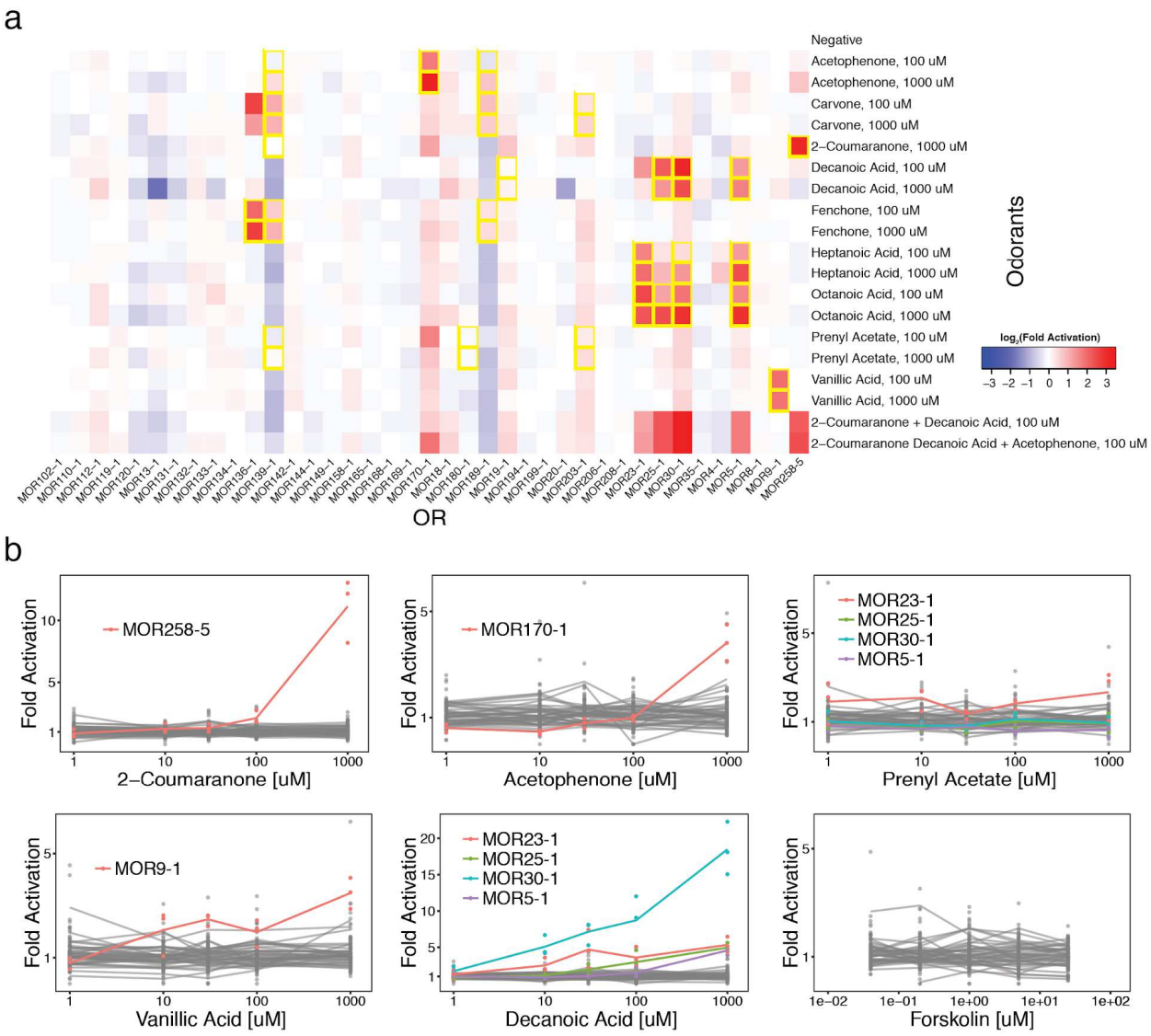


Figure 2.7. Pilot-Scale Recapitulation of Odorant Response in Multiplex. (**a**) Heatmap displaying 39 pooled receptors activity against 9 odorants and 2 mixtures. Interactions are colored by the $\log_2$-fold activation of the genetic reporter (see methods). Odorant interactions previously identified[12] are boxed in yellow. (**b**) Dose-response curves for odorants or forskolin (adenylate cyclase stimulator) at 5 concentrations screened against the OR library. Curves for ORs known to interact with the odorant are colored. Stimulation with forskolin does not show substantial differential activity between ORs in our assay.

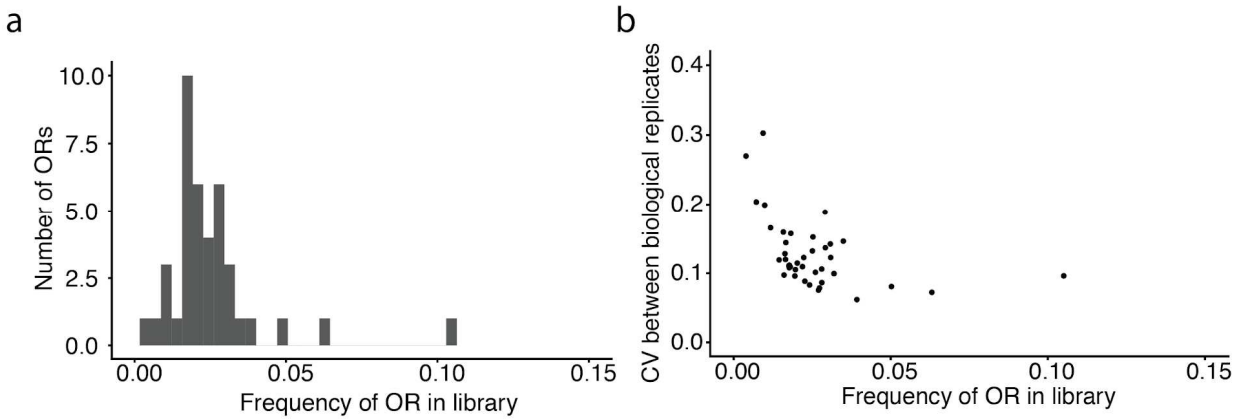


Figure 2.8. Library Representation. Representation of individual ORs in the library for the 39/42 ORs that had sufficient cellular coverage (see Methods). (**a**) Frequency of each OR as a fraction of the library determined by the relative activation of each reporter stimulated with DMSO. (**b**) The relationship between frequency of each OR in the library and the average coefficient of variation between biological replicate measurements of reporter activation for all conditions.

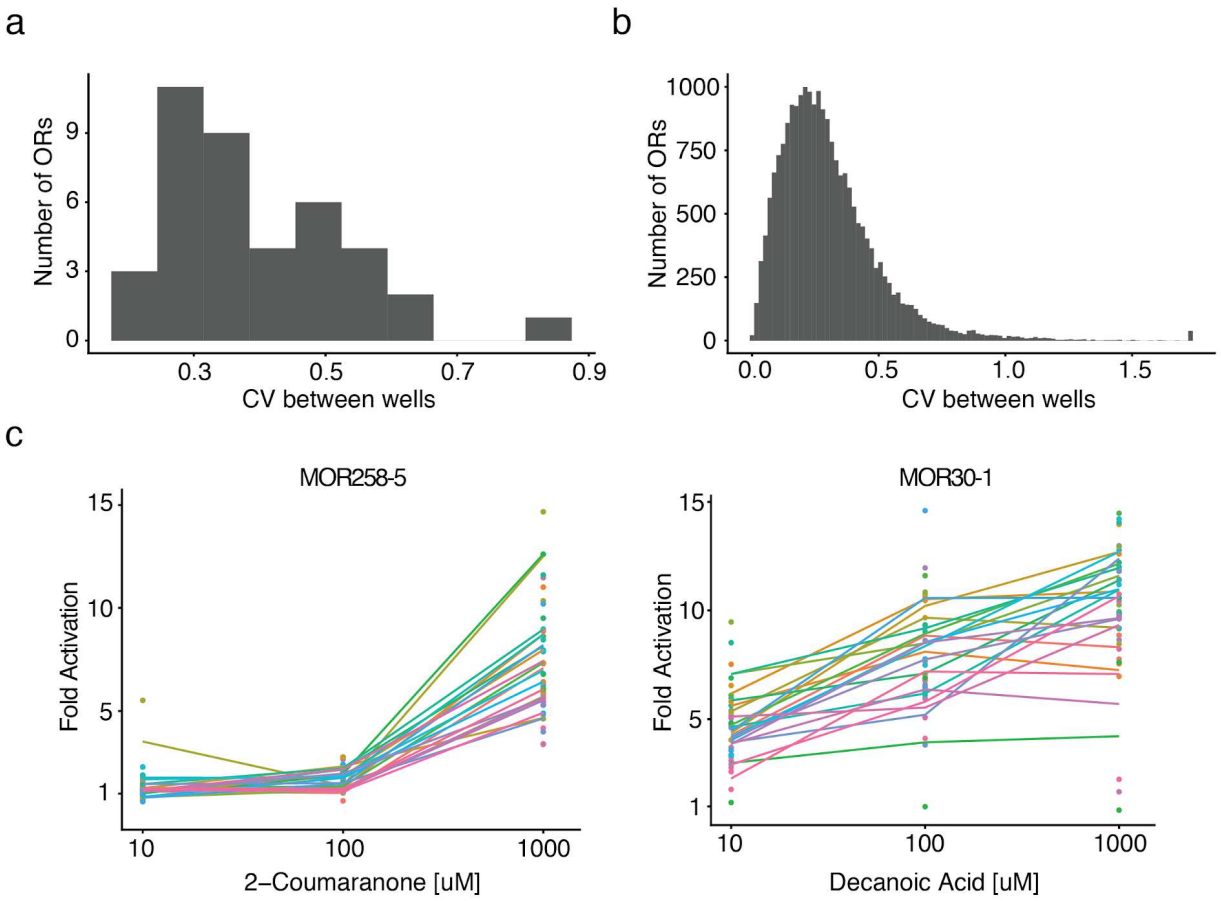


Figure 2.9. Replicability of the Large-Scale Multiplexed Screen. (**a**) Histogram displaying the distribution of the coefficient of variation for the OR library when stimulated with DMSO. (**b**) Histogram displaying the distribution of the coefficient of variation for the OR library against all conditions assayed. (**c**) Dose-response curves for the control odorants included on each 96-well plate assayed. Each color represents a different plate.

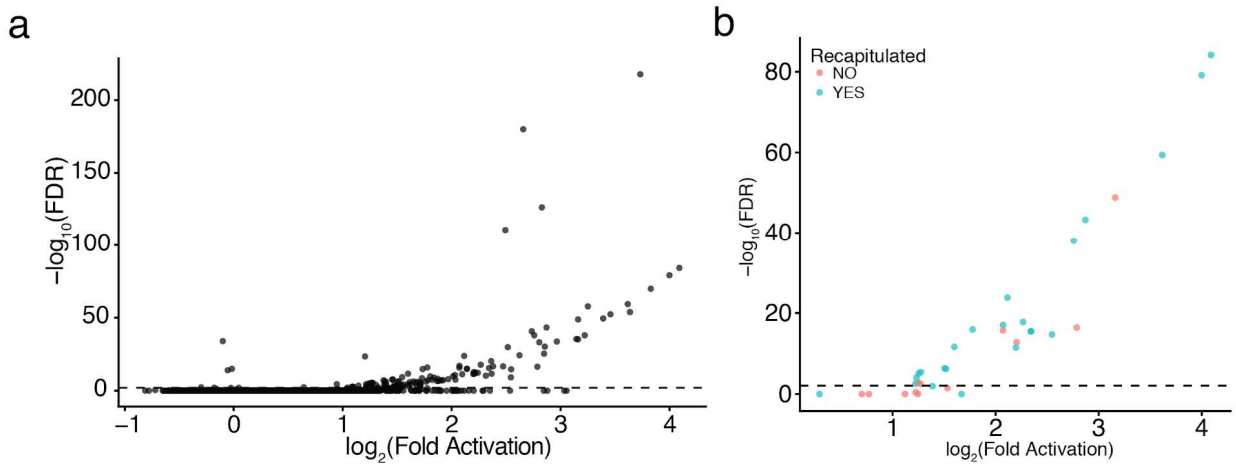
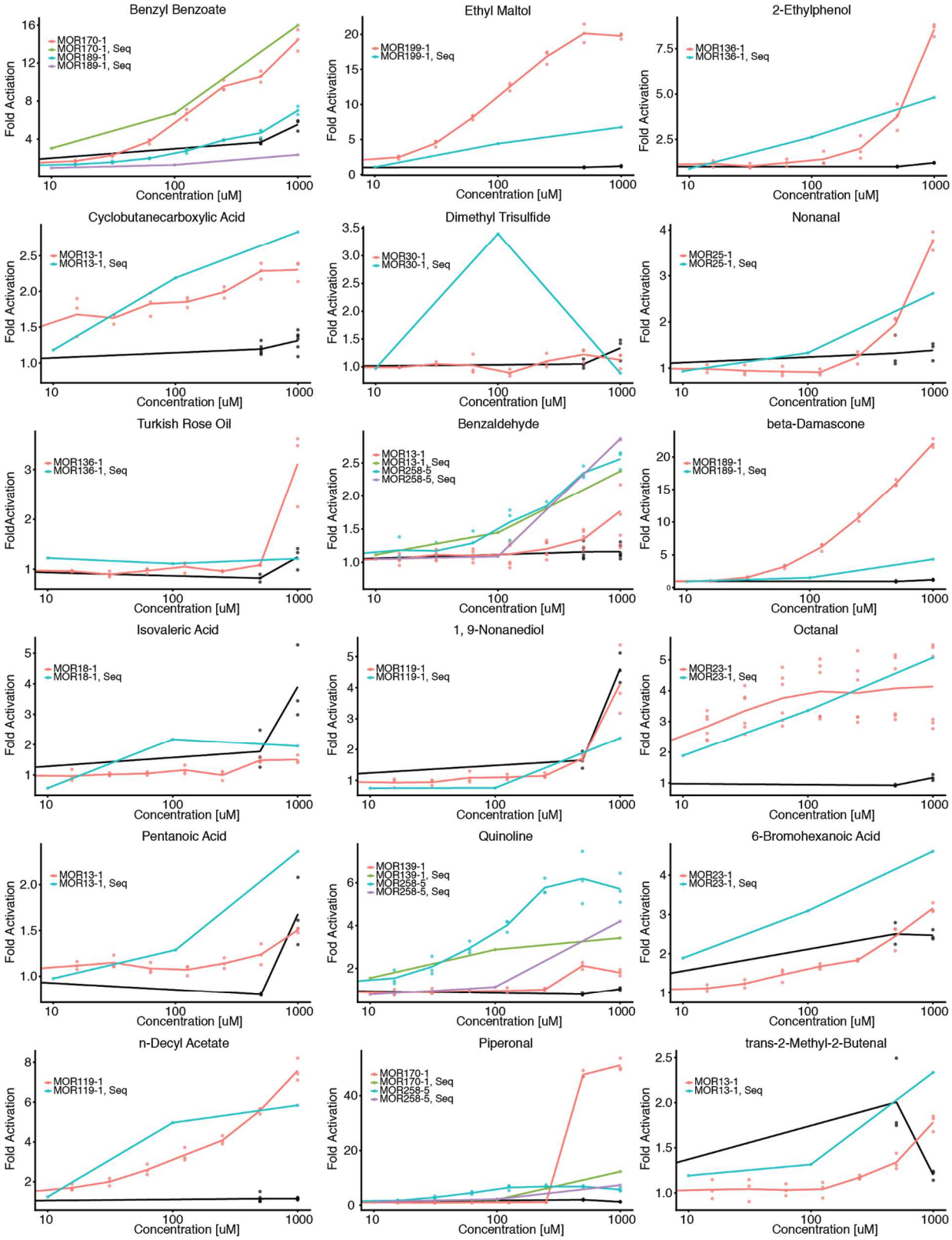


Figure 2.10. Significance and Fold Change of High-Throughput Assay Data. (**a**) The False Discovery Rate (FDR; Benjamini-Hochberg corrected, see Methods) plotted against the fold change for each OR-odorant interaction. The dashed line represents the 1% FDR, the cutoff used to identify positive interactions. (**b**) The subset of interactions tested by a follow-up orthogonal luciferase assay (color indicates whether it was recapitulated in the orthogonal system). Of the interactions passing a 1% FDR, 20 of 27 also showed interaction in the orthogonal follow-up assay.


Benzyl Benzoate
MOR170-1
MOR170-1, Seq
MOR189-1
MOR189-1, Seq
Fold Actiation
Concentration [uM]
Ethyl Maltol
MOR199-1
MOR199-1, Seq
Fold Activation
Concentration [uM]
2-Ethylphenol
MOR136-1
MOR136-1, Seq
Fold Activation
Concentration [uM]
Cyclobutanecarboxylic Acid
MOR13-1
MOR13-1, Seq
Fold Activation
Concentration [uM]
Dimethyl Trisulfide
MOR30-1
MOR30-1, Seq
Fold Activation
Concentration [uM]
Nonanal
MOR25-1
MOR25-1, Seq
Fold Activation
Concentration [uM]
Turkish Rose Oil
MOR136-1
MOR136-1, Seq
FoldActivation
Concentration [uM]
Benzaldehyde
MOR13-1
MOR13-1, Seq
MOR258-5
MOR258-5, Seq
Fold Activation
Concentration [uM]
beta-Damascone
MOR189-1
MOR189-1, Seq
Fold Activation
Concentration [uM]
Isovaleric Acid
MOR18-1
MOR18-1, Seq
Fold Activation
Concentration [uM]
1, 9-Nonanediol
MOR119-1
MOR119-1, Seq
Fold Activation
Concentration [uM]
Octanal
MOR23-1
MOR23-1, Seq
Fold Activation
Concentration [uM]
Pentanoic Acid
MOR13-1
MOR13-1, Seq
Fold Activation
Concentration [uM]
Quinoline
MOR139-1
MOR139-1, Seq
MOR258-5
MOR258-5, Seq
Fod Activation
Concentration [uM]
6-Bromohexanoic Acid
MOR23-1
MOR23-1, Seq
Fold Activation
Concentration [uM]
n-Decyl Acetate
MOR119-1
MOR119-1, Seq
Fold Activation
Concentration [uM]
Piperonal
MOR170-1
MOR170-1, Seq
MOR258-5
MOR258-5, Seq
Fold Activation
Concentration [uM]
trans-2-Methyl-2-Butenal
MOR13-1
MOR13-1, Seq
Fold Activation
Concentration [uM]

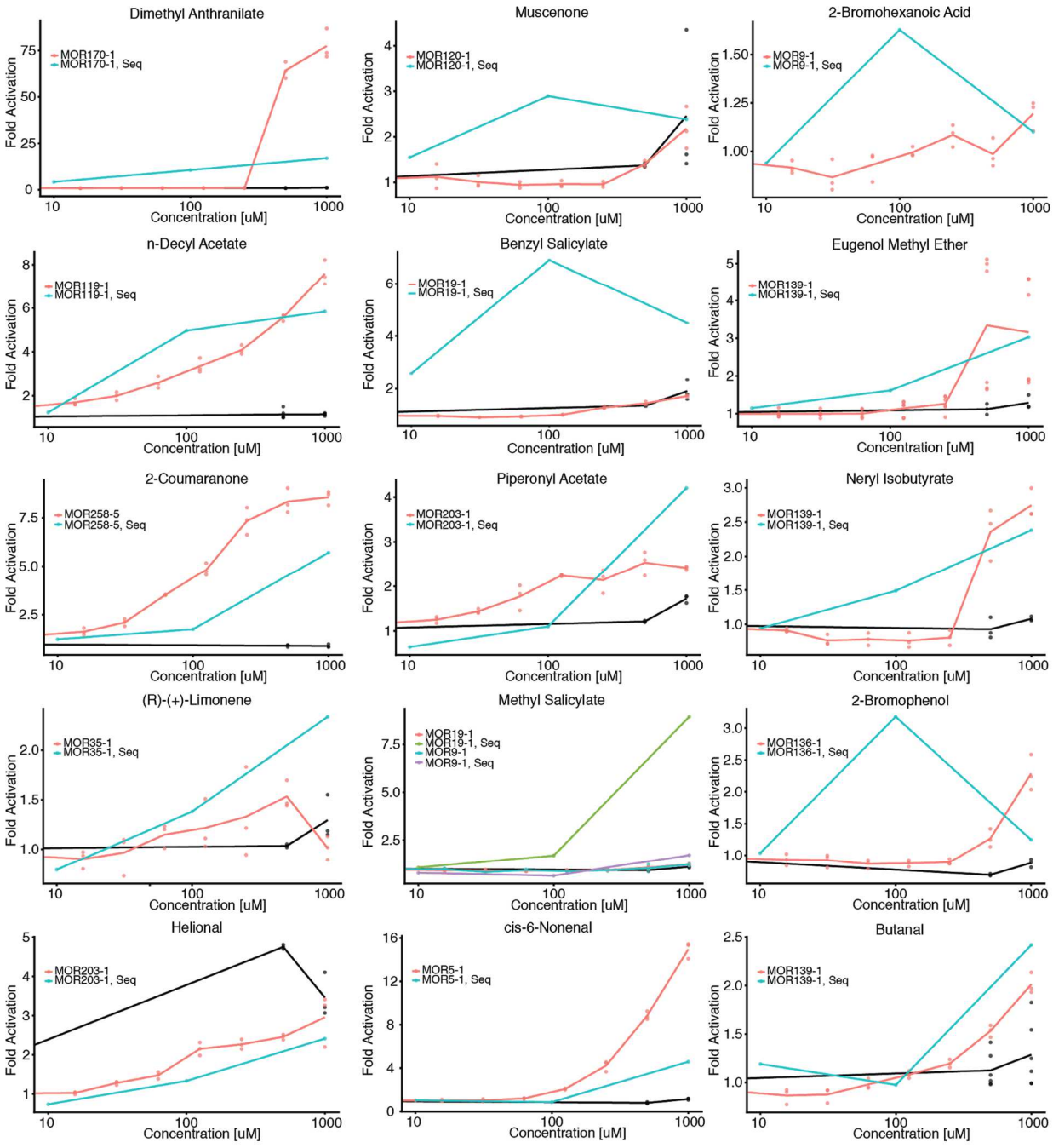


Figure 2.11. Recapitulation of the Screen in a Transient, Orthogonal System. Secondary screen of chemicals in a transient OR reporter activation system[26] with a luciferase reporter gene readout. Each plot shows the behavior of a control cell line expressing the reporter gene but no OR (black line), as well as a cell line expressing a specific OR and reporter gene. In addition, data from the high throughput screen (labeled as Seq) is plotted for reference.

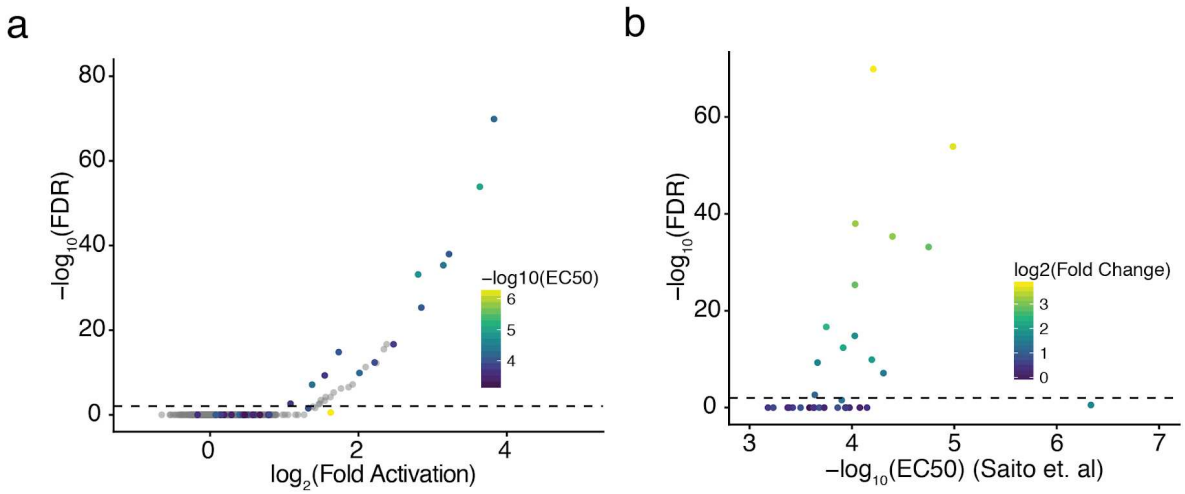


Figure 2.12. Assay Correspondence with Previously Screened Odorant-Receptor Pairs. (**a**) FDR plotted against fold induction for the 540 odorant-OR interactions that were previously tested by *Saito et al.*[12]. Points are colored by the EC_{50} of the interaction in the previous work. Grey points represent interactions not identified in the previous screen. Comparing the results from transient versus integrated luciferase assays revealed that, in some cases, the integrated system required a higher concentration of odorant to achieve significant activation, likely because of the lower DNA copy number of the CRE-driven luciferase and receptor. Since the highest concentration of odorant assayed was 1 mM, low affinity interactions may not have been detectable in this screen. (**b**) The FDR in the assay related to the EC_{50} of the hit from the previous screen, colored by the fold activation from the multiplexed screen.

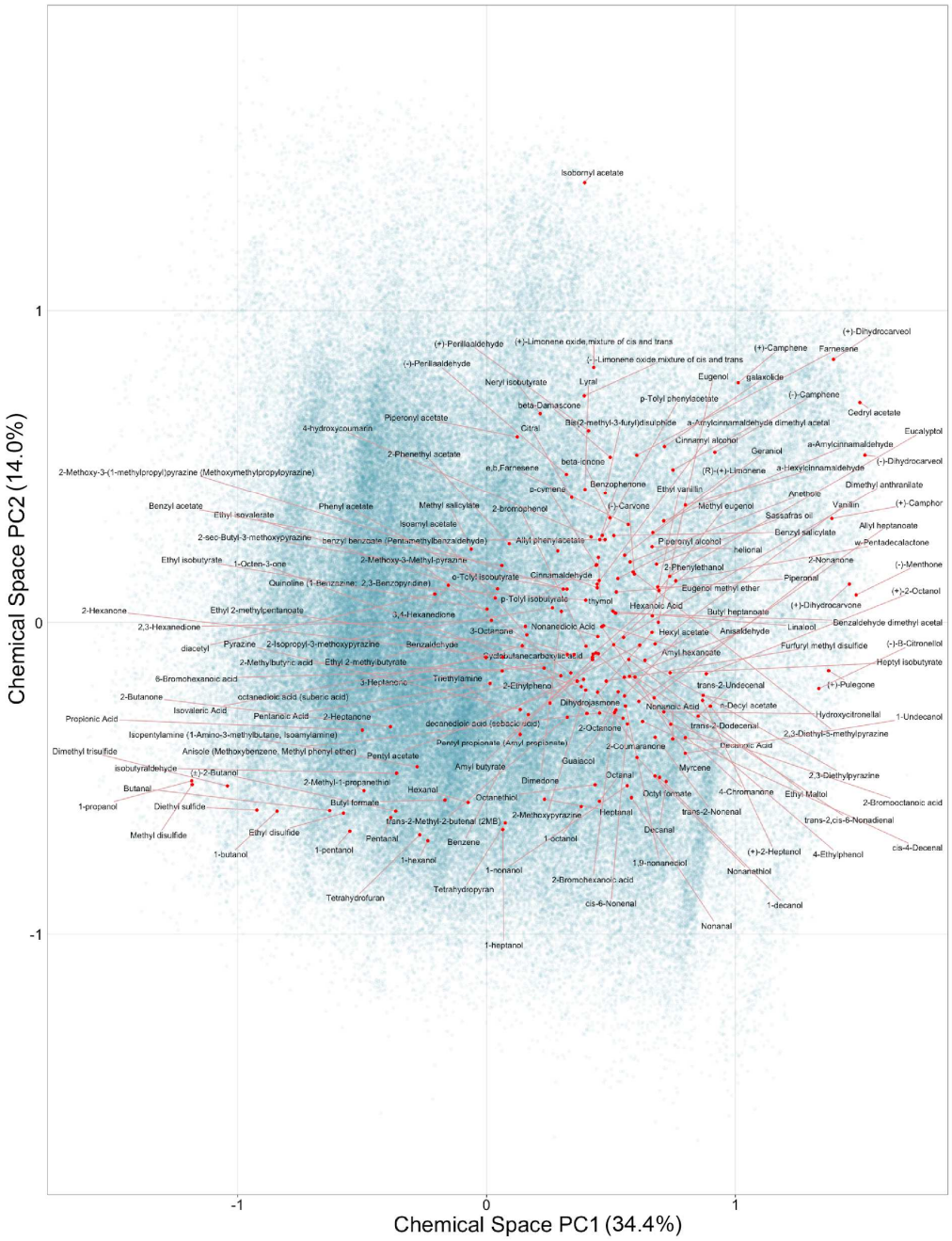


Figure 2.13. Location of Odors Tested with Respect to a Learned Chemical Space. Locations of the chemicals tested in this assay in chemical space. The molecular autoencoder[27] was used to generate a 292-dimensional representations of 250,000 randomly sampled molecules from the ChEMBL 23 database (blue) as well as the chemicals tested in our assay (red) projected onto two dimensions with Principal Component Analysis (PCA).

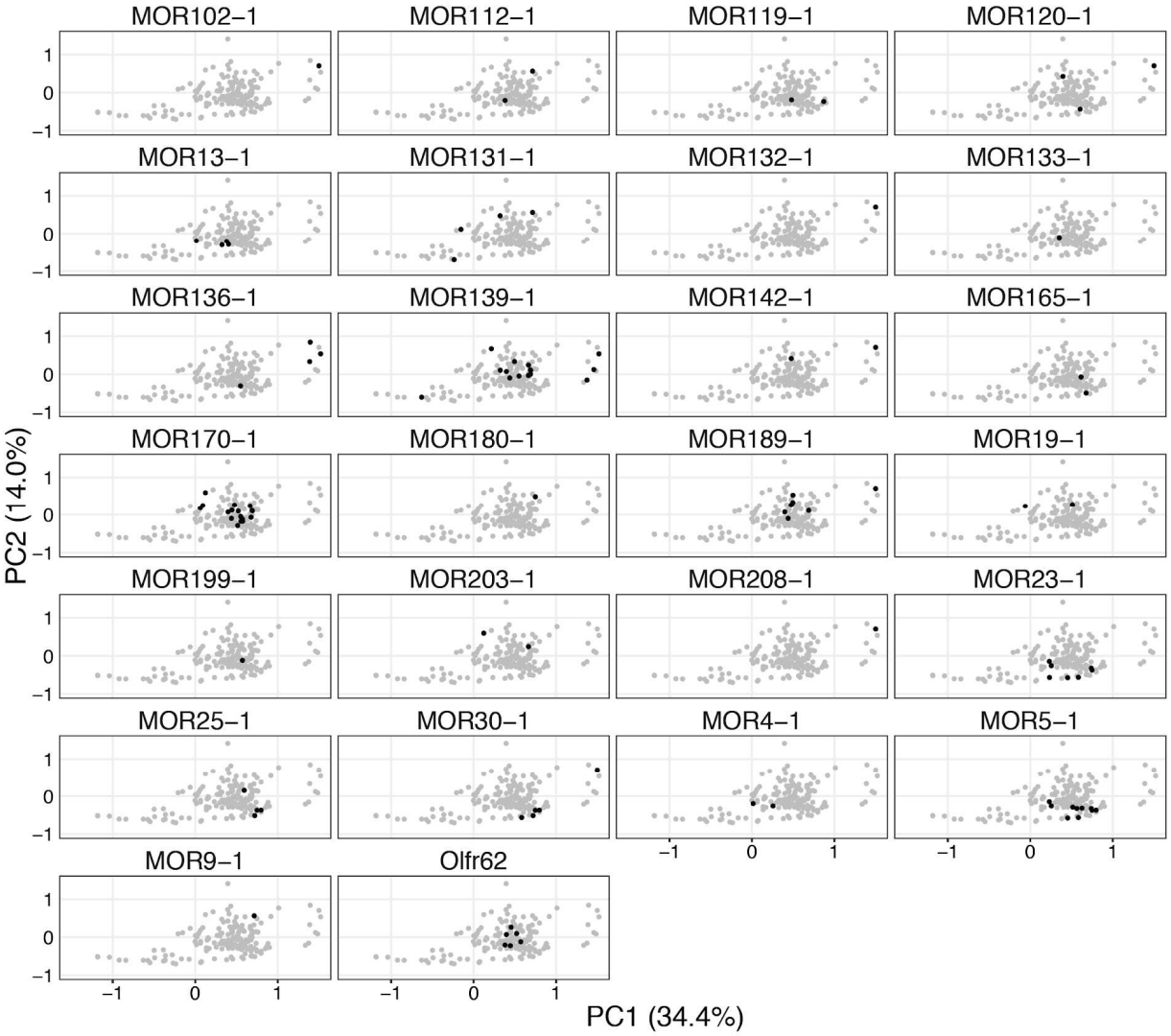


Figure 2.14. Clustering of Odorant Response for Receptors. The locations of any hits (black) with respect to other chemicals tested (grey) for each OR on the PCA projection as shown in Supplementary Fig. 10.

Materials and Methods:

Odorant-Receptor Activation Luciferase Assay (Transient)

The Dual-Glo Luciferase Assay System (Promega) was used to measure OR-odorant responses as previously described[26]. HEK293T cells (ATCC #11268) were plated in poly-D-lysine coated white 96-well plates (Corning) at a density of 7,333 cells per well in 100 ul DMEM+10% FBS (Thermo Fisher Scientific). 24 hours later, cells were transfected using lipofectamine 2000 (Thermo Fisher Scientific) with 5 ng/well of plasmids encoding ORs and 10 ng/well of luciferase driven by a cyclic AMP response element or 10 ng/well of a plasmid encoding both the OR and the luciferase gene, and in both cases 5 ng/well of a plasmid encoding *Renilla* luciferase. Experiments conducted with accessory factors included 5 ng/well of plasmids encoding RTP1S (Gene ID: 132112) and RTP2 (Gene ID: 344892). Inducibly expressed ORs were transfected with 1 ug/ml doxycycline (Sigma-Aldrich) added to the transfection media. 10-100 mM odorant stocks were established in DMSO or ethanol. 24 h after transfection, transfection medium was removed and replaced with 25 ul/well of the appropriate concentration of odorant diluted from the stocks into CD293 (Thermo Fisher Scientific). Four hours after odorant stimulation, the Dual-Glo Luciferase Assay kit was administered according to the manufacturer's instructions. Luminescence was measured using the M1000 plate reader (Tecan). All luminescence values were normalized to *Renilla* luciferase activity to control for transfection efficiency in a given well. Data were analyzed with Microsoft Excel and R.

Odorant-Receptor Activation Luciferase Assay (Integrated)

HEK293T and HEK293T derived cells integrated with the combined receptor/reporter plasmids were plated at a density of 7333 cells/well in 100 uL DMEM+10in poly-D-lysine coated 96-well plates. 24 hours later, 1 ug/ml doxycycline was added to the well medium. Odorant stimulation, luciferase reagent addition, and luminescence measurements were carried out in the same

manner as the transient assays. Constitutively expressed ORs were assayed in the same manner without doxycycline addition. Data were analyzed with Microsoft Excel and R.

Odor Stimulation and RNA Extraction for Pilot-Scale Multiplexed Odorant Screening

HEK293T and HEK293T-derived cells transposed with the combined receptor/reporter plasmid were plated at a density of 200k cells/well in a 6 well plate in 2 mL DMEM+10%FBS. 24 hours later, 1 ug/ml doxycycline was added to the well medium. 10-100 mM odorant stocks were diluted in DMSO or ethanol. 24 hours after doxycycline addition, odorants were diluted in OptiMEM and media was aspirated and replaced with 1 mL of the odorant-OptiMEM solution. 3 hours after odor stimulation, odor media was aspirated and 600 uL of buffer RLT (Qiagen) was added to each well. Cells were lysed with the Qiashredder Tissue and Cell Homogenizer (Qiagen),and RNA was purified using the RNEasy MiniPrep Kit (Qiagen) with the optional on-column DNAse step according to the manufacturer's protocol.

Pilot Scale Library Preparation and RNA-seq

5 ug of total RNA per sample was reverse transcribed with Superscript IV (Thermo-Fisher) using a gene specific primer for the barcoded reporter gene (OL003). The reaction conditions are as follows: annealing: [65°C for 5 min, 0°C for 1 min] extension: [52°C for 60 min, 80°C for 10 min]. 10% of the cDNA library volumes were amplified for 5 cycles (OL004F and R) using HiFi Master Mix (Kapa Biosystems). The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 5 cycles of 98°C for 20 seconds, 59°C for 15 seconds, and 72°C for 10 seconds, followed by an extension of 72°C for 1 minute. The PCR products were purified using the DNA Clean & Concentrator kit (Zymo Research) into 10 uL and 1 uL of each sample was amplified (OL005F and R) using the SYBR FAST qPCR Master mix (Kapa Biosystems) with a CFX Connect Thermocycler (Biorad) to determine the number of PCR cycles necessary for library amplification. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes,

40 cycles of 95°C for 3 seconds and 60°C for 20 seconds. After qPCR, 5 uL of the pre-amplified cDNA libraries were amplified a second time at the same cycling conditions as the first amplification with the same primers used for qPCR for 4 cycles greater than the previously determined Cq. The PCR products were then gel isolated from a 1% agarose gel with the Zymoclean Gel DNA Recovery Kit (Zymo Research). Library concentrations were quantified using a TapeStation 2200 (Agilent) and loaded at equimolar ratios onto a HiSeq 3000 with a 20% PhiX spike-in and sequenced with custom primers: Read 1 (OL003) and i7 Index (OL006).

Pilot Scale Data Analysis

To determine fold activation of each OR treated with each chemical, we first calculated the fraction of barcodes (composition) corresponding to each OR in the control treatment (DMSO). Then, we calculated the fold change in the composition of each OR in each a specific condition. As the barcode reads from activated ORs can dominate the composition of all reads and change the effective library size, we then normalized the activation of each OR by the median activation for each well. To be effective, this normalization assumes that fewer than half of the ORs are activated by an odorant.

OR Library Cloning

The backbone plasmid (all genetic elements except the OR and barcode) was created using isothermal assembly with the Gibson Assembly HiFi Mastermix (SGI-DNA). A short fragment was amplified with a primer containing 15 random nucleotides to create the barcode sequence (OL007F and R) using HiFi Master Mix. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 35 cycles of 98°C for 20 seconds, 60°C for 15 seconds, and 72°C for 20 seconds, followed by an extension of 72°C for 1 minute. The amplicon and the backbone plasmid were digested with restriction enzymes MluI and AgeI (New England Biolabs) and

ligated together with T4 DNA ligase (New England Biolabs). DH5α *E.coli* competent cells (New England Biolabs)
were transformed directly into liquid culture with antibiotic to maintain the diversity of the barcode library.

OR genes were received as a gift from Hiro Matsunami. OR genes were amplified individually with primers (OL008) adding homology to the barcoded backbone plasmid using HiFi Master Mix. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 35 cycles of 98°C for 20 seconds, 61°C for 15 seconds, and 72°C for 30 seconds, followed by an extension of 72°C for 1 minute. The amplified ORs were purified with DNA Clean and Concentrator Kit (Zymo Research) and pooled together. The barcoded backbone plasmid was digested with NdeI and SbfI and the OR amplicon pool was cloned into it using isothermal assembly with the Gibson Assembly Hifi Mastermix. DH5α *E.coli* competent cells were transformed with the assembly and antibiotic resistant clones were picked and grown up in 96-well plates overnight. The plasmid DNA was prepped with the Zyppy -96 Plasmid Miniprep Kit (Zymo Research). Plasmids were Sanger sequenced (OL109-111) both to associate the barcode with the reporter gene and identify error-free ORs.

OR Library Genomic Integration

HEK293T cells and HEK293T-derived cells were seeded at a density of 350k cells/well in a 6-well plate in 2 mL DMEM+10% FBS. 24 hours after seeding, cells were transfected with plasmids encoding receptor/reporter transposon and the Super PiggyBac Transposase (Systems Bioscience) according to the manufacturer's instructions. 1 ug of transposon DNA and 200 ng of transposase DNA were transfected per well with Lipofectamine 3000 (Thermo Fisher Scientific). 3 days after transfection, cells were passaged 1:10 into a 6-well plate, and one day after passaging 8 ug/mL blasticidin were added to the cells. Cells were grown with selection for

7-10 days. The OR library was transposed individually and pooled together at equal cell numbers.

Accessory Factor Cell Line Generation

HEK293T derived cells were transposed with plasmids encoding the accessory factor genes RTP1S, RTP2, $G\alpha_{olf}$ (Gene ID: 2774), and Ric8b (Gene ID: 237422) inducibly driven by the Tet-On promoter pooled equimolar according to the transposition protocol in the OR Library Integration section. Cells were selected with 2 ug/mL puromycin (Thermo Fisher). After selection, cells were seeded in a 96-well plate at a density of 0.5 cells/well. Wells were examined for single colonies after 3 days and expanded to 24-well plates after 7 days. Clones were screened for accessory factor expression by screening them for robust activation of MOR258-5/Olfr62 and OR7D4 with a transient luciferase assay (Supplementary Fig. 1). The clone with the highest fold activation for both receptors and no salient growth defects was established for the multiplexed screen.

Transposon Copy Number Verification

gDNA was purified from cells transposed with the OR reporter vector and from cells containing the single copy landing pad with the Quick-gDNA Miniprep kit. 50 ng of gDNA was amplified with primers annealing to the regions of the exogenous DNA from each sample using the SYBR FAST qPCR Master Mix (Kapa Biosystems) on a CFX Connect Thermocycler using the manufacturer's protocol. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 40 cycles of 95°C for 3 seconds and 60°C for 20 seconds. Cq values for the transposed ORs were normalized to the single copy landing pad to determine copy number.

Lentiviral Transduction

Lentiviral vector was produced by transient transfection of 293T cells with lentiviral transfer plasmid, pCMVΔR8.91 and pCAGGS-VSV-G using Mirus TransIT-293. HEK293T cells were transduced to express the m2rtTA transcription factor (Tet-On) at 50% confluency and seeded one day prior to transduction. Clones were isolated by seeding cells in a 96-well plate at a density of 0.5 cells/well. Wells were examined for single colonies after 7 days and expanded to 24 well plates. Clones were assessed for m2rtTA expression by screening for robust activation of MOR42-3 (Gene ID: 257926) with a transient luciferase assay.

High-throughput Odorant Screening

The OR library cell line was thawed from a liquid nitrogen frozen stock into a T-225 flask (Corning) three days before seeding into a 96-well plate for screening. The library was seeded at 6,666 cells per well in 100 uL of DMEM+10% FBS. 24 hours later a working concentration of 1 ug/mL of doxycycline in DMEM+10% FBS was added to the wells. 24 hours after induction, the media was removed from each plate and replaced with 25 ul of odorant diluted in OptiMEM. Each odor was added at three different concentrations (10 uM, 100 uM, 1 mM) in triplicate with the same amount of final DMSO (1%). Each plate contained two control odorants at a three concentration (10 uM, 100 uM, 1 mM) in triplicate and three wells containing 1% DMSO dissolved in media. The library was incubated with odorants for three hours in a cell culture incubator with the lids removed.

After odor incubation, media was pipetted out of the plates and cells were lysed by adding 25 uL of ice-cold Cells-to-cDNA II Lysis Buffer (Thermo Fisher) and pipetting up and down to homogenize and lyse cells. The lysate was then heated to 75°C for 15 minutes and flash frozen with liquid nitrogen and kept at -80C until further processing. Then 0.5 uL DNase I (New England Biolabs) was added to lysate, and incubated at 37°C for 15 minutes. To anneal the RT primer, 5 ul of lysate from each well was combined with 2.5 uL of 10 mM dNTPs (New England

Biosciences), 1 uL of 2 uM gene specific RT primer (OL003), and 1.5 uL of H2O. The reaction was heated to 65°C for 5 min and cooled back down to 0°C. After annealing, 1 uL of M-MuLV Reverse Transcriptase (Enzymatics), 1 uL of buffer, and 0.25 ul of RNase Inhibitor (Enzymatics) were added to each reaction. Reactions were incubated at 42°C for 60 min and the RT enzyme was heat inactivated at 85°C for 10 min.

For each batch, qPCR was performed on a few wells (OL005F and OL013) with SYBR FAST qPCR Mastermix to determine the number of cycles necessary for PCR based library preparation. The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 40 cycles of 95°C for 3 seconds and 60°C for 20 seconds. After qPCR, 5 uL of each RT reaction was combined with 0.4 uL of 10 uM primers containing sequencing adaptors (OL005F and OL013), 10 uL of NEB-Next Q5 Mastermix (New England Biosciences) and 4.2 uL H2O, the PCR was carried out according to the manufacturer's protocol. The forward primer contains the P7 adaptor sequence and an index identifying the well in the assay and the reverse primer contains the P5 adaptor sequence and an index identifying the plate in the assay. PCR products were pooled together by plate and purified with the DNA Clean and Concentrator Kit. Library concentrations were quantified using a TapeStation 2200 and a Qubit (Thermo Fisher). The libraries were sequenced with two index reads and a single end 75-bp read on a NextSeq 500 in high-output mode (Illumina).

Analysis of Next-Generation Sequencing Data

Samples were identified via indexing by their PCR index adapters unique for each well (5' end) and unique for each plate (3' end). The well barcodes followed the 7bp indexing scheme in (Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing Matthias Meyer, Martin Kircher, Cold Spring Harb Protoc; 2010; doi:10.1101/pdb.prot5448). The plate indexing scheme followed the Illumina indexing scheme. Sequencing data was

demultiplexed, and 15bp barcode sequences were counted with only exact matches by custom python and bash scripts.

Statistical Methods for Calling Hits

Count data was then analyzed using the differential expression package EdgeR[24]. To filter out ORs with low representation, we empirically set a cutoff that an OR had to contain at least 0.5% of the reads from more than 399 of the 1954 test samples. This filtered out 3 of 42 ORs which were underrepresented in the cell library (MOR172-1, MOR176-1 and MOR181-1).

Normalization factors were determined using the EdgeR package function calcNormFactors, and glmFit was used with the dispersion set to the tagwise dispersion, since only 39 ORs were present in the library and trended dispersion values did fit the data well. By fitting a generalized linear model to the count data to determine if odorants stimulated specific ORs, we were able to determine both the mean activation for each OR-odorant interaction and the p-value. We then corrected this p-value for multiple hypothesis testing using the built in p.adjust function with the Benjamini & Hochberg correction[25], yielding a False Discovery Rate (FDR). We set a cutoff of 1% to determine interacting odorant-OR pairs. For each interaction between an odorant and an OR, we further required that an OR-odorant interaction was above the cutoff in two different concentrations of odorant or in just the 1000 uM concentration.

Molecular Autoencoder

We used an autoencoder as described in Gómez-Bombarelli *et al.* to visualize OR-chemical interactions in the context of chemical space[27]. Following the authors' advice, we used a reimplementation of autoencoder as the original implementation requires a defunct Python package (https://github.com/chembl/autoencoder_ipython). This model comes pre-trained to a validation accuracy of 0.99 on the entire ChEMBL 23 database with the exception of molecules

whose SMILES are longer than 120 characters. We used this pretrained model to generate the latent representations of the 168 chemicals for which we could find SMILES representations and 250,000 randomly sampled chemicals from ChEMBL 23. We then used scikit-learn[31] to perform principal component analysis to project the resulting matrix onto two dimensions.

Supplementary Tables

Table 2.1: Olfactory receptors screened in this study

Mouse Olfactory Receptor Convention	Olfactory Receptor Convention
MOR102-1	Olfr1325
MOR110-1	Olfr812
MOR112-1	Olfr790
MOR119-1	Olfr214
MOR120-1	Olfr459
MOR13-1	Olfr644
MOR131-1	Olfr161
MOR132-1	Olfr24
MOR133-1	Olfr406
MOR8-1	Olfr575
MOR20-1	Olfr613
MOR203-1	Olfr992
MOR206-1	Olfr1098
MOR208-1	Olfr1413
MOR23-1	Olfr599
MOR25-1	Olfr554
MOR30-1	Olfr569

MOR35-1	Olfr686
MOR4-1	Olfr620
MOR5-1	Olfr638
MOR168-1	Olfr916
MOR169-1	Olfr902
MOR170-1	Olfr895
MOR18-1	Olfr558
MOR180-1	Olfr1019
MOR189-1	Olfr1079
MOR19-1	Olfr616
MOR194-1	Olfr1046
MOR199-1	Olfr1032
MOR258-5	Olfr62
MOR134-1	Olfr356
MOR136-1	Olfr340
MOR139-1	Olfr1352
MOR142-1	Olfr1356
MOR144-1	Olfr39
MOR149-1	Olfr828
MOR158-1	Olfr362
MOR165-1	Olfr909
MOR9-1	Olfr609

Table 2.2: Odorants screened in this study

Pentanoic Acid	Hexanoic Acid	1-nonanol	Nonanal
4-hydroxycoumarin	Dimedone	1-decanol	Decanal
4-Chromanone	(-)-Menthone	(+)-2-Heptanol	Citral
2-Butanone	beta-ionone	(+)-2-Octanol	Hydroxycitronellal
2-Hexanone	Pentyl acetate	(-)-B-Citronellol	Lyral

2-Heptanone	Allyl heptanoate	Geraniol	Acetophenone
3-Heptanone	Amyl hexanoate	Linalool	Control_1
2-Octanone	Nonanoic Acid	1-Undecanol	Control_2
3-Octanone	Amyl butyrate	Allyl phenylacetate	Decanoic_Acid
Propionic Acid	Butyl heptanoate	Benzene	DMSO
2_coumaranone	Heptyl isobutyrate	Benzyl acetate	Prenyl_Acetate
2-Nonanone	Hexyl acetate	Phenyl acetate	Vanillic_Acid
2,3-Hexanedione	Butyl formate	Octanethiol	a-Amylcinnamaldehyde
3,4-Hexanedione	Ethyl isobutyrate	Nonanedioic Acid	Eucalyptol
(-)-Carvone	1-butanol	Nonanethiol	Pentyl propionate (Amyl propionate)
(+)-Dihydrocarvone	Isovaleric Acid	Butanal	Dihydro Myrcenol
(+)-Camphor	1-propanol	Pentanal	Muscenone
Dihydrojasmone	1-hexanol	Hexanal	ethyl maltol
Benzophenone	1-heptanol	Heptanal	calone
(+)-Pulegone	1-octanol	Octanal	Sandalwood Mysone
Iso E Super	w-Pentadecalactone	benzyl benzoate (Pentamethylbenzaldehyde)	Ethyl 2-methylbutyrate
Olibanum Coeur MD	2-Phenylethanol	Piperonyl alcohol	trans-2-Dodecenal
Turkish Rose Oil	2-Phenethyl acetate	Piperonyl acetate	Cedryl acetate
Angel Eau de parfum (10 uM)	Piperonal	Tetrahydrofuran	1-Octen-3-one
a-Hexylcinnamaldehyde	Pyrazine	Tetrahydropyran	2-Bromohexanoic acid
Dior Jadone Eau de parfum	Sassafras oil	Benzaldehyde dimethyl acetal	6-Bromohexanoic acid
Flowerbomb Viktor and Rolf	thymol	Â 2-Methyl-1-propanethiol	2-Bromooctanoic acid
Chanel No 5	Triethylamine	(+)-Dihydrocarveol	Furfuryl methyl disulfide
Axe	L-Turpentine	(-)-Dihydrocarveol	Ethyl isovalerate
Aedione	Anisaldehyde	(+)-Perillaaldehyde	Bis(2-methyl-3-furyl)disulphide)
Isobornyl acetate	[Di]ethyl sulfide	(-)-Perillaaldehyde	Dimethyl trisulfide

a-Amylcinnamaldehyde dimethyl acetal	Eugenol	Benzyl salicylate	trans-2,cis-6-Nonadienal
p-Tolyl isobutyrate	Eugenol methyl ether	(+)-Limonene oxide,mixture of cis and trans	trans-2-Nonenal
o-Tolyl isobutyrate	4-Ethylphenol	(-)-Limonene oxide,mixture of cis and trans	Cinnamyl alcohol
p-Tolyl phenylacetate	Ethyl vanillin	(R)-(+)-Limonene	n-Decyl acetate
2-Methoxy-3-Methyl-pyrazine	Vanillin	(-)-Camphene	Dimethyl anthranilate
2-Methoxypyrazine	2-Ethylphenol	(+)-Camphene	trans-2-Undecenal
Methyl salicylate	Guaiacol	2,3-Diethyl-5-methylpyrazine	Neryl isobutyrate
Anethole	2-bromophenol	Ethyl disulfide	cis-4-Decenal
Myrcene	Benzaldehyde	Methyl disulfide	Octyl formate
(Â±)-2-Butanol	2,3-Diethylpyrazine	trans-2-Methyl-2-butenal (2MB)	p-cymene
2-Isopropyl-3-methoxypyrazine	2-Methylbutyric acid	diacetyl	helional
2-sec-Butyl-3-methoxypyrazine	Cyclobutanecarboxylic acid	galaxolide	1,9-nonanediol
cis-6-Nonenal	Isopentylamine (1-Amino-3-methylbutane, Isoamylamine)	isobutyraldehyde	octanedioic acid (suberic acid)
Cinnamaldehyde	Quinoline (1-Benzazine; 2,3-Benzopyridine)	Ethyl 2-methylpentanoate	decanedioic acid (sebacic acid)
beta-Damascone	Farnesene	e,b,Farnesene	Anisole (Methoxybenzene, Methyl phenyl ether)

Table 2.3: Odorant-receptor pairs called as hits.

OR	**Odorant**	**Minimum Activating Concentration (uM)**	**Previously Orphan Receptor?**

MOR102-1	Cedryl acetate	1000	YES
MOR112-1	Benzaldehyde	1000	YES
MOR112-1	galaxolide	100	YES
MOR119-1	Axe (10 uM)	1000	YES
MOR119-1	Furfuryl methyl disulfide	1000	YES
MOR119-1	n-Decyl acetate	100	YES
MOR120-1	Cedryl acetate	1000	YES
MOR120-1	Lyral	1000	YES
MOR120-1	Nonanethiol	1000	YES
MOR13-1	Benzaldehyde	1000	YES
MOR13-1	Cyclobutanecarboxylic acid	1000	YES
MOR13-1	Pentanoic Acid	1000	YES
MOR13-1	trans-2-Methyl-2-butenal (2MB)	1000	YES
MOR131-1	(-)-Perillaldehyde	1000	YES
MOR131-1	1-hexanol	1000	YES
MOR131-1	3,4-Hexanedione	1000	YES
MOR131-1	galaxolide	1000	YES
MOR132-1	Cedryl acetate	1000	YES
MOR133-1	3-Octanone	1000	YES
MOR134-1	Chanel No 5 (10 uM)	1000	YES
MOR136-1	(-)-Dihydrocarveol	1000	NO
MOR136-1	(+)-Camphor	100	NO

MOR136-1	(+)-Dihydrocarveol	1000	NO
MOR136-1	2-Ethylphenol	100	NO
MOR136-1	Olibanum Coeur MD	1000	NO
MOR139-1	(-)-Dihydrocarveol	1000	NO
MOR139-1	(+)-Dihydrocarvone	1000	NO
MOR139-1	(+)-Pulegone	1000	NO
MOR139-1	2-sec-Butyl-3-methoxypyrazine	1000	NO
MOR139-1	4-Chromanone	1000	NO
MOR139-1	beta-ionone	1000	NO
MOR139-1	Butanal	1000	NO
MOR139-1	Dihydrojasmone	1000	NO
MOR139-1	Dimethyl anthranilate	1000	NO
MOR139-1	Eugenol	1000	NO
MOR139-1	Eugenol methyl ether	1000	NO
MOR139-1	helional	1000	NO
MOR139-1	Neryl isobutyrate	1000	NO
MOR139-1	Quinoline (1-Benzazine; 2,3-Benzopyridine)	100	NO
MOR142-1	Bis(2-methyl-3-furyl)disulphide)	1000	YES
MOR142-1	Cedryl acetate	1000	YES

MOR158-1	Iso E Super	1000	YES
MOR165-1	decanedioic acid (sebacic acid)	1000	YES
MOR165-1	Octyl formate	1000	YES
MOR170-1	2-Bromohexanoic acid	1000	NO
MOR170-1	2-Phenethyl acetate	1000	NO
MOR170-1	4-Chromanone	100	NO
MOR170-1	4-Ethylphenol	1000	NO
MOR170-1	Anisaldehyde	1000	NO
MOR170-1	Benzyl acetate	1000	NO
MOR170-1	benzyl benzoate (Pentamethylbenzaldehyde)	10	NO
MOR170-1	Chanel No 5 (10 uM)	1000	NO
MOR170-1	Cinnamyl alcohol	1000	NO
MOR170-1	Dimethyl anthranilate	10	NO
MOR170-1	ethyl maltol	1000	NO
MOR170-1	Eugenol methyl ether	10	NO
MOR170-1	helional	1000	NO
MOR170-1	Piperonal	1000	NO

MOR170-1	Piperonyl acetate	1000	NO
MOR170-1	Quinoline (1-Benzazine; 2,3-Benzopyridine)	100	NO
MOR170-1	Vanillin	1000	NO
MOR180-1	a-Amylcinnamaldehyde dimethyl acetal	1000	NO
MOR180-1	Axe (10 uM)	1000	NO
MOR189-1	4-Chromanone	1000	NO
MOR189-1	benzyl benzoate (Pentamethylbenzaldehyde)	1000	NO
MOR189-1	beta-Damascone	1000	NO
MOR189-1	beta-ionone	1000	NO
MOR189-1	Cedryl acetate	1000	NO
MOR189-1	Eugenol methyl ether	1000	NO
MOR189-1	Quinoline (1-Benzazine; 2,3-Benzopyridine)	1000	NO
MOR19-1	Benzyl salicylate	10	YES
MOR19-1	Methyl salicylate	1000	YES
MOR199-1	ethyl maltol	100	YES
MOR203-1	helional	1000	NO
MOR203-1	Piperonyl acetate	1000	NO
MOR208-1	Cedryl acetate	1000	YES
MOR23-1	2-Bromooctanoic acid	1000	NO

MOR23-1	6-Bromohexanoic acid	100	NO
MOR23-1	Heptanal	1000	NO
MOR23-1	Hexanoic Acid	1000	NO
MOR23-1	Nonanal	1000	NO
MOR23-1	Nonanoic Acid	1000	NO
MOR23-1	Octanal	100	NO
MOR25-1	(-)-Carvone	1000	NO
MOR25-1	Decanal	1000	NO
MOR25-1	Decanoic-Acid	100	NO
MOR25-1	Nonanoic Acid	1000	NO
MOR30-1	Cedryl acetate	1000	NO
MOR30-1	Decanal	100	NO
MOR30-1	Decanoic-Acid	10	NO
MOR30-1	Nonanal	1000	NO
MOR30-1	Nonanoic Acid	100	NO
MOR4-1	Hexanoic Acid	1000	NO
MOR4-1	Pentanoic Acid	1000	NO
MOR5-1	2-Bromohexanoic acid	1000	NO
MOR5-1	2-Bromooctanoic acid	1000	NO

MOR5-1	6-Bromohexanoic acid	1000	NO
MOR5-1	cis-4-Decenal	1000	NO
MOR5-1	cis-6-Nonenal	1000	NO
MOR5-1	Decanoic-Acid	1000	NO
MOR5-1	Hexanoic Acid	1000	NO
MOR5-1	Nonanal	1000	NO
MOR5-1	Nonanoic Acid	100	NO
MOR5-1	Octanal	1000	NO
MOR5-1	Olibanum Coeur MD	1000	NO
MOR258-5	2-coumaranone	1000	NO
MOR258-5	Benzaldehyde	1000	NO
MOR258-5	Benzophenone	1000	NO
MOR258-5	ethyl maltol	1000	NO
MOR258-5	Piperonal	1000	NO
MOR258-5	Quinoline (1-Benzazine; 2,3-Benzopyridine)	1000	NO
MOR9-1	galaxolide	1000	NO

Supplementary Table 2.4: Primers and Sequences Used in This Study

Primer	Sequence	Description
OL001	CCCTTTAATCAGATGCGTCG	Gene Specific RT, Reporter Gene, for Q-RTPCR
OL002	CTGCCTGCTTCACCACCTTC	Gene Specific RT, GAPDH
OL003	AAGTGCCTTCCTGCCCTTTAATCAGATGCGTCG	Gene Specific RT, Reporter Gene, for RNA-seq, Also NGS Read1 Primer
OL004F	CGCCGAAGTGAAAACCACCTA	Pilot-Scale RNA-seq Round 1 Library Prep Amplification

OL004R	AAGTGCCTTCCTGCCCTTTAA	Pilot-Scale RNA-seq Round 1 Library Prep Amplification
OL005F	CAAGCAGAAGACGGCATACGAG AT NNNNNNNN CGAAGTGAAAACCACCTA	P7+i7index+primer for RNAseq library amplification
OL005R	AATGATACGGCGACCACCGAGA TCTACACAAGTGCCTTCCTGCCC TTTAA	P5+Read1+primer for pilot-scale RNAseq library amplification
OL006	CGGGTTTCTTGGCCTTGTAGGT GGTTTTCACTTCG	i7 index read primer, pilot-scale experiment
OL007F	ggaataACGCGTNNNNNNNNNNN NNNNCGACGCATCTGATTAAAG GG	Amplification of fragment containing barcode to be cloned into reporter plasmid
OL007R	ggaaggACCGGTtctagtcaaggcactat acat	Amplification of fragment containing barcode to be cloned into reporter plasmid
OL008F	tgctcctggccctgctgaccctaggcctggctC ATATGAATGGCACAGAAGGCCC	Amplification of fragment containing the OR to be cloned into the reporter plasmid
OL008R	AGTCGGCCCTGCTGAGGAGTCT TTCCACCTGCAGGTCTTATCATG TCTGCTCGAA	Amplification of fragment containing the OR to be cloned into the reporter plasmid
OL009	CTTCTACGTGCCCTTCTC	Sequencing and linking barcodes/ORs in the reporter vector
OL010	CCTGCAGGTCTTATCATGTC	Sequencing and linking barcodes/ORs in the reporter vector
OL011	TACAGGCGGAATGGACGAG	Sequencing and linking barcodes/ORs in the reporter vector
OL012F	AAGTGAAAACCACCTACAAGG	QPCR of the transposon for copy number analysis
OL012R	CCCTTTAATCAGATGCGTCG	QPCR of the transposon for copy number analysis

OL013	AATGATACGGCGACCACCGAGATCTACAC NNNNNNNN AAGTGCCTTCCTGCCCTTTAA	P5+i5+Read1+primer, for large-scale library amplification
LP001F	TGGGCAGTTCCAGGCTTATAGTC	Genomic Amplification of the H11 locus with the landing pad
LP001R	GGGCGTACTTGGCATATGATACAC	Genomic Amplification of the H11 locus with the landing pad
List of indices used for Pilot-Scale Screen (i7)		
Name	Index	
TBSC01	ATCACG	
TBSC02	CGATGT	
TBSC03	TTAGGC	
TBSC04	TGACCA	
TBSC05	ACAGTG	
TBSC06	GCCAAT	
TBSC07	CAGATC	
TBSC08	ACTTGA	

TBSC09	GATCAG	
TBSC10	TAGCTT	
TBSC11	GGCTAC	
TBSC12	CTTGTA	
TBSC13	AGTCAA	
TBSC14	AGTTCC	
TBSC15	ATGTCA	
TBSC16	CCGTCC	
TBSC17	GTAGAG	
TBSC18	GTCCGC	
TBSC19	GTGAAA	
TBSC20	GTGGCC	
TBSC21	GTTTCG	
TBSC22	CGTACG	
TBSC23	GAGTGG	
TBSC24	GGTAGC	
TBSC25	ACTGAT	

TBSC26	ATGAGC	
TBSC27	ATTCCT	
TBSC28	CAAAAG	
TBSC29	CAACTA	
TBSC30	CACCGG	
TBSC31	CACGAT	
TBSC32	CACTCA	
TBSC33	CAGGCG	
TBSC34	CATGGC	
List of Indices Used for Large-Scale Odorant Screen		
Well	Plate	
Index 1 (i7 side)	Index 2 (i5 side)	
CCTGCGA	CTCTCTAT	
TGCAGAG	TATCCTCT	

ACCTAGG	GTAAGGAG	
TTGATCC	ACTGCATA	
ATCTTGC	AAGGAGTA	
TCTCCAT	CTAAGCCT	
CATCGAG	CGTCTAAT	
TTCGAGC	TCTCTCCG	
AGTTGGT	CTAGTCGA	
GTACCGG	AGCTAGAA	
CGGAGTT	ACTCTAGG	
ACTTCAA	TCTTACGC	
TGATAGT	CTTAATAG	
GATCCAA		
CAGGTCG		
CGCATTA		
GGTACCT		
GGACGCA		
GAGATTC		

GAGCATG		
GTTGCGT		
CCAATGC		
CGAGATC		
CATATTG		
GACGTCA		
TGGCATC		
GTAATTG		
CCTATCT		
CAATCGG		
GCGGCAT		
AGTACTG		
TACTATT		
CCGGATG		
ACCATGA		
CGGTTCT		
TATTCCA		

CCTCCTG		
AGGTATT		
GCATTCG		
TTGCGAA		
TTGAATT		
CTGCGCG		
AGACCTT		
GTCCAGT		
ACCTGCT		
CCGGTAC		
CTTGACC		
CATCATT		
TCTGACT		
TCTAGTT		
GCCATAG		
ACCGTCG		
CTTGGTT		

TACGCCG		
GGACTGC		
GCGCGAG		
GTCGCAG		
CATACGT		
TCAGTAT		
CTAAGTA		
TTAGCTT		
CGCCGTC		
GTCTTCT		
GCCGGAC		
AAGCTGA		
GCGCTCT		
CGTAGGC		
ATGATTA		
GCAGGTT		
AATCGTC		

CGGCCTA		
CTATGCC		
GGTTGAA		
GAGTTAA		
TAGACTA		
TCATGCA		
GCTTATT		
CAAGGCT		
AGGTTGG		
CTTCTGC		
TAATTCT		
GATGCTG		
CCTAGAA		
CTAGAGG		
TATCCGG		
AGGCGGC		
GGTCGTT		

CCGCTGG		
GGAACTA		
ATTGCCA		
ATATACG		
GATTAGC		
AGAAGTC		
ATAGTAC		
GATCTCG		
GGCTGCG		

References

1. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3,** 353–359 (2004).

2. Reddy, A. S. & Zhang, S. Polypharmacology: drug discovery for the future. *Expert Rev. Clin. Pharmacol.* **6,** 41–47 (2013).

3. Fang, J., Liu, C., Wang, Q., Lin, P. & Cheng, F. In silico polypharmacology of natural products. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbx045

4. Anighoro, A., Bajorath, J. & Rastelli, G. Polypharmacology: challenges and opportunities in drug discovery. *J. Med. Chem.* **57,** 7874–7887 (2014).

5. Malnic, B., Hirono, J., Sato, T. & Buck, L. B. Combinatorial receptor codes for odors. *Cell* **96,** 713–723 (1999).

6. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65,** 175–187 (1991).

7. 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).

8. Niimura, Y., Matsui, A. & Touhara, K. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome Res.* **24,** 1485–1496 (2014).

9. Peterlin, Z., Firestein, S. & Rogers, M. E. The state of the art of odorant receptor deorphanization: a report from the orphanage. *J. Gen. Physiol.* **143,** 527–542 (2014).

10. Lu, M., Echeverri, F. & Moyer, B. D. Endoplasmic reticulum retention, degradation, and aggregation of olfactory G-protein coupled receptors. *Traffic* **4,** 416–433 (2003).

11. Bushdid, C., de March, C. A., Matsunami, H. & Golebiowski, J. Numerical Models and In Vitro Assays to Study Odorant Receptors. in *Olfactory Receptors: Methods and Protocols* (eds. Simoes de Souza, F. M. & Antunes, G.) 77–93 (Springer New York, 2018).

12. Saito, H., Chi, Q., Zhuang, H., Matsunami, H. & Mainland, J. D. Odor coding by a Mammalian receptor repertoire. *Sci. Signal.* **2,** ra9 (2009).

13. Mainland, J. D. *et al.* The missense of smell: functional variability in the human odorant receptor repertoire. *Nat. Neurosci.* **17,** 114–120 (2014).

14. Botvinik, A. & Rossner, M. J. Linking cellular signalling to gene expression using EXT-encoded reporter libraries. *Methods Mol. Biol.* **786,** 151–166 (2012).

15. Galinski, S., Wichert, S. P., Rossner, M. J. & Wehr, M. C. Multiplexed profiling of GPCR activities by combining split TEV assays and EXT-based barcoded readouts. *Sci. Rep.* **8,** 8137 (2018).

16. Zhuang, H. & Matsunami, H. Synergism of accessory factors in functional expression of mammalian odorant receptors. *J. Biol. Chem.* **282,** 15284–15293 (2007).

17. Cook, B. L., Ernberg, K. E., Chung, H. & Zhang, S. Study of a synthetic human olfactory receptor 17-4: expression and purification from an inducible mammalian cell line. *PLoS One* **3,** e2920 (2008).

18. Belloir, C., Miller-Leseigneur, M.-L., Neiers, F., Briand, L. & Le Bon, A.-M. Biophysical and functional characterization of the human olfactory receptor OR1A1 expressed in a mammalian inducible cell line. *Protein Expr. Purif.* **129,** 31–43 (2017).

19. Saito, H., Kubota, M., Roberts, R. W., Chi, Q. & Matsunami, H. RTP family members induce functional expression of mammalian odorant receptors. *Cell* **119,** 679–691 (2004).

20. Von Dannecker, L. E. C., Mercadante, A. F. & Malnic, B. Ric-8B, an olfactory putative GTP exchange factor, amplifies signal transduction through the olfactory-specific G-protein Galphaolf. *J. Neurosci.* **25,** 3793–3800 (2005).

21. Shepard, B. D., Natarajan, N., Protzko, R. J., Acres, O. W. & Pluznick, J. L. A cleavable N-terminal signal peptide promotes widespread olfactory receptor surface expression in HEK293T cells. *PLoS One* **8,** e68758 (2013).

22. Krautwurst, D., Yau, K. W. & Reed, R. R. Identification of ligands for olfactory receptors by functional expression of a receptor library. *Cell* **95,** 917–926 (1998).

23. Li, X. *et al.* piggyBac transposase tools for genome engineering. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E2279–87 (2013).

24. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40,** 4288–4297 (2012).

25. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57,** 289–300 (1995).

26. Zhuang, H. & Matsunami, H. Evaluating cell-surface expression and measuring activation of mammalian odorant receptors in heterologous cells. *Nat. Protoc.* **3,** 1402–1413 (2008).

27. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **4,** 268–276 (2018).

28. Antebi, Y. E. *et al.* Combinatorial Signal Perception in the BMP Pathway. *Cell* **170,** 1184–1196.e24 (2017).

29. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14,** 111–129 (2015).

30. Colwell, L. J. Statistical and machine learning approaches to predicting protein-ligand interactions. *Curr. Opin. Struct. Biol.* **49,** 123–128 (2018).

Acknowledgements

We thank the Kosuri Lab for helpful discussions, the UCLA Broad Stem Cell Research Center Sequencing Core, and the Technology Center for Genomics and Bioinformatics for providing next-generation sequencing. We thank Mengjue Ni, Jeong Hoon Ko, and Aaron Cooper for expert technical assistance. We thank Jason Kakoyiannis for providing perfumes as a gift.

Funding: National Science Foundation, Brain Initiative (1556207 to S.K and 1556207/151801 to H.M.), the Jane Coffin Child Foundation (R.J.), Ruth L. Kirschstein National Research Service Award (GM007185 to N.L.), the USPHS National Research Service Award (5T32GM008496 to E.J.), the NIH (DP2GM114829 to S.K. and DC014423 to H.M.) and UCLA. **Author Contributions:** E.J., S.K., R.C., and H.M. conceptualized the experiments. E.J., R.J., D.C., N.L., J.W., M.S., and R.C. performed the experiments. E.J., R.J., N.L., J.B., H.M. and S.K. analyzed the results. C.dM. provided essential reagents, R.J., E.J. and S.K. wrote the manuscript. H.M., J.B., and C. dM. edited the manuscript. **Competing interests:** E.J., R.J., N.L., J.B. and S.K. consult for and hold equity in Octant Inc. to which patents based on this work have been licensed. All other authors declare that they have no competing interests. **Data and materials availability:** All data and analysis scripts are available on https://github.com/KosuriLab/olfaction. Plasmids are available from Addgene and cell lines upon request.

CHAPTER THREE

Deep Mutational Scanning of the Beta-2 Adrenergic Receptor

Title: Deep Mutational Scanning of the Beta-2 Adrenergic Receptor

Abstract

The G protein-coupled receptors (GPCRs) are a uniquely important protein family in human physiology. Their dynamic structure is critical to their myriad functions, but makes biophysical characterization challenging. We developed a platform to characterize large libraries of GPCRs in human cell lines, and use it to functionally assess all possible single amino acid substitutions to the beta-2-adrenergic receptor across several agonist concentrations. Cumulatively, we find that residues with similar mutational profiles reflect their structural and functional organization, and we identify both known and novel residues critical for function. In addition, we describe a previously uncharacterized, conserved extracellular “structural latch” maintained in both the inactive and active state of the receptor. Our approach enables mutational scanning for most GPCRs and other human proteins where function can be linked to a genetic reporter.

Introduction

G protein-coupled receptors (GPCRs) are central mediators of mammalian cells’ ability to sense and respond to their environment. In humans, the ~800 GPCRs respond to a wide range of chemical stimuli such as hormones, odors, natural products, and drugs by modulating a set of prototypical pathways that affect cell physiology. Their central role in altering relevant cell states makes them ideal targets for therapeutic intervention, with ~34% of all U.S. Food and Drug Administration (FDA)-approved drugs targeting the GPCR superfamily[1].

Understanding GPCR signal transduction is difficult for several reasons. First, GPCRs exist in a large conformational landscape, making traditional biophysical characterizations difficult. Consequently, most GPCR structures are truncated, non-native, or artificially stabilized. Even when structures exist, most are of inactive states, and only ~18 receptors have active state

structures available[2]. Second, GPCR dynamics are critical for their function. Static structures from both X-ray crystallography and Cryo Electron Microscopy do not directly probe receptor dynamics[3]. Towards this end, tools such as spectroscopy and computational simulation have aided our interpretation[4].

Alternatively, mutagenesis has long been a foundation of protein biochemistry and, when coupled with a phenotypic screen, provides a robust approach to directly investigate GPCR signaling and function[5–7]. Historically, technical constraints restricted the number of mutations that could be generated and characterized. Recent advancements in DNA synthesis, genome editing, and next-generation sequencing (NGS) have enabled Deep Mutational Scanning (DMS), a method to functionally assay all possible missense variants of a given protein[8,9].

However, deep mutational scans often suffer from not being generalizable across different protein targets or screen a phenotype that is not informative of protein function. GPCRs are particularly susceptible to this problem because they bind a variety of signaling effectors that activate distinct pathways[10–12]. Here we report a novel platform to globally dissect the functional consequences of missense variation in GPCRs expressed in human cell lines. By constructing a method to assay genetic reporters in multiplex, we are able to link GPCR activation to cAMP production, a direct output of G protein signaling. Furthermore, genetic reporters are modular and can be easily exchanged to suit DMS targets of disparate function[13].

Results

To comprehensively probe the structure-function relationship of the β_2-adrenergic receptor (β_2AR), we developed technology to generate and simultaneously profile the receptor's 7,828 possible missense variants for differences in functional activity in HEK293T cells (Fig. 1a, b, Supplementary Fig. 1, 2). The activity of each variant is linked to expression of a cAMP

responsive genetic reporter enabling an accurate depiction of receptor signaling capability. Variant identity is encoded in a short barcode sequence appended to the 3' UTR of the reporter gene. Using RNA-seq we read the activity of the entire variant library in multiplex. We screened the mutant library under four conditions of the β_2AR agonist isoproterenol: vehicle control, an empirically determined EC_{50}, EC_{100}, and beyond saturation of the WT receptor, and report measurements for 99.6% of possible missense variants.

To validate our assay, we recorded the activity of 6 mutants that are stably and individually expressed at single copy, the same configuration as the multiplexed assay, with a luciferase reporter gene[14–16]. These measurements largely agree with the results of our multiplexed assay (Supplementary Fig. 3). Our variant generation approach, microarray-based oligo synthesis, often produces single base deletions that introduced a plethora of frameshift mutations into our library[17]. As expected, frameshift mutations have consistently lower activity than missense mutations. Furthermore, frameshifts occurring in the C-terminus have a significantly diminished reduction of activity (Supplementary Fig. 3).

The heatmap representation of mutant activity reveals the helices are more sensitive to substitution than the termini or loops, and this effect becomes more pronounced at higher agonist concentration (Fig. 1c). In general, the transmembrane domain is especially sensitive to proline substitution (Supplementary Fig. 3). In lieu of large-scale functional data, two indications of the effect a potential mutation will have on protein function are sequence conservation and co-variation. While conservation is highly correlated (ρ = -0.747) with mutational tolerance, the aggregate fitness for all substitutions at a given position, it does not apply to specific substitutions([18–20]). EVmutation, a predictor of mutational effects from sequence covariation, correlates well (ρ = 0.521) with our variant-level data (Fig. 1c)[18–20]. Of note, correlation between our data and both predictors increases with agonist concentration up to EC_{100}, suggesting our phenotypic screen is evolutionarily relevant (Supplementary Fig. 4).

Our data spans thousands of mutations of varying severity across multiple agonist conditions. We hypothesized unsupervised learning methods could reveal hidden regularities within groups' of residues response to mutation. We applied Uniform Manifold Approximation and Projection (UMAP) to learn multiple different lower-dimensional representations of our data and clustered the output with HDBSCAN[21,22] (Supplementary Fig. 5). We found residues consistently separated into 6 clusters that exhibit distinct responses to mutation (Fig. 2a, 2b). Clusters 1 and 2 are globally intolerant to all substitutions, whereas Cluster 3 is affected by proline and hydrophilic substitutions. Cluster 4 is particularly inhibited by negatively charged substitutions, while Cluster 5 is uniquely intolerant to proline, and Cluster 6 is unaffected by substitution.

Mapping these clusters onto a 2D snake plot representation shows Clusters 1-5 primarily comprise the transmembrane helices, while Cluster 6 mainly resides in ICL3 and the termini (Fig. 2c) [2]. These flexible regions are often truncated before crystal structure determination to minimize conformational variability[23]. Surprisingly, a number of residues from Cluster 5 also map there. Given that residues in Cluster 5 are uniquely intolerant to proline substitutions, we hypothesize these regions may become structured in one or more receptor conformations.

Next, we projected the clusters onto the hydroxybenzyl isoproterenol bound structure (Supplementary Fig. 6; PDB: 4LDL). The globally intolerant Clusters 1 and 2 segregate to the core of the protein, while Cluster 3, intolerant to polar residue substitution, is enriched in the lipid-facing portion. This suggested that differential response to hydrophobic and charged substitutions could correlate with side chain orientation within the transmembrane domain. Indeed, residues that are uniquely charge sensitive are significantly more lipid-facing than those that are sensitive to both hydrophobic and charged mutations (Fig 3a; Supplementary Fig. 6) [24]. Taken together, DMS and unsupervised learning methods provide a way to determine patterns of mutational constraint between cohorts of residues. From this, we can learn structural

features, such as side chain orientation and secondary structure, even without a crystal structure.

Decades of research have revealed many GPCRs couple ligand binding to G protein activation through a series of conserved motifs[25]. The globally intolerant UMAP clusters (1 and 2) highlight many residues from these motifs and suggest novel residues for further investigation (Fig. 2c, Fig. 3b). We can further resolve the significance of individual residues within these motifs by ranking the mutational tolerance of every position at EC_{100} (Supplementary Fig. 7). In fact, 8 of the 10 most intolerant positions belong to known structural motifs. However, the most mutationally intolerant residue is the uncharacterized G315. In the active state, G315's alpha carbon points directly at W286 of the CWxP motif, and any substitution at G315 will likely clash with W286 (Fig. 3c). Additionally, W286 is the second most intolerant position, reinforcing its essentiality for receptor function. Recent simulations suggest networks of water-mediated hydrogen bonds play a critical role in GPCR function[26]. Y326 of the NPxxY motif, the 5th most intolerant position, switches between two such networks during the active state transition. In the inactive state, Y326 networks with N51 and D79, two of the top 20 most intolerant positions (Fig. 3d).

Next, we wondered if residues in the orthosteric site that directly contact isoproterenol would respond differently to mutation than residues that contact other agonists. Using the crystal structure of the β_2AR bound to hydroxybenzyl isoproterenol, we find that positions responsible for binding the derivatized hydroxybenzyl tail are significantly less sensitive to mutation than residues that contact the catecholamine head common to both molecules at EC_{100} ($p = 0.0162$; Fig. 3e, Supplementary Fig. 7). Given this discrimination, we believe DMS can be a powerful tool for mapping ligand-receptor contacts.

The numerous β_2AR crystal structures in various complexes and conformational states enable us to evaluate the functional consequences of predicted intermolecular interactions. For example, cholesterol is an important modulator of β_2AR and the timolol-bound inactive state structure elucidated the coarse location of a cholesterol binding site (PDB: 3D4S; [27]). As previously predicted, W158$^{4.50x50}$ is the most mutationally intolerant of the residues (Supplementary Fig. 7). Furthermore, the relative contribution of most individual residues for stabilizing the G_s-β_2AR interface is unknown[7,28–34]. Interestingly, most residues are tolerant to substitution, but three of the most intolerant positions are I135, V222, and Q229 respectively (Fig. 3f). Q229 appears to coordinate polar interactions between D381 and R385 of the α5 helix of G_s, whereas V222 and I135 form a hydrophobic pocket on the receptor surface.

The mutationally intolerant UMAP clusters also highlight residues from tolerant regions of the structure. For instance, the uncharacterized W99$^{23.50x50}$, of ECL1, is proximal to the disulfide bond C106-C191, which is important for stabilization of the high-affinity receptor state (Fig. 4a)[35,36]. Aromatic residues are known to facilitate disulfide bond formation, but our data suggest only tryptophan is tolerated[37]. We hypothesize W99$^{23.50x50}$'s indole group hydrogen bonds with the backbone carbonyl of neighboring G102, positioning W99$^{23.50x50}$ towards the disulfide bond. Other aromatic residues are unable to hydrogen bond and are less likely to be positioned properly.

This observation led us to inquire whether the structural latching between W99$^{23.50x50}$ and the disulfide bond is specific to human β2AR or generic to all class A GPCRs. Comparing over 25 high-resolution structures of class A GPCRs from five functionally different sub-families and six different species revealed that position the trp and disulfide bond consistently contact each other (Fig. 4a). Three exceptions to this trend are the human S1P1 sphingosine receptor, A2A adenosine receptor and bovine rhodopsin. Expectedly, the S1P1 sphingosine receptor lacks the conserved disulfide bond and has a relatively long ECL1, uncharacteristic properties of class A

GPCRs. For both the human A2A adenosine receptor and the bovine rhodopsin, the trp is substituted by another aromatic residue, phenylalanine. In addition to the trp-disulfide bond non-covalent interaction, we also observe the backbone geometry of ECL1 is highly similar among the class A GPCRs (data not shown). Based on the evolutionary coupling analysis and structural comparison of class A GPCRs, we find the trp in ECL1 together with the disulfide bond connecting ECL2 and the extracellular end of TM3 form a conserved “extracellular structural latch” that is maintained consistently in different GPCRs spanning diverse molecular functions and phylogenetic origins.

Next, we were interested in understanding the dynamics of the structural latch. While the overall RMSD between the inactive and active states for human β2AR (PDB: 2RH1 vs. 3P0G) and M2 muscarinic receptor (PDB: 3UON vs. 4MQS) are 1.32 and 1.78 respectively, the RMSD of the latch is nearly identical in both receptors (Fig. 4b). Additionally, we examined 100 frames each sampled from deactivating simulations of the human β2AR to investigate whether the latch is maintained during the conformational transition between the two states. The residues forming the structural latch in human β2AR (W99$^{23.50x50}$, C106, and C191) are locked in their chi1 angles, as seen in the inactive (2RH1) or active (3P0G) state structures, with standard deviations around 7-8 degrees each (data not shown). The low variability of the side chain geometry of these residues during the conformational transition between the active and inactive states asserts that the extracellular structural latch is rigid and indeed conformation independent. Furthermore, the majority of 15 other residues that undergo very low chi1 rotamer changes are in proximity to the extracellular structural latch (data not shown). This suggests that the extracellular structural latch is a part of a larger rigid plug present at the interface of the transmembrane and extracellular region, which could be important for the structural integrity and thereby function of the receptor.

Conformation dictates GPCR function, therefore identifying individual mutations that stabilize particular states provides insight into the biochemistry of receptor activation. We filtered for mutations that lead to greater than WT activity without agonist stimulation to search for variants with increased basal activation rates or expression levels. Mapping these mutations onto the 2D snake plot reveals they are not uniformly distributed throughout the protein; rather they are enriched in the termini, TM1, TM5, ICL3, and Helix 8 (Supplementary Fig. 8). Concentration at the N- and C-termini is unsurprising, as these regions have known involvement in surface expression[38]. Similarly, the enrichment of mutants in ICL3 reiterates its role in G protein binding[39–41]. Of note, a group of mutations in TM5 face TM6, which undergoes a large conformational change during receptor activation (Supplementary Fig. 8). The activating mutant E62R of ICL1 is also salient as R63 and L64 are both highly intolerant, suggesting an underappreciated role of ICL1 in receptor activation (Supplementary Fig. 7). Lastly, understanding how human variation affects β_2AR signaling and GPCRs in general is critical. We find approximately 60% of reported variants in the genome Aggregation Database (gnomAD) result in a loss of function of the β_2AR at EC_{100} (Supplementary Fig. 8).

Discussion

Our findings showcase a new generalizable approach for deep mutational scanning of human protein targets with transcriptional reporters. Genetic reporters enable precise measurements of gene-specific phenotypes that can be widely applied across the proteome. We show comprehensive mutagenesis can allude to the structural organization of the protein and the local environment of individual residues. These results suggest deep mutational scanning can work in concert with other techniques (e.g. X-ray crystallography and Cryo-EM) to augment our understanding of GPCR structure. Moreover, we identify key residues for β2AR function including uncharacterized positions that inform about receptor stability and activation.

Looking forward, our method is well poised to investigate many outstanding questions in GPCR biology. First, individual GPCRs are known to signal through multiple pathways: both through interactions with multiple G protein alpha subunits as well as beta-arrestin signaling[42]. Through systematic mutational interrogation across these pathways' genetic reporters, we can understand the different mechanisms that underpin their signal transduction and the molecular basis for biased signaling[43]. Second, GPCRs are often targeted by synthetic molecules with either unknown or predicted binding sites. We find ligands imprint a mutational signature on their receptor contacts and each ligand's mutational profile can reveal their molecular contacts either in the case of orthosteric ligands or allosteric modulators. Lastly, the identification of mutations that can stabilize specific conformations or increase receptor expression can aid in GPCR structural determination[44,45].

Figures

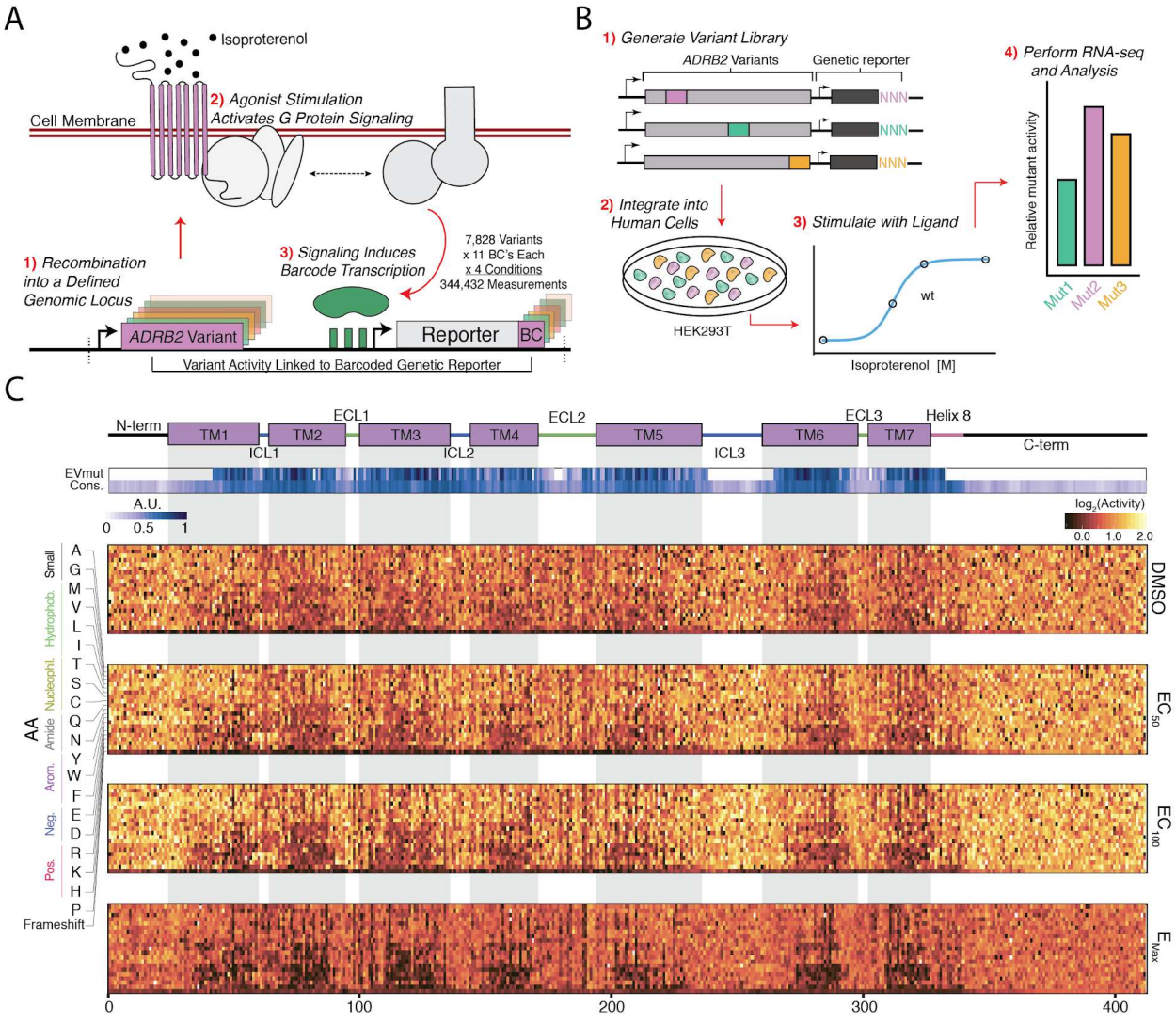


Fig. 3.1. Platform for Deep Mutational Scanning of GPCRs and Variant-Activity Landscape. A. Graphical Display of Multiplexed GPCR Activity Assay. *ADRB2* variants with their barcoded genetic reporter are integrated into a defined genomic locus such that one variant is integrated per cell. Upon isoproterenol agonization, G protein signaling induces transcription of the cAMP-responsive genetic reporter and the barcode. The barcode sequence in the 3' UTR of the reporter encodes the identity of the receptor within the same cell. **B.** Overview of workflow for Multiplexed GPCR Activity Assay. The variant library is generated, barcoded, and cloned into a vector with a genetic reporter. The library is then integrated into HEK293T cells and agonized with various concentrations of isoproterenol. After stimulation, mutant activity is determined by measuring the relative abundance of each variant's barcoded cAMP-responsive genetic reporter transcripts with RNA-seq. **C.** Top: Secondary structure diagram represents the N and C termini in black, the transmembrane domains as blocks, and the intra- and extracellular domains in blue and green respectively. The EVmutation track displays average effect of every mutation as predicted by EVmutation. The Conservation track displays the sequence conservation of each residue. The shaded guides represent positions of the protein in the transmembrane domain. Bottom: The heatmap representation of the activity of

every missense mutation and frameshift at each agonist condition. Cells are colored by the relative activity to the mean frameshift mutation.

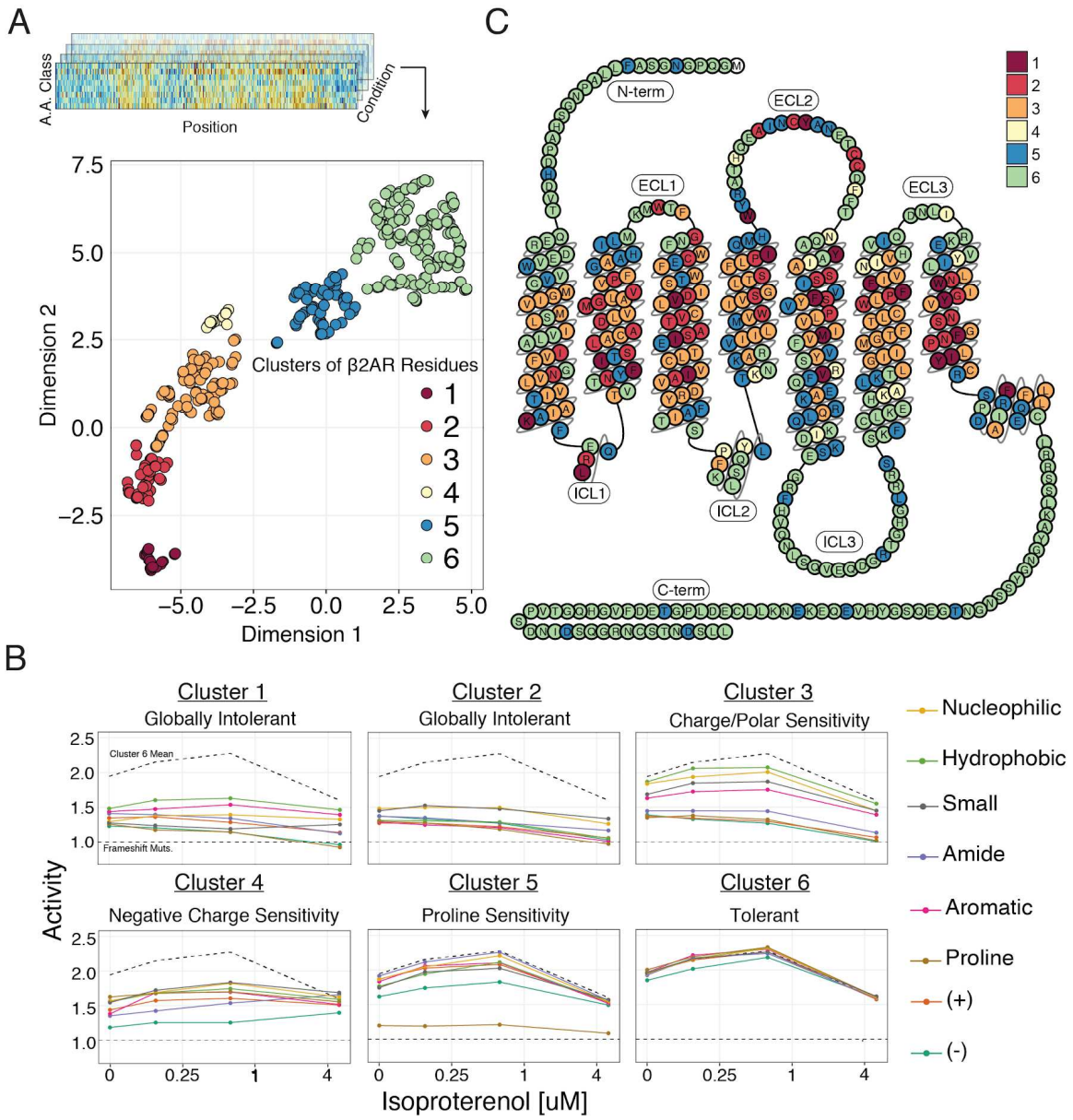


Fig. 3.2. Unsupervised Learning Segregates Residues into Clusters of Distinct Responses to Mutation. A. We averaged amino acid substitutions into classes based on their physicochemical properties. We then used Uniform Manifold Approximation and Projection (UMAP) to learn a 2D representation of every residue's response to these classes of substitutions across all agonist conditions. Each residue is assigned into one of six clusters using HDBSCAN (see Supplementary Fig. 5). **B.** The class averages of each of these cluster reveals their distinct responses to mutation. The upper dashed line represents the mean of the Cluster 6 and the lower dashed line represents the mean activity of frameshifted mutants. **C.** A 2D snake plot representation of β_2AR secondary structure with each residue colored by cluster.

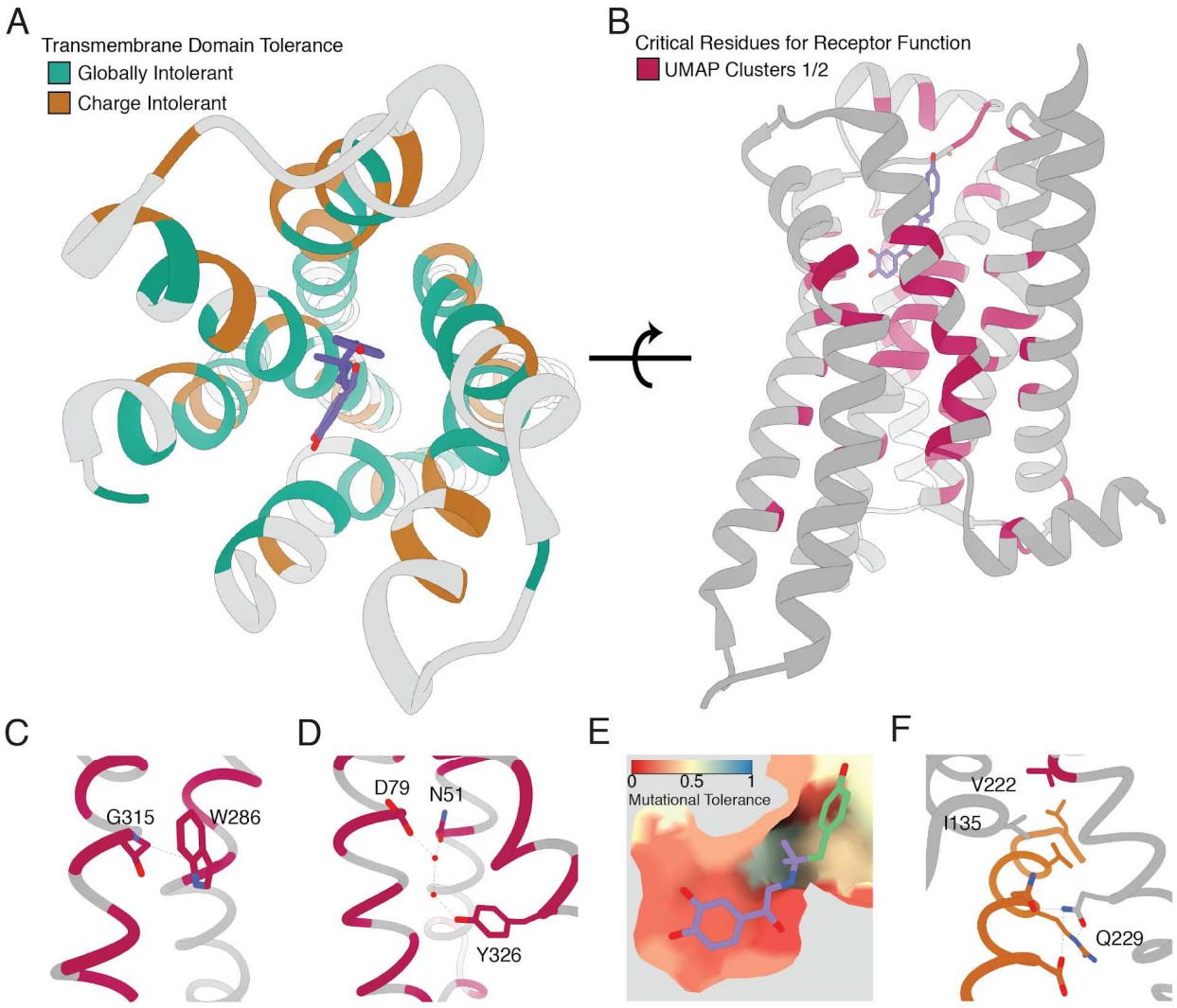


Fig. 3.3. Cluster Identity Elucidates Broad Structural Features and Critical Residues of the β_2AR. A. Residues within the transmembrane domain colored by their tolerance to particular substitutions. Teal residues are intolerant to both hydrophobic and charged amino acids (globally intolerant), and brown residues are tolerant to hydrophobic amino acids but intolerant to charged amino acids. These charge sensitive positions tend to point into the membrane, while the globally intolerant positions face into the core of the protein. **B.** The crystal structure of the hydroxybenzyl isoproterenol-activated state of the β_2AR (PDB: 4LDL) with residues from the mutationally intolerant Clusters 1 and 2 highlighted in magenta. **C-F.** Selected vignettes of residues from the mutationally intolerant UMAP clusters. **C.** W286 of the CWxP motif and the neighboring G315 are positioned in close proximity. Substitutions at G315 are likely to cause a steric clash with W286 (PDB: 4LDL). **D.** An inactive state water-mediated hydrogen bond network (red) associates N51 and Y326 (PDB: 2RH1). Disruption of this network may destabilize the receptor **E.** The ligand-bound orthosteric site surface colored by mutational tolerance the unique sensitivity of the receptor-ligand contacts and displays the assay's discriminatory power between agonists (PDB: 4LDL). **F.** Mutationally intolerant β_2AR residues at the G protein interface from the β_2AR-Gs complex crystal structure (PDB: 3SN6), V222, I135, and Q229.

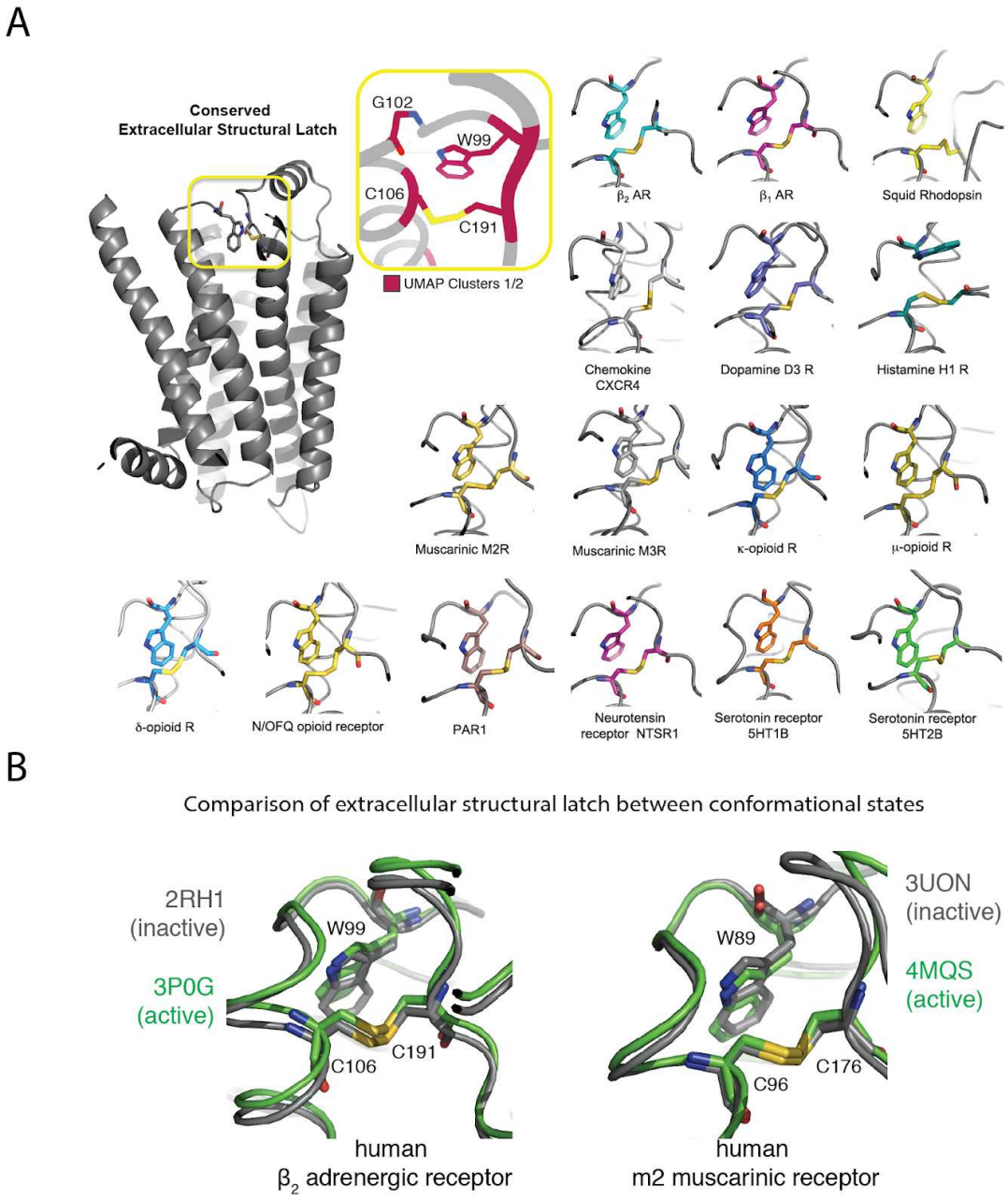


Fig. 3.4. A Conserved Extracellular Tryptophan-Disulphide 'Structural Latch' in Class A GPCRs is Rigid and Conformation-Independent. A. $W99^{23.50x50}$ is mutationally intolerant and appears to be contacting the C106-C191 disulfide bond of the ECL1. A structural comparison of Class A GPCR structures reveals the Trp-disulfide bond contact is conserved in 22 of the 25 receptors. **B.** The Trp-disulfide bond contact is maintained in both the inactive and active state structures for the β_2AR and M2 muscarinic receptor.

Supplementary Information

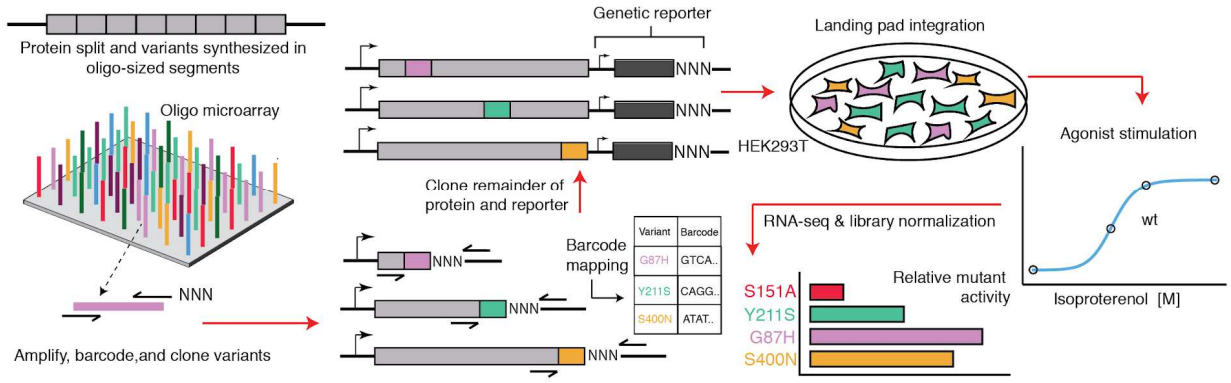


Fig. 3.5. Schematic of Generation, Functional Assessment, and Analysis of All 7,828 Missense Variants of the β2AR. We synthesized missense variants on an oligonucleotide microarray, amplified the oligos and appended random DNA barcode sequences, and cloned the variants into WT background vectors. We then mapped barcode-variant pairs with next-generation sequencing and cloned the remainder of the WT receptor and genetic reporter into the construct. Next, we integrated the variant library *en masse* into a serine recombinase landing pad engineered at the H11 locus of Δ*ADRB2* HEK293T cells. The recombination strategy ensures a single receptor variant/genetic reporter is integrated per cell to avoid crosstalk between genetic reporters. After selection, we stimulated the library with various concentrations of ADRB2 agonist, isoproterenol. Finally, we determined mutant activity by measuring the relative abundance of each variant's barcoded cAMP-responsive genetic reporter transcripts with RNA-seq.

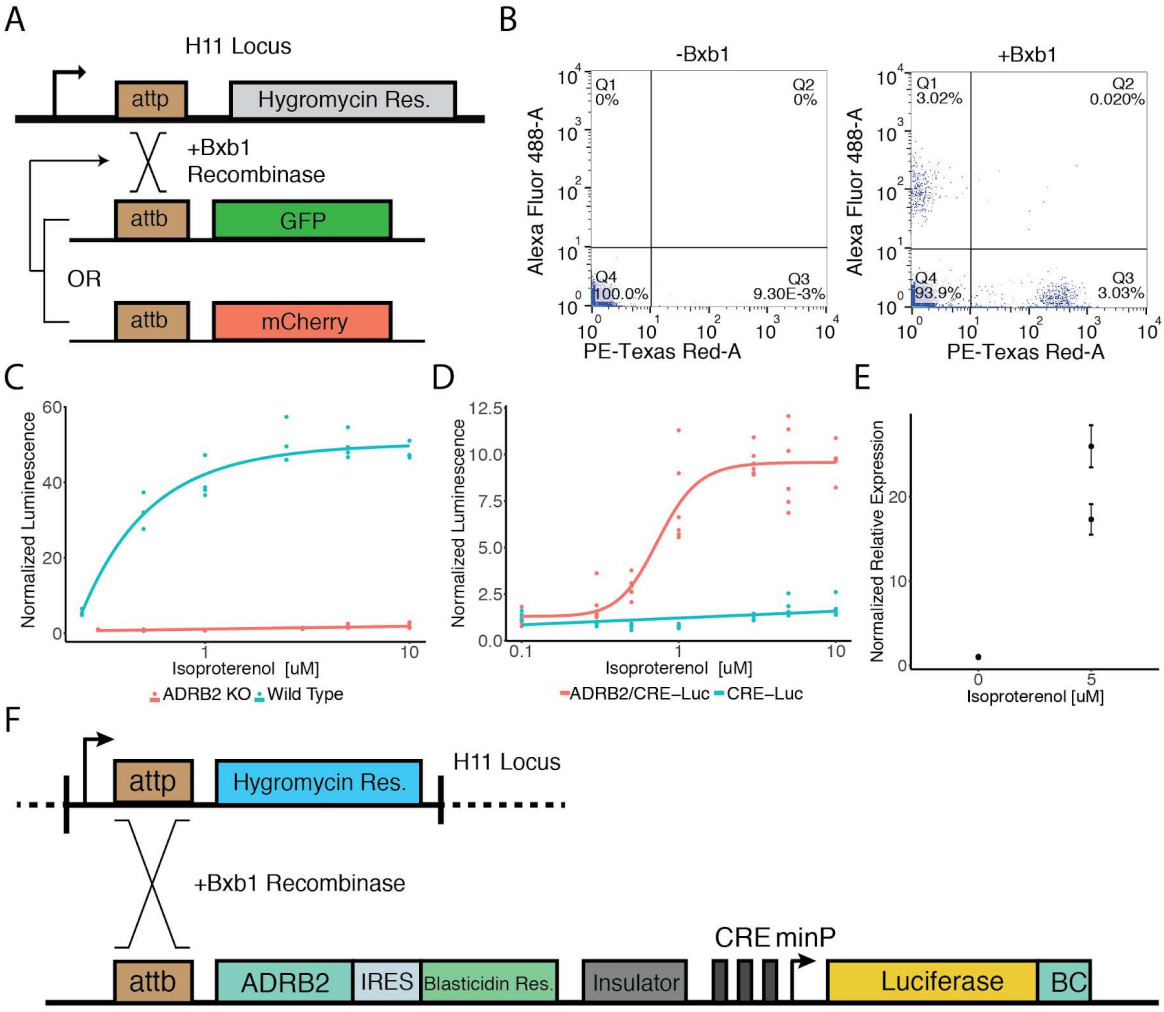


Fig. 3.6. Engineering HEK293T Cells for Clonal and Functional Integration of an ADRB2 Genetic Reporter. A. Schematic of functional assay to ensure the landing pad is present at single copy in the genome and can recombine a single donor plasmid per cell. Single copy integration is essential to ensure receptor's of variable functionality do not activate barcoded reporters mapped to other variants. Upon co-transfection of the promoterless GFP and mCherry plasmids with bxb1 recombinase sites, a cell line with a single landing pad will exclusively integrate one cassette. Therefore, cells will be either GFP^+ or $mCherry^+$ but never both. **B.** Flow Cytometry plots detailing the percentage of GFP^+ and $mCherry^+$ cells when transfected with an equimolar ratio of promoterless GFP and mCherry expression cassettes with or without Bxb1 recombinase expression. **C.** Activation of a cAMP-responsive genetic reporter via a luciferase assay integrated in the landing pad when stimulated with isoproterenol in a WT or ΔADRB2 background. Activation of the reporter in the WT background emphasizes the importance for generation of the Δ*ADRB2* for the purpose of multiplexed experiment. **D.** Activation of a genetic reporter with or without exogenous *ADRB2* expression via a luciferase assay integrated in the landing pad when stimulated with isoproterenol in Δ*ADRB2* cells. **E.** Activation of an equivalent integrated genetic reporter/*ADRB2* cassette via qRT-PCR of the reporter transcript in Δ*ADRB2*

cells. **F.** Schematic detailing the recombination of the reporter/receptor expression plasmid into the landing pad locus.

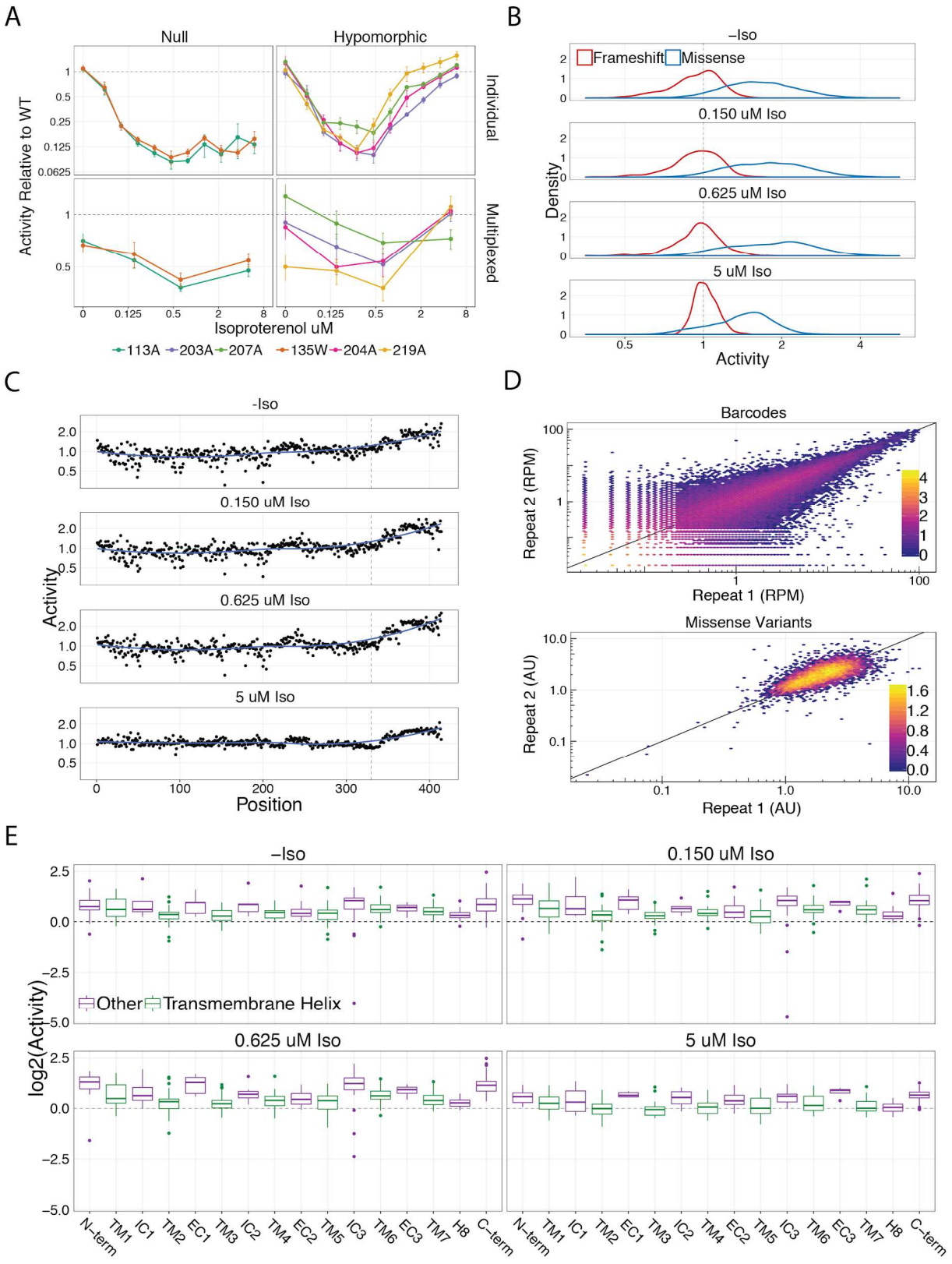

A
Null
Hypomorphic
Individual
Multiplexed
Activity Relative to WT
Isoproterenol uM
113A 203A 207A 135W 204A 219A
B
−Iso
Frameshift Missense
0.150 uM Iso
0.625 uM Iso
5 uM Iso
Density
Activity
C
-Iso
0.150 uM Iso
0.625 uM Iso
5 uM Iso
Activity
Position
D
Barcodes
Repeat 2 (RPM)
Repeat 1 (RPM)
Missense Variants
Repeat 2 (AU)
Repeat 1 (AU)
E
−Iso
0.150 uM Iso
0.625 uM Iso
5 uM Iso
log2(Activity)
Other Transmembrane Helix
N−term TM1 IC1 TM2 EC1 TM3 IC2 TM4 EC2 TM5 IC3 TM6 EC3 TM7 H8 C−term

Fig. 3.7. Individual and Global Multiplexed Assay Validation. A. To validate our genetic reporter, we compared the measured mutant activity screened individually in the landing pad locus via a luciferase assay to via the multiplexed mutational scan. We recapitulated results individually observed in the DMS for both null and hypomorphic mutations. **B.** The distribution of activity for frameshifts are significantly different that the distribution of our designed missense mutations across increasing isoproterenol concentrations ($p << 0.001$). **C.** We also find the relative activity for frameshift mutations mapped to each codon in the ADRB2 sequence is markedly decreased in the C-terminus of the protein (dotted line), and is consistent across agonist concentration. Blue line represents the LOESS fit. **D.** The measurements between barcodes at the RNA-seq level are well correlated ($r = 0.867$, $r = 0.871$, $r = 0.864$, $r = 0.868$) at all agonist concentrations (0, 0.150, 0. 625, and 5 uM Iso). Similarly, the mean forskolin-normalized values for each variant are correlated at every concentration ($r = 0.657$, $r = 0.686$, $r = 0.729$, $r = 0.750$). Bars represent log10 counts per hex-bin. **E.** The activity of proline mutations stratified by the protein domain each residue belongs to reveals a proline sensitivity in the transmembrane domain across all agonist conditions.

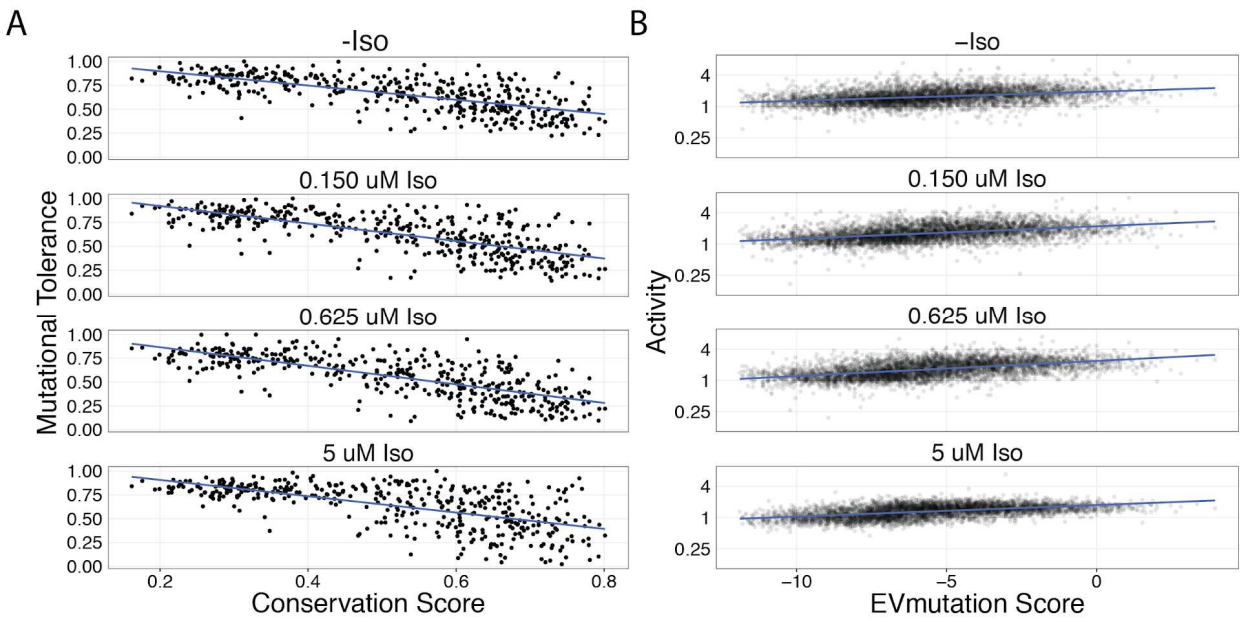


Fig. 3.8. Correlation with Sequence Conservation and Covariation. A. Mutational tolerance is highly correlated with sequence conservation and is maximized at EC_{100} (ρ = -0.689, ρ = -0.719, ρ = -0.747, ρ = -0.634 for -Iso, 0.150 uM Iso, 0.625 uM Iso, and 5 uM Iso, respectively). Here we calculated sequence conservation using the Jensen-Shannon divergence from a multiple alignment of 55 ADRB2 orthologs from the OMA database. The blue line is the least squares fit. **B.** Similarly, our measure of relative activity for individual substitutions is well correlated with the predictions from EVMutation, and is maximized at EC_{100} (ρ = 0.370, ρ = 0.460, ρ = 0.521, ρ = 0.504). The blue line is the least squares fit.

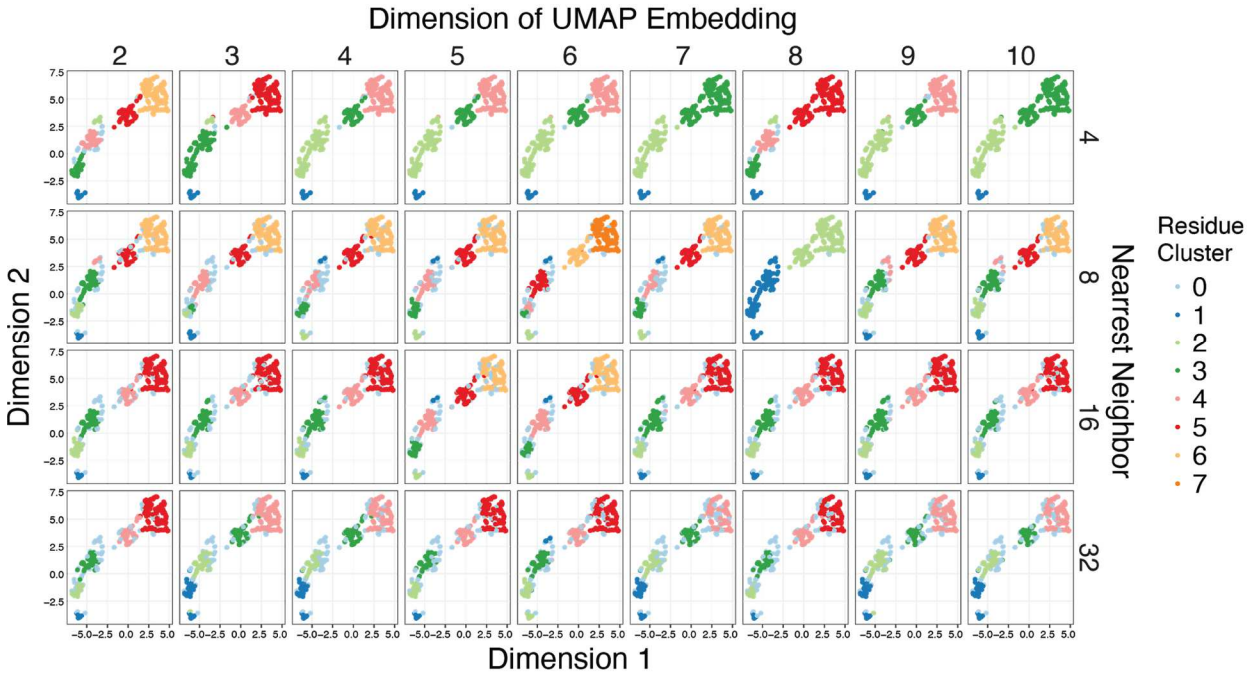


Fig. 3.9. Cluster Assignment is Robust Across Different UMAP Embeddings. Given the high dimensionality of our data, we used UMAP to learn lower-dimension representations of our data before clustering with HDBSCAN (minimum cluster size = 10). To ensure that the clustering results are not biased by a particular UMAP embedding, we ran a hyperparameter search over the dimension and nearest neighbor parameters of UMAP. We then plot the HDBSCAN cluster assignments on a 2D UMAP embedding to ease visualization. Points that HDBSCAN does not assign to a cluster are colored powder blue. We find that groups of residues reliably cluster together regardless of the UMAP embedding, and manually assign all residues to six distinct clusters following the robust HDBSCAN assignment.

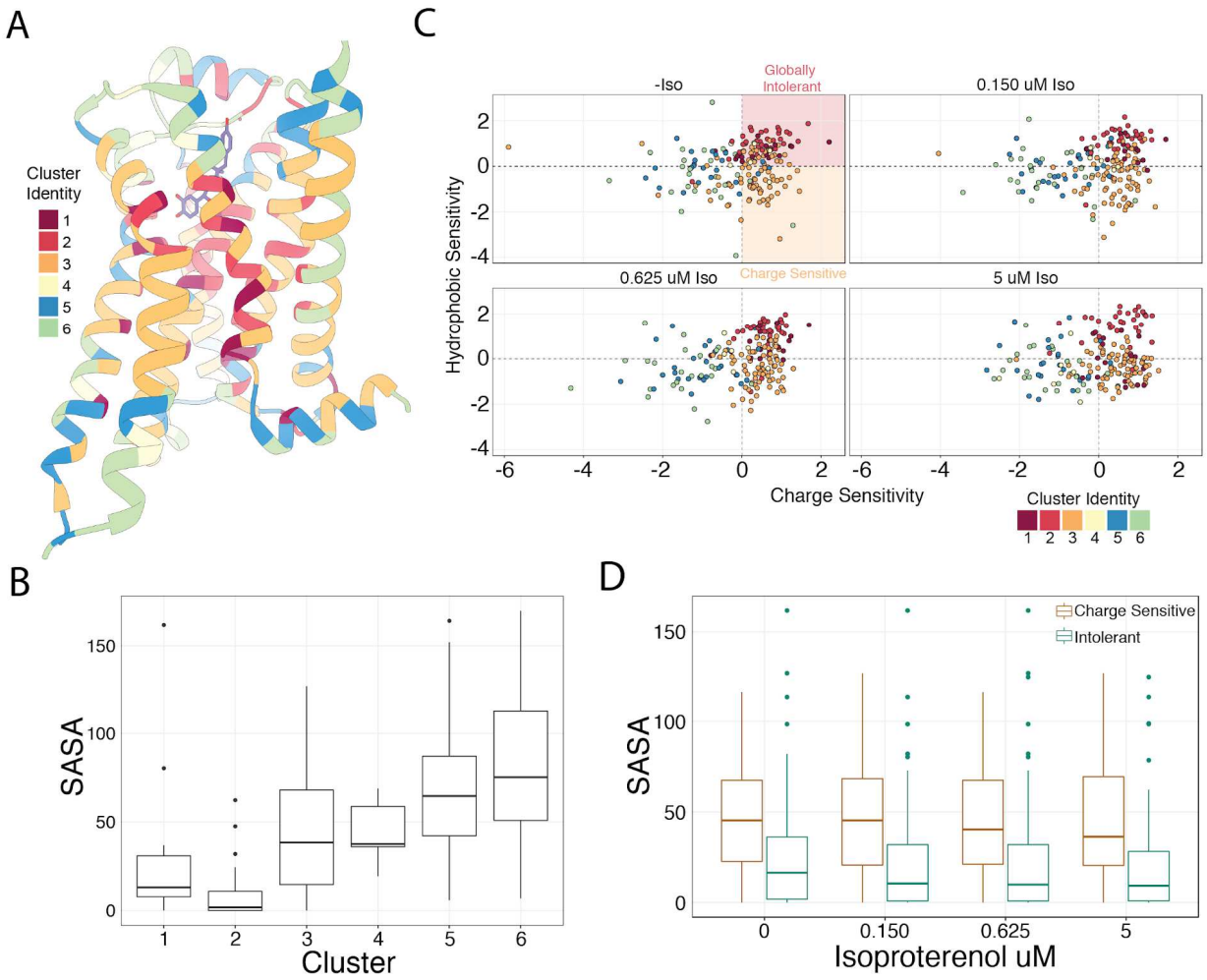


Fig. 3.10. Mutational Profile Suggests Side Chain Orientation and Environment. A. The crystal structure of the hydroxybenzyl isoproterenol-activated state of the β_2AR (PDB: 4LDL) with residues colored by UMAP cluster identity. **B.** Distributions of Solvent Accessible Surface Area (SASA) for each cluster at EC_{100}. **C.** Hydrophobic versus Charge Sensitivity across all drug conditions. Points are colored by cluster identity. We define residues to be globally intolerant to substitution if their Hydrophobic and Charge Sensitivity is greater than 0. Similarly, we define residues to be uniquely charge sensitive if their Hydrophobic Sensitivity is less than 1 and their Charge Sensitivity is greater than 1. **D.** Distributions of SASA for intolerant and charge sensitive clusters are significantly different across all drug concentrations (all $p < 0.0005$).

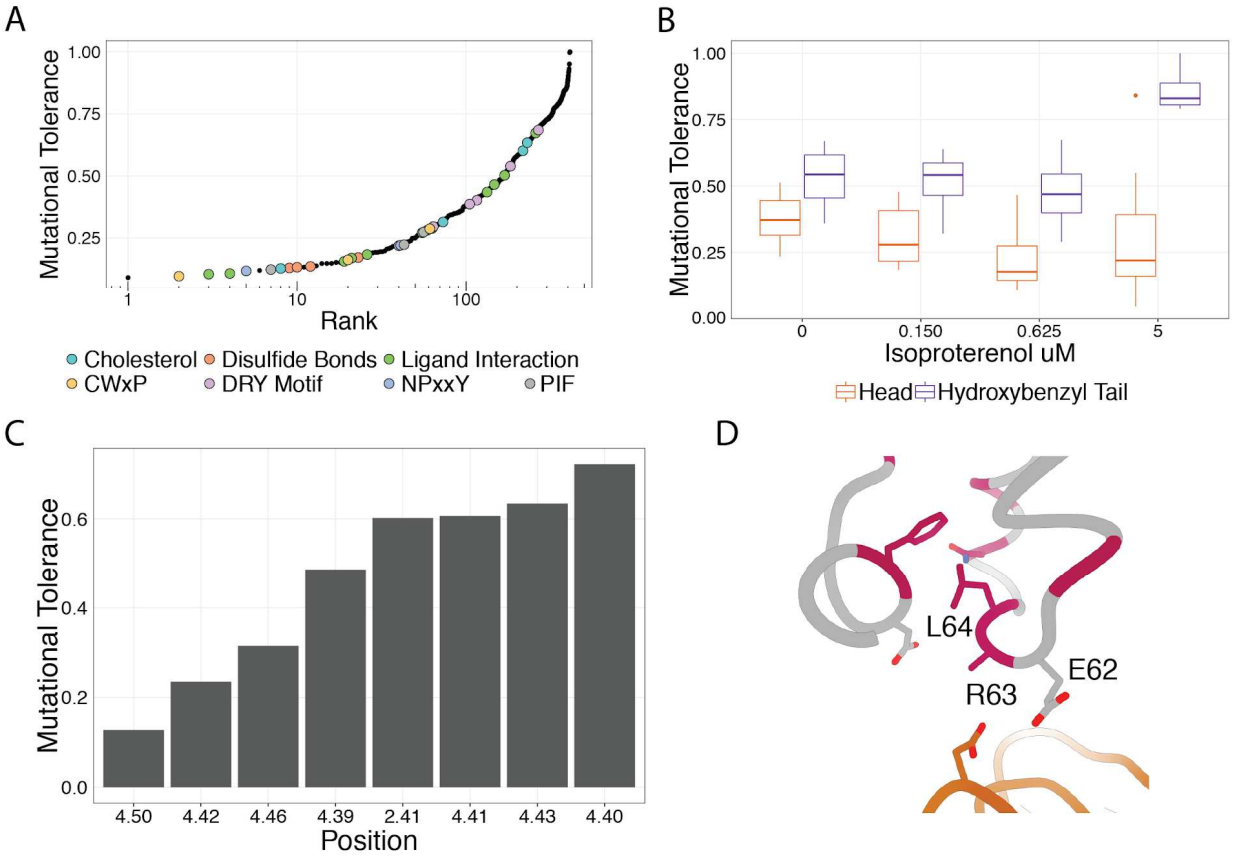


Fig. 3.11. Inspection of Mutationally Intolerant Residues. A. Rank order plot of mutational tolerance at 0.625 uM isoproterenol for all 412 β_2AR residues mutagenized. Residues in known structural motifs (colored points) are significantly more sensitive to mutation than other positions on the protein ($p << 0.001$). **B.** Residues that interact with the head (orange) of hydroxybenzyl isoproterenol have significantly lower mutational tolerance than those that interact with the hydroxybenzyl functional group on the tail (purple). These differences are significantly different at EC_{50} ($p = 0.028$), EC_{100} ($p = 0.016$), and saturating agonist concentration ($p = 0.008$). **C.** Box plot displaying the mutational tolerance of all predicted contacts of the cholesterol binding pocket determined in the timolol-bound structure of the β_2AR inactive state (PDB: 3D4S). The highly conserved $W158^{4.50}$ is the most constrained residue. **D.** ECL1, with residues belonging to cluster 1 and 2 colored magenta, contains a region of sensitivity where R63 and L64 are both intolerant to substitution. However, neighboring E62 displays greater than WT activity at multiple individual mutations (PDB: 3SN6).

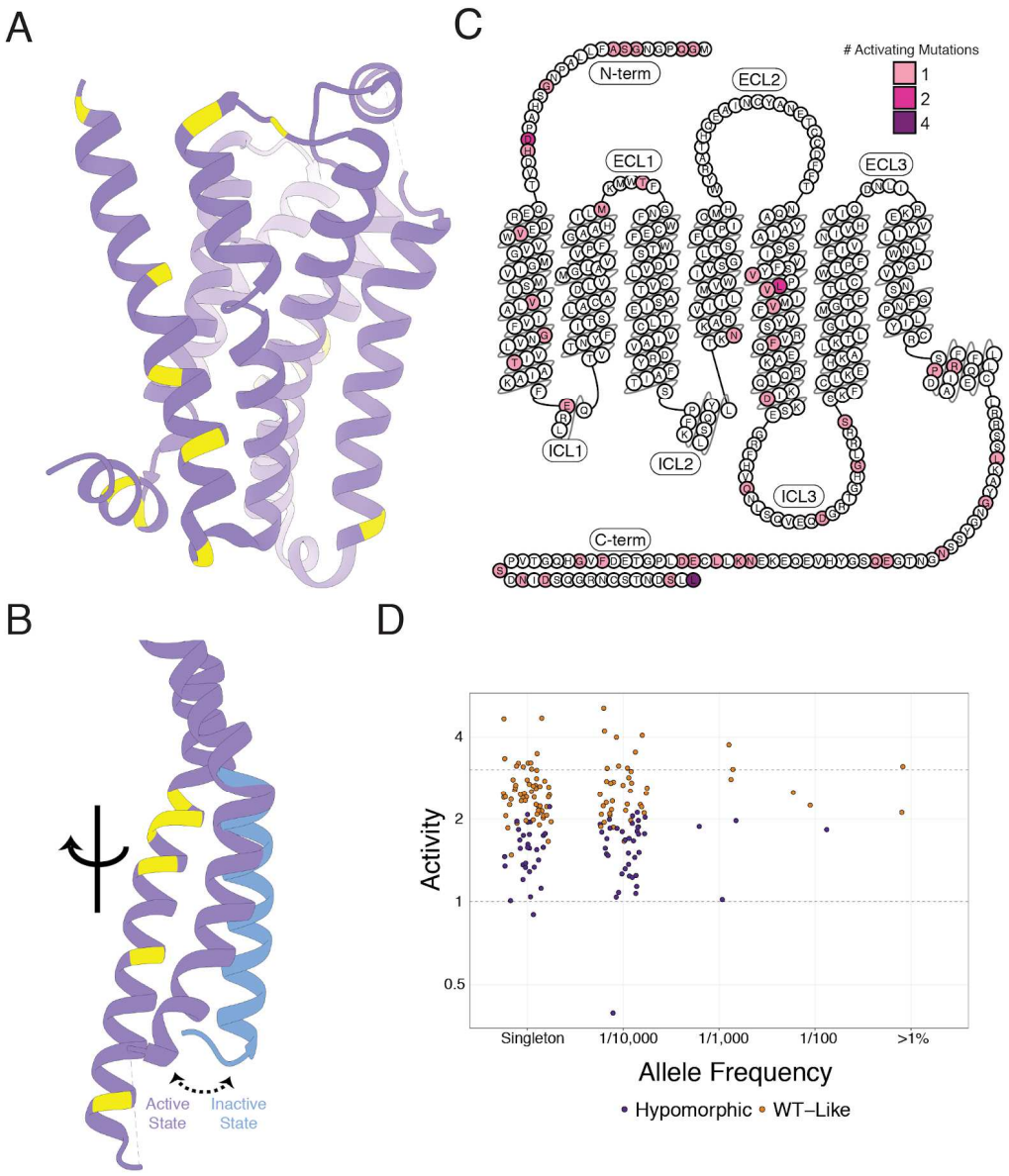


Fig. 3.12. Evaluation of Individual Missense Variants. A. The inactive state β_2AR structure highlighted in regions where residues display greater than WT activity without agonist stimulation for at least one individual mutation (yellow). These mutations localize to the extracellular membrane interface of TM1, TM2, and ECL1. **B.** Other concentrations of these mutants are found in the lower half of TM1, helix 8, and the TM5-TM6 interface. The blue colored structure represents the shift in TM6 upon adoption of the active state. **C.** 2-D snake plot with residues colored by the number of individual mutations that lead to greater than WT activity in the no agonist condition. These residues are enriched in the loops and termini which are truncated in the crystal structures. **D.** Activity of all ADRB2 mutants present in the gnomAD database plotted against to their allele frequency. We classified variants into four categories as follows: null mutants (purple) are variants whose mean plus a standard deviation (SD) are less than 1 (the mean frameshift); activating mutants (orange) are variants whose mean minus a SD are greater than the mean synonymous mutant (dashed line); hypomorphic mutants (periwinkle) are variants whose mean plus a SD are less than the mean synonymous variant; the rest of the variants are considered WT-like (white).

Methods

Experimental Methods

Endogenous ADRB2 Deletion using CRISPR/Cas9

Cas9 and sgRNAs targeting the sole exon of ADRB2 were cloned and transfected into HEK293T cells according to the protocol outlined in Ran et al. (2013) (Supplementary Table X). After transfection, cells were seeded in a 96-well plate at a density of 0.5 cells/well. Wells were examined for single colonies after 3 days and expanded to 24-well plates after 7 days. Clones were screened for ADRB2 deletion by screening them for the inability to endogenously activate a cAMP genetic reporter when stimulated with the ADRB2 ligand isoproterenol. Clones were seeded side by side wild type HEK293T cells at a density of 7,333 cells/well in a pol 96-well plate. 24 hours later, cells were transfected with 10 ng/well of a plasmid encoding luciferase driven by a cyclic AMP response element and 5 ng/well of a plasmid encoding Renilla luciferase with lipofectamine 2000. 24 hours later, media was removed and cells were stimulated with 25 µl of a range of 0 to 10 µM isoproterenol (Sigma-Aldrich) in CD293 (Thermo Fisher Scientific) for 4 hours. After agonist stimulation, the Dual-Glo Luciferase Assay kit was administered according to the manufacturer's instructions. Luminescence was measured using the M1000 plate reader (Tecan). All luminescence values were normalized to Renilla luciferase activity to control for transfection efficiency in a given well. Data were analyzed with Microsoft Excel and R.

Landing Pad Genome Editing

The H11 locus was edited using TALEN plasmids received from Addgene (#51554, #51555). HEK293T cells were seeded at a density of 75k cells in a 24-well plate. 24 hours after seeding

cells were transfected with 50 ng LT plasmid, 50 ng RT plasmid, and 400 ng of the Linearized Landing Pad using Lipofectamine 2000. 2 days after transfection, cells were expanded to a 6-well plate and one day after expansion 500 ug/ml hygromycin B (Thermo FIsher Scientific) was added to the media. Cells were grown under selection for 10 days. After selection, cells were seeded in a 96-well plate at a density of 0.5 cells/well. Wells were examined for single colonies after 3 days and expanded to 24-well plates after 7 days. gDNA was purified using the Quick-gDNA Miniprep kit (Zymo Research) from the colonies and PCR was performed with Hifi Master Mix to ensure the landing pad was present at the correct locus (LP001F and R). The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 35 cycles of 98°C for 20 seconds, 63°C for 15 seconds, and 72°C for 40 seconds, followed by an extension of 72°C for 2 minutes. To ensure a single landing pad was present per cell, HEK293T cell lines with both singly and doubly-integrated landing pads along with untransduced (WT) HEK293T cells were plated at 4 x 10^5 cells per 6-well. All landing pad cells were transfected the next day with 1.094 µg of both an attB-containing eGFP and mCherry donor plasmid and 0.3125 µg of the BxB1 expression vector or a pUC19 control. Two singly-integrated landing pad cell samples were also transfected with 2.1875 µg of either an attB-containing eGFP and mCherry donor plasmid with 0.3125 µg of the BxB1 expression vector. Cells were transfected at a 1:1.5 DNA:Lipofectamine ratio with Lipofectamine 3000. 2 days later cells were passaged at 1:10 and were analyzed using flow cytometry 10 days later after 4 total passages. Samples were flown using the LSRII at the UCLA Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research Flow Cytometry Core. Cytometer settings were adjusted to the settings: FSC – 183 V, SSC – 227 V, PE-Texas Red – 336 V, Alexa Fluor 488 – 275 V.

Individual Donor Bxb1 Recombinase Plasmid Integrations

HEK293T derived cells engineered to contain the Bxb1 Recombinase site at the H11 locus were seeded at a density of 350k cells in a 6-well plate (Corning). 24 hours after seeding cells were

transfected with 2 ug Donor plasmid and 500 ng plasmid encoding the Bxb1 recombinase using Lipofectamine 3000 (Thermo Fisher Scientific). 3 days after transfection cells were expanded to a T-75 flask (Corning) and 8 ug/ml blasticidin (Thermo Fisher Scientific) was added one day after expansion. Cells were kept under selection 7-10 days and passaged twice 1:10 to ensure removal of transient plasmid DNA.

Ligand-Receptor Activation Luciferase Assay for Genomically Integrated Receptor/Reporter Constructs

HEK293T and HEK293T derived cells integrated with the combined receptor/reporter plasmids were plated at a density of 7333 cells/well in 100 uL DMEM in poly-D-lysine coated 96-well plates. 48 hours later, media was removed and cells were stimulated with 25 µl of a range of isoproterenol concentrations in CD293 for 4 hours. After agonist stimulation, the Dual-Glo Luciferase Assay kit was administered according to the manufacturer's instructions. Luminescence was measured using the M1000 plate reader. Data were analyzed with Microsoft Excel and R.

Ligand-Receptor Activation q-RT PCR Assay for Genomically Integrated Receptor/Reporter Constructs

HEK293T and HEK293T derived cells integrated with the combined receptor/reporter plasmids were plated at a density of 200k cells/well in 2 mL DMEM in 6-well plates. 48 hours after seeding, media was removed and cells were induced with various concentrations of either forskolin (Sigma-Aldrich) or isoproterenol diluted in 1 ml of OptiMEM (Thermo Fisher) per plate for 3 hours. After stimulation, media was removed and 600 uL of RLT buffer (Qiagen) was added to each well to lyse cells. Lysate from each sample were homogenized with the QIAshredder kit (Qiagen) and total RNA was prepared from each sample using the RNeasy Mini Kit with the optional on-column DNAse step (Qiagen). 5 ug of total RNA per sample was reverse

transcribed with Superscript III (Thermo-Fisher) using a gene specific primer for the reporter gene and GAPDH (Supplementary Table X) according the manufacturer's protocol. The reaction conditions are as follows: Annealing: [65°C for 5 min, 0°C for 1 min] Extension: [52°C for 60 min, 70°C for 15 min]. 10% of the RT reaction was amplified in triplicate for both genes, the reporter gene and GAPDH (Supplementary Table X), using the SYBR FAST qPCR Master mix (Kapa Biosystems) with a CFX Connect Thermocycler (Biorad). The reaction and cycling conditions are optimized as follows: 95°C for 3 minutes, 40 cycles of 95°C for 3 seconds and 60°C for 20 seconds. Reporter gene expression was normalized to GAPDH expression for each sample. Data were analyzed with Microsoft Excel and R.

Variant Library Generation and Cloning

The ADRB2 missense variant library was created by splitting the protein coding sequence into 8 distinct segments (~52 a.a. each) and synthesizing all single amino acid substitutions for each segment separately as an oligonucleotide library (Agilent). 500 pg of the oligonucleotide library was amplified with biotinylated primers unique for each segment (Supplementary Table X) with the Real-Time Library Amplification Kit (Kapa Biosystems) on a CFX Connect Thermocycler (Biorad). The reaction and cycling conditions are as follows: 98°C for 45 seconds, X cycles of 98°C for 15 seconds, 65°C for 30 seconds, and 72°C for 30 seconds, followed by an extension of 72°C for 1 minute. The number of cycles for the amplification was determined to ensure the amplification was in the exponential phase at least two cycles before the amplification reached saturation. The PCR products were cleaned up with the DNA Clean and Concentrator Kit (Zymo Research) and digested with restriction enzymes BamHI and BspQI, BbsI and BspQI, or BbsI and NheI (New England Biolabs). Digestions were cleaned up with the DNA Clean and Concentrator Kit and digested ends of the amplified library were removed by performing a streptavidin bead cleanup with the Dynabeads M-280 and the DynaMag (Thermo Fisher). Each library segment was to be cloned into a different vector that includes components of the ADRB2

reporter and the wild type sequence portion of ADRB2 upstream of the segment being cloned. These eight different base vectors were digested with restriction enzymes BamHI and BspQI, BbsI and BspQI, or BbsI and NheI. The base vectors were cleaned up with the DNA Clean and Concentrator Kit and the library segments were ligated into the base vectors with T4 DNA ligase (New England Biolabs). The ligations were transformed into 5-alpha Electrocompetent cells (New England Biolabs) directly into liquid culture. Cultures were grown at 30° C overnight to maintain library diversity and dilutions were plated on agarose plates to ensure transformation efficiency was high enough to cover the entire library (>100 transformants per library member). DNA was prepared 16 hours later with the DNA Miniprep Kit (Qiagen). The vectors were digested with BspQI and AgeI or NheI and AgeI (Qiagen). Vectors containing unique sequences corresponding to each library segment that complete the ADRB2 protein sequence and reporter were digested with the same restriction enzymes. These fragments were gel isolated from a 1% agarose gel using the Zymoclean Gel DNA Recovery Kit (Zymo Research). These secondary fragments were cloned into the library vectors with the same protocol as the previous cloning step. DNA was prepared 16 hours later with the Plasmid Plus DNA Maxiprep Kit (Qiagen).

Variant-Barcode Mapping

After the initial cloning of the variant fragments from the oligonucleotide library into each segment's corresponding base vector, the random barcode attached to each variant was associated to its variant with paired-end sequencing. Each plasmid was amplified with 2 rounds of PCRs with distinct primer sets for each segment (Supplementary Table X) with HiFi DNA Master Mix (Kapa Biosystems). For the first round of amplification, the reaction and cycling conditions were optimized as follows: 98°C for 30 seconds, 10 cycles of 98°C for 8 seconds, 64°C for 15 seconds, and 72°C for 10 seconds, followed by an extension of 72°C for 2 minutes. These amplicons were gel isolated from a 1% agarose gel using the Zymoclean Gel DNA Recovery Kit. Prior to the second round of amplification, the number of cycles to amplify was

determined by performing qPCR with the SYBR FAST QPCR Master Mix (Kapa) on the CFX Connect Thermocycler according to the manufacturer's instructions. The Cq determined from the QPCR plus an addition two cycles was used to as the number of cycles to amplify the libraries for the second round of amplification. For the second round of amplification, the reaction and cycling conditions were optimized as follows: 98°C for 30 seconds, X cycles of 98°C for 8 seconds, 62°C for 15 seconds, and 72°C for 10 seconds, followed by an extension of 72°C for 2 minutes. These amplicons were gel isolated from a 1% agarose gel using the Zymoclean Gel DNA Recovery Kit. Kit. Library concentrations were quantified using a TapeStation 2200 (Agilent) and a Qubit (Thermo Fisher). The libraries were sequenced with paired end 150-bp reads on a NextSeq 500 in medium-output mode and paired end 250-bp reads on a MiSeq (Illumina).

Variant Library Bxb1 Recombinase Plasmid Integrations

HEK293T derived cells engineered to contain the Bxb1 Recombinase site at the H11 locus and deletion of endogenous ADRB2 were seeded at a density of 2.13 million cells per dish in 6 100 mm x 20 mm tissue-culture treated culture dishes (Corning). 24 hours after seeding cells were transfected with 11.5 ug Donor plasmid and 2.9 ug plasmid encoding the Bxb1 recombinase using Lipofectamine 3000. 3 days after transfection cells were expanded to T-225 flasks (Corning) and 8 ug/ml blasticidin was added one day after expansion. Cells were kept under selection 7-10 days and passaged 1:10 four times to ensure removal of transient plasmid DNA.

Multiplexed Variant Functional Assay Agonist Stimulation, RNA Preparation and Sequencing

HEK293T derived cells engineered to contain the Bxb1 Recombinase site at the H11 locus, deletion of endogenous ADRB2, and integration of the ADRB2 mutagenic library were seeded at a density of 3,237,868 cells per dish in 150 mm x 25 mm tissue-culture treated culture dishes. 10 dishes were seeded for each biological replicate of each drug condition. 48 hours after

seeding, media was removed and cells were induced with various concentrations of either forskolin or isoproterenol diluted in 9 ml of OptiMEM per plate for 3 hours. After stimulation, media was removed and 3.24 ml of RLT buffer was added to each well to lyse cells. Lysate from dishes belonging to the same replicate were pooled and vortexed thoroughly. 5 ml of lysate from each sample were homogenized with the QIAshredder kit and total RNA was prepared from each sample using the RNeasy Midi Kit with the optional on-column DNAse step (Qiagen) and eluted into 500 ul H_2O. 40 reverse transcriptase reactions were carried out for each sample using the Superscript IV RT kit (Thermo Fisher). For each reaction 11 ul of total RNA were added to 1 ul dNTPs (Qiagen) and 1 ul 2 uM RT primer (Supplementary Table X). The primers were annealed to the template by heating to 65°C for 5 minutes and cooling down to 0°C for 1 minute. After annealing, 4 ul of RT buffer, 1 ul DTT, 1 ul of RNAseOUT, and 1 ul SSIV were added to the mixture and cDNA synthesis was performed. The reaction and cycling conditions are as follows: 52°C for 1 hour, 80°C for 10 minutes. cDNA from the same sample was pooled together and treated with 100 ug/ml RNAse A (Thermo Fisher) and 200 U of RNase H (Enzymatics) at 37°C for 30 minutes. cDNA was concentrated using the Amicon Ultra 0.5 mL 30k Centrifugal Filter (Millipore) according to the manufacturer's instructions with a final spin step time of 15 minutes. To determine the number of cycles necessary for library amplification in preparation for RNA-seq, 1 ul of cDNA from each sample was amplified with SYBR FAST QPCR Master Mix according to the manufacturer's instructions using primers for library amplification and adaptor addition (Supplementary Table X). Each sample was subsequently amplified for 4 cycles more than the Cq calculated in the QPCR run adjusting for sample volume. The entire volume of concentrated cDNA for each sample was amplified with sequencing adaptors using NEB-Next High-Fidelity 2x PCR Master Mix (New England Biolabs): 25 ul Master Mix, 2.5 ul of both 10 uM forward and reverse primer (Supplementary Table X), 4 ul of cDNA, and 16 ul H_2O. The reaction and cycling conditions are as follows: 98°C for 30 seconds, X cycles of 98°C for 8 seconds, 66°C for 20 seconds, and 72°C for 10 seconds,

followed by an extension of 72°C for 2 minutes. Amplified DNA was purified with the DNA Clean and Concentrator kit and gel isolated from a 1% agarose gel with the Zymoclean Gel DNA Recovery Kit. Library concentrations were quantified using a TapeStation 2200 and a Qubit. The libraries were sequenced with an i7 index read and a single end 75-bp read on a NextSeq 500 in high-output mode.

Quantification and Statistical Analysis

Barcode Mapping

We used the BBTools suite(https://jgi.doe.gov/data-and-tools/bbtools/) of programs to process our sequencing data using the default settings unless otherwise noted. First, we used BBDuk2 to filter out any reads matching PhiX (k=23, mink=11, hdist=1) and to trim off any Illumina sequencing adapters. We then used BBMerge to merge our paired end reads. We performed another round of trimming with BBDuk2 to ensure no adapters were left over after merging and to remove any sequence with an N base call. After merging and trimming the reads, we used a custom Python script (bcmap.py) to generate a consensus nucleotide sequence for each barcode.

Briefly the script works as follows. First, we split each read into the 15 nt barcode and its corresponding variant. We then generate a dictionary that maps each barcode to its list of unique sequences and their counts. To enable majority basecall we drop any barcode that has less than 3 reads. We then pass the barcodes through a series of filters to eliminate potential errors introduced by barcodes that are mapped to multiple variants. Since we barcoded and mutagenized the ADRB2 gene in separate pieces, barcodes can be associated to variants from different pieces. We address this case by using BBMap to align every barcode's sequences to the ADRB2 reference and consider that barcode to be contaminated if any sequence aligns >5

nt away from the most common sequence. Another source of contamination comes from our chip-synthesized library itself, which contains a significant number of single base deletions. We consider a barcode contaminated if it has any sequences of different lengths as it is unlikely that a single base deletion will come from an Illumina sequencer by chance. However, these filters would not catch the case where a barcode is contaminated with variants from the same piece of ADRB2. As we only synthesized the missense variants, we expect variants within the same piece of ADRB2 to be a Levenshtein distance of 4 from each other on average (approximately two changes to WT and two changes to a new codon). Thus, we drop any barcode that has a sequence with >1 read at a Levenshtein distance of 4 away from that barcode's most common sequence. Lastly, we generate a consensus sequence by taking the majority base call at each position and call an N at any ties.

After we associate each barcode with its consensus sequence, we use a series of different alignments to determine that sequence's identity. To find the designed missense variants in our library, we use BBMap to search for barcodes that an exact alignment to them. To find frameshift mutations, we use BBMap to align the consensus sequences to the ADRB2 reference and parse the resulting CIGAR strings for indels with a simple python script (classify-negs.py). Finding synonymous mutants required more processing as each sub-library did not start at a complete codon. We first used the rough BBMap alignment to determine what ADRB2 chunk each sequence was associated with. We then used a custom python script (synon-filter.py) to trim up to the last whole clonal codon, as the first few codons of each sequence were part of the clonally backbone and are unlikely to have any errors. Finally, we translated the resulting sequences, aligned the protein sequence to the ADRB2 coding sequence with a Smith-Waterman aligner from the Parasail library (https://github.com/jeffdaily/parasail), and retained perfect translations with the correct length.

Data Normalization

We incubated our cellular library with forskolin to activate the cAMP reporter in each cell, providing an agonist-independent measurement of maximal reporter activity. This measurement can be used to approximate cellular copy number. To ensure that barcodes with low cellular representation are excluded from our analyses we require that all forskolin barcodes be present in both repeats, we normalize our read counts to sequencing depth, we average the two repeats together, and filter out any barcodes less than 0.2 RPM (~8-10 reads at our sequencing depth). Next, we use this list of barcodes to control for copy-number variation in our measurements. We first require that all of the barcodes in the forskolin condition are also present in our drug conditions, and add a pseudocount that is scaled relative to the condition with the fewest number of reads (N/min(N)). This explicitly sets missing barcodes to the pseudocount. We then normalize each condition to its read depth (including added pseudocounts) and divide this value by its associated forskolin value. We also excluded barcodes with high forskolin counts (>= 10 RPM) as they are systematically less induced in the drug conditions relative to other barcodes.

With a filtered set of barcodes in place, we averaged together all measurements for each variant (median 11 barcodes per variant), keeping the repeats separate. To make our values more interpretable, defined activity as the ratio of these values to the mean frameshift. We then averaged the relative activities of the two repeats together and used propagation of uncertainty to combine their standard deviations.

Conservation, EVMutation, and gnomAD

To calculate sequence conservation, we aligned 55 ADRB2 orthologs from the OMA database (entry: HUMAN24043) using MAFFT with the default settings (mafft --reorder --auto). We then used the Jensen-Shannon Divergence to score this alignment and majority basecall to generate a consensus sequence ignoring any gaps if they made up < 35% of the alignment at

that position. Using EMBOSS Needle to align this consensus sequence back to the ADRB2 reference, we found the consensus sequence had a two nt insertion at positions 360 and 361. We excluded these positions for the purposes of our analyses. For both EVMutation and gnomAD, we simply downloaded the results for ADRB2.

Unsupervised Learning

We performed a number of preprocessing steps before running UMAP on our data. First, we grouped amino acids into 8 different classes based on their physiochemical properties ((+) - H, K; (-) - D, E; Aromatic - F, W, Y; Amide - N, Q; Nucleophilic - C, S, T; Hydrophobic - I, L, V, M; Small - G, A; Proline - P) and averaged their relative activities. Next, we standardized the log2 relative activity values of each group and used mean imputation to model missing data for any missing AA groups at a given position. Finally, we combined the data from every drug condition into a 412 x 32 design matrix in which the columns are an AA group at a specific condition and the rows are the positions in the protein.

With our data processed, we used the R implementation of UMAP to run hyperparameter search of all combinations of UMAP embeddings with the parameters n_neighbors = (4, 8, 16, 32) and n_components = (2, 3, 4, 5, 6, 7, 8, 9, 10), holding min_dist=0 and n_epochs=2000 constant. This provided a variety of different representations of our data that we used HDBSCAN to search for clusters in these embeddings (R package dbscan; minPts = 10). To ease interpretation of the clustering, we plotted the HDBSCAN results onto a 2D UMAP embedding with the following parameters: n_neighbors=4, min_dist=0, n_components=2, n_epochs=2000, and random_state=3308004 using the Python implementation (https://github.com/lmcinnes/umap). We found the cluster assignments to be largely robust across the different embeddings, and used them to guide our manual cluster assignment.

Identification of Activating and Hypomorphic Mutations

We defined a variant to be “activating” if its mean activity minus its standard deviation were greater than the mean synonymous variant. Similarly, we defined a variant to be hypomorphic if its mean activity plus its standard deviation were less than the mean synonymous variant.

Mutational Tolerance

We defined mutational tolerance as a given residue’s ability to accommodate all amino acid substitutions. To calculate this, we first capped the maximum our activity values at WT-like activity (the mean of the synonymous barcodes). Similarly, we capped the minimum activity at the mean of the frameshifts (1 on our activity scale). By limiting our activity measurements to this range, we ensure that individual substitutions do not have an outsized effect on the mutational tolerance. Next, we averaged the activities of every amino acid substitution for each position in the β_2AR. Finally, we scaled the mutational tolerance values to lie on a 0-1 scale.

Statistical Tests

All statistical tests unless otherwise noted are the two-sided Mann-Whitney U and were performed in R (version 3.5.x) using the wilcox.test function.

Structural Modeling and Solvent Accessible Surface Area

All molecular graphics and analyses were performed with the UCSF Chimera package. To determine if a given in the β_2AR points into the core of the protein or into the lipid membrane, we used FreeSASA (version 2.0.3) to calculate the Solvent Accessible Surface Area (SASA) of the G_s-bound β_2AR (PDB: 3SN6). The G_s occludes the intracellular surface of the β_2AR thereby reducing the SASA of residues on the intracellular surface. Similarly, the extracellular surface is mostly blocked by the extracellular loops. Finally, we used the Orientations of Proteins in Membranes (OPM) database to filter out any residues outside of the lipid membrane from our

analyses. To quantify charge sensitivity, we calculated the average activity for H, K, R, D, and E substitutions at each agonist concentration for residues in the lipid membrane. We then multiplied the values by -1 and standardized the results within each concentration such that the values were mean-centered and scaled by their standard deviation. We calculated hydrophobic sensitivity (I, L, V, M) in an analogous manner. Next, we classified residues that had above average charge sensitivity and below average hydrophobic sensitivity as being exclusively charge sensitive. Conversely, we classified residues that had above average charge sensitivity and above average hydrophobic sensitivity as being intolerant.

References

1. 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).

2. Isberg, V. *et al.* GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **44,** D356–64 (2016).

3. Granier, S. & Kobilka, B. A new era of GPCR structural and chemical biology. *Nat. Chem. Biol.* **8,** 670–673 (2012).

4. Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. GPCR Dynamics: Structures in Motion. *Chem. Rev.* **117,** 139–155 (2017).

5. Schönegge, A.-M. *et al.* Evolutionary action and structural basis of the allosteric switch controlling β 2 AR functional selectivity. *Nat. Commun.* **8,** 2169 (2017).

6. Pei, G. *et al.* A constitutively active mutant beta 2-adrenergic receptor is constitutively desensitized and phosphorylated. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 2699–2702 (1994).

7. Valentin-Hansen, L. *et al.* The arginine of the DRY motif in transmembrane segment III functions as a balancing micro-switch in the activation of the β2-adrenergic receptor. *J. Biol. Chem.* **287,** 31973–31982 (2012).

8. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11,** 801–807 (2014).

9. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101,** 315–325 (2017).

10. Hilger, D., Masureel, M. & Kobilka, B. K. Structure and dynamics of GPCR signaling complexes. *Nat. Struct. Mol. Biol.* **25,** 4–12 (2018).

11. Luttrell, L. M. Reviews in molecular biology and biotechnology: transmembrane signaling by G protein-coupled receptors. *Mol. Biotechnol.* **39,** 239–264 (2008).

12. DeWire, S. M., Ahn, S., Lefkowitz, R. J. & Shenoy, S. K. β-Arrestins and Cell

Signaling. *Annu. Rev. Physiol.* **69,** 483–510 (2007).

13. Azimzadeh, P., Olson, J. A., Jr & Balenga, N. Reporter gene assays for investigating GPCR signaling. *Methods Cell Biol.* **142,** 89–99 (2017).

14. Sato, T., Kobayashi, H., Nagao, T. & Kurose, H. Ser203 as well as Ser204 and Ser207 in fifth transmembrane domain of the human beta2-adrenoceptor contributes to agonist binding and receptor activation. *Br. J. Pharmacol.* **128,** 272–274 (1999).

15. Elling, C. E., Thirstrup, K., Holst, B. & Schwartz, T. W. Conversion of agonist site to metal-ion chelator site in the beta(2)-adrenergic receptor. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 12322–12327 (1999).

16. Shenoy, S. K. *et al.* beta-arrestin-dependent, G protein-independent ERK1/2 activation by the beta2 adrenergic receptor. *J. Biol. Chem.* **281,** 1261–1273 (2006).

17. LeProust, E. M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38,** 2522–2540 (2010).

18. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35,** 128–135 (2017).

19. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23,** 1875–1882 (2007).

20. Altenhoff, A. M. *et al.* The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.* **46,** D477–D485 (2018).

21. McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).

22. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. in *Advances in Knowledge Discovery and Data Mining* 160–172 (Springer Berlin Heidelberg, 2013).

23. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).

24. Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res.* **5,** 189 (2016).

25. Weis, W. I. & Kobilka, B. K. The Molecular Basis of G Protein-Coupled Receptor Activation. *Annu. Rev. Biochem.* **87,** 897–919 (2018).

26. Venkatakrishnan, A. J. *et al.* Stable networks of water-mediated interactions are conserved in activation of diverse GPCRs. *bioRxiv* 351502 (2018). doi:10.1101/351502

27. Hanson, M. A. *et al.* A specific cholesterol binding site is established by the 2.8 A structure of the human beta2-adrenergic receptor. *Structure* **16,** 897–905 (2008).

28. Rasmussen, S. G. F. *et al.* Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature* **477,** 549–555 (2011).

29. Moro, O., Lameh, J., Högger, P. & Sadée, W. Hydrophobic amino acid in the i2 loop plays a key role in receptor-G protein coupling. *J. Biol. Chem.* **268,** 22273–22276 (1993).

30. Sheikh, S. P. *et al.* Similar Structures and Shared Switch Mechanisms of the β2-Adrenoceptor and the Parathyroid Hormone Receptor: Zn(II) BRIDGES BETWEEN HELICES III AND VI BLOCK ACTIVATION. *J. Biol. Chem.* **274,** 17033–17041 (1999).

31. O'Dowd, B. F. *et al.* Site-directed mutagenesis of the cytoplasmic domains of the human beta 2-adrenergic receptor. Localization of regions involved in G protein-receptor coupling. *J. Biol. Chem.* **263,** 15985–15992 (1988).

32. Valiquette, M., Parent, S., Loisel, T. P. & Bouvier, M. Mutation of tyrosine-141 inhibits insulin-promoted tyrosine phosphorylation and increased responsiveness of the human beta 2-adrenergic receptor. *EMBO J.* **14,** 5542–5549 (1995).

33. Swaminath, G., Lee, T. W. & Kobilka, B. Identification of an Allosteric Binding Site for Zn2+on the β2 Adrenergic Receptor. *J. Biol. Chem.* **278,** 352–356 (2003).

34. Jensen, A. D. *et al.* Agonist-induced Conformational Changes at the Cytoplasmic Side of Transmembrane Segment 6 in the β2 Adrenergic Receptor Mapped by Site-selective Fluorescent Labeling. *J. Biol. Chem.* **276,** 9279–9290 (2001).

35. Noda, K., Saad, Y., Graham, R. M. & Karnik, S. S. The high affinity state of the beta 2-adrenergic receptor requires unique interaction between conserved and non-conserved extracellular loop cysteines. *J. Biol. Chem.* **269,** 6743–6752 (1994).

36. Dohlman, H. G., Caron, M. G., DeBlasi, A., Frielle, T. & Lefkowitz, R. J. Role of extracellular disulfide-bonded cysteines in the ligand binding function of the beta 2-adrenergic receptor. *Biochemistry* **29,** 2335–2342 (1990).

37. Bhattacharyya, R., Pal, D. & Chakrabarti, P. Disulfide bonds, their stereospecific environment and conservation in protein structures. *Protein Eng. Des. Sel.* **17,** 795–808 (2004).

38. Dong, C., Filipeanu, C. M., Duvernay, M. T. & Wu, G. Regulation of G protein-coupled receptor export trafficking. *Biochim. Biophys. Acta* **1768,** 853–870 (2007).

39. Ozgur, C., Doruker, P. & Akten, E. D. Investigation of allosteric coupling in human β2-adrenergic receptor in the presence of intracellular loop 3. *BMC Struct. Biol.* **16,** 9 (2016).

40. Ozcan, O., Uyar, A., Doruker, P. & Akten, E. D. Effect of intracellular loop 3 on intrinsic dynamics of human β2-adrenergic receptor. *BMC Struct. Biol.* **13,** 29 (2013).

41. West, G. M. *et al.* Ligand-dependent perturbation of the conformational ensemble for the GPCR β2 adrenergic receptor revealed by HDX. *Structure* **19,** 1424–1432 (2011).

42. Galandrin, S., Oligny-Longpré, G. & Bouvier, M. The evasive nature of drug efficacy: implications for drug discovery. *Trends Pharmacol. Sci.* **28,** 423–430 (2007).

43. Reiter, E., Ahn, S., Shukla, A. K. & Lefkowitz, R. J. Molecular mechanism of β-arrestin-biased agonism at seven-transmembrane receptors. *Annu. Rev. Pharmacol. Toxicol.* **52,** 179–197 (2012).

44. Tate, C. G. & Schertler, G. F. X. Engineering G protein-coupled receptors to facilitate their structure determination. *Curr. Opin. Struct. Biol.* **19,** 386–395 (2009).

45. Serrano-Vega, M. J., Magnani, F., Shibata, Y. & Tate, C. G. Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 877–882 (2008).

46. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11,** 499–507 (2014).

47. Duportet, X. *et al.* A platform for rapid prototyping of synthetic gene networks in mammalian cells. *Nucleic Acids Res.* **42,** 13440–13451 (2014).

48. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* **45,** e102 (2017).

CHAPTER FOUR

Conclusions and Future Directions

Summary of Novel Technology

In the prior two chapters, I describe a new high-throughput method for measuring receptor-ligand interactions. By engineering human cell lines to stably express receptor libraries uniquely tagged with short DNA barcodes, we can measure their activation in multiplex with RNA-seq via genetic reporters. First, I presented its applicability for screening large sets of chemicals against moderate-sized receptor libraries cost effectively at scale. Second, I demonstrate the platform's ability to quantitatively screen massive receptor libraries (thousands of receptors) against a single input. Broadly, this technology can be applied to the disparate biological niches that receptors occupy. I believe it will be a powerful tool for mapping receptor-ligand interactions and understanding receptor biochemistry.

Summary of Findings

In Chapter 2, we measure the interactome between 39 murine olfactory receptors and 182 chemicals (~7100 interactions) in 20 96-well plates. If assayed the traditional way, individually, the screen would require ~800 96-well plates and be overwhelmingly cost- and labor-prohibitive. From this screen, we identified ~79 novel interactions between 28 receptors. From this set, 15 of the receptors were orphan -- they had no previously known ligands mapped to them. Additionally, through both manual inspection and utilization of a molecular autoencoder, we were able to identify features of the chemical specificity for individual ORs.

In Chapter 3, we systematically mutagenize the beta-2 adrenergic receptor and generated all ~8000 possible missense variants of the protein. From there, we measured the relative fitness for each variant, using isoproterenol-induced g protein signaling as a proxy. We demonstrated our platform's capability to reveal structural and environmental features of regions and residues of the receptor: profiling side chain accessibility and presence of order. These analyses can be applied as a structural aid in conjunction with experimentally determined structures or in cases

where such structures are unavailable. Furthermore, we identify critical residues for receptor function, both recapitulating the importance of known motifs and identifying novel positions for investigation. We follow up on one such residue, W99, an extracellular tryptophan that has a conserved contact with a disulfide bond across class A GPCRs. From there, we combine evolutionary coupling and structural analysis to determine the contact is preserved throughout receptor activation. This enabled us to formulate a mechanistic hypothesis: the tryptophan-disulfide bond contact is a structural latch for the inner helix movement during activation. Lastly, we highlight regions of the receptor that harbor mutations that lead to gain of function and increased expression. These mutations, while rare, are informative for understanding receptor allostery and useful for protein crystallization. In summary, we show deep mutational scanning is a powerful tool for biochemical characterization of a GPCR.

Future Directions for Olfaction and GPCR Mutagenic Screens

Our initial 39 receptor x 182 chemical screen revealed many novel interactions. However, it was largely a pilot-scale screen with the intention of showcasing the new method. The larger overall goal is to understand the molecular basis for odor perception. To do so, requires screening larger receptor libraries against larger chemical panels. The human olfactory receptor repertoire has 396 functional members[1]. Our next goal is to screen the entire human repertoire against a 1000-member chemical panel. This screen is about 60 times greater in scale than our assay from Chapter 2, however it is well within the scope of the technology. Accomplishing such a goal would likely deorphanize a significant number of the 86% of human receptors currently without a mapped ligand[2]. Additionally, it would provide a large enough data set to perform accurate computational predictions of receptor-odorant specificity. Bridging those two knowledge gaps would be very enabling for researchers trying to understand the neurobiological processes governing the sense of smell.

Comprehensive mutagenesis and functional screening of GPCRs is an apt approach for answering many questions within GPCR biology. Biased signaling, the phenomenon where GPCRs can signal through their various effectors at different rates, is a highly active area of GPCR research[3,4]. Briefly, individual ligands are known to differentially activate signal pathways for a given GPCR and this has tremendous therapeutic implications[5]. For example, the mu opioid receptor, when targeted by morphine, activates both g protein and arrestin signaling[6]. The analgesic properties of the molecule largely arise from g protein signaling where negative side effects such as respiratory depression, nausea, etc stem from arrestin signaling. The development of molecules that can preferentially activate individual signaling pathways is of great interest for the pharmaceutical industry. By comparing the mutagenic profiles of both arrestin and g protein signaling, we can identify the different allosteric networks and protein-protein interactions responsible for both signaling pathways. A better understanding of the distinct biochemical mechanisms for both signaling pathways will increase our ability to precisely target them.

Industry Applications of High-throughput Receptor-Ligand Screening

A critical first step of the drug discovery pipeline is to perform high-throughput screening (HTS) against a protein target of interest against large chemical libraries. These chemical libraries can often include hundreds of thousands to millions of compounds. Following HTS, compounds deemed hits undergo more rigorous characterization to determine efficacy and safety in cell culture, mouse, and ultimately human models. Often, in this long, laborious process compounds fail because of safety including modes such as on-target toxicity. On-target toxicity occurs when a molecule induces toxicity by modulating its protein target too acutely or in an unintended cell type.

Polypharmacology is a natural biological mechanism for avoiding such toxicity. Individual cell types express different repertoires of receptors and cell-type specificity can be achieved by molecules that target sets of receptors[7–9]. By weakly binding individual target receptors, but producing a strong response when encountering multiple cognate receptors, a molecule can trigger an appropriate therapeutic response in selected cell and tissue types. In fact, the polypharmacological nature of so-called 'dirty drugs' is often critical for the efficacy. Clozapin, an atypical antipsychotic, is one of the most potent treatments for schizophrenia[10]. It binds many receptors and is a dopaminergic and serotonergic antagonist and more targeted compounds that hit individual receptors have been shown to have little efficacy[11].

Performing HTS in search of molecules with affinity for multiple receptors is difficult when screening receptor ligand pairs individually. Like in the case of olfactory receptors, the number of potential interactions to screen is costly and labor intensive. However, our screening platform is uniquely suited to overcome this challenge. The ability to practically develop molecules that target cohorts of receptors has potential to be transformative for the pharmaceutical industry and biomedical research in general.

References

1. Malnic, B., Godfrey, P. A. & Buck, L. B. The human olfactory receptor gene family. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 2584–2589 (2004).

2. *Springer Handbook of Odor*. (Springer International Publishing, 2017).

3. Reiter, E., Ahn, S., Shukla, A. K. & Lefkowitz, R. J. Molecular mechanism of β-arrestin-biased agonism at seven-transmembrane receptors. *Annu. Rev. Pharmacol. Toxicol.* **52,** 179–197 (2012).

4. Zhou, X. E., Melcher, K. & Xu, H. E. Understanding the GPCR biased signaling through G protein and arrestin complex structures. *Curr. Opin. Struct. Biol.* **45,** 150–159 (2017).

5. Hodavance, S. Y., Gareri, C., Torok, R. D. & Rockman, H. A. G Protein-coupled Receptor Biased Agonism. *J. Cardiovasc. Pharmacol.* **67,** 193–202 (2016).

6. Madariaga-Mazón, A. *et al.* Mu-Opioid receptor biased ligands: A safer and painless discovery of analgesics? *Drug Discov. Today* **22,** 1719–1729 (2017).

7. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.* **5,** 993–996 (2006).

8. Boran, A. D. W. & Iyengar, R. Systems approaches to polypharmacology and drug discovery. *Curr. Opin. Drug Discov. Devel.* **13,** 297–309 (2010).

9. Zhang, W., Bai, Y., Wang, Y. & Xiao, W. Polypharmacology in Drug Discovery: A Review from Systems Pharmacology Perspective. *Curr. Pharm. Des.* **22,** 3171–3181 (2016).

10. Joober, R. & Boksa, P. Clozapine: a distinct, poorly understood and under-used molecule. *J. Psychiatry Neurosci.* **35,** 147–149 (2010).

11. Roth, B. L., Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* **3,** 353–359 (2004).