A hydrophobic ratchet entrenches molecular complexes

https://doi.org/10.1038/s41586-020-3021-2

Received: 28 November 2019

Accepted: 20 October 2020

Published online: 09 December 2020

Check for updates

Georg K. A. Hochberg[1], Yang Liu[2], Erik G. Marklund[3], Brian P. H. Metzger[1], Arthur Laganowsky[2] & Joseph W. Thornton[1,4] ✉

Most proteins assemble into multisubunit complexes[1]. The persistence of these complexes across evolutionary time is usually explained as the result of natural selection for functional properties that depend on multimerization, such as intersubunit allostery or the capacity to do mechanical work[2]. In many complexes, however, multimerization does not enable any known function[3]. An alternative explanation is that multimers could become entrenched if substitutions accumulate that are neutral in multimers but deleterious in monomers; purifying selection would then prevent reversion to the unassembled form, even if assembly per se does not enhance biological function[3–7]. Here we show that a hydrophobic mutational ratchet systematically entrenches molecular complexes. By applying ancestral protein reconstruction and biochemical assays to the evolution of steroid hormone receptors, we show that an ancient hydrophobic interface, conserved for hundreds of millions of years, is entrenched because exposure of this interface to solvent reduces protein stability and causes aggregation, even though the interface makes no detectable contribution to function. Using structural bioinformatics, we show that a universal mutational propensity drives sites that are buried in multimeric interfaces to accumulate hydrophobic substitutions to levels that are not tolerated in monomers. In a database of hundreds of families of multimers, most show signatures of long-term hydrophobic entrenchment. It is therefore likely that many protein complexes persist because a simple ratchet-like mechanism entrenches them across evolutionary time, even when they are functionally gratuitous.

To understand why multimeric interfaces persist and change over evolutionary time, we studied the evolution of steroid receptors (SRs), a protein family in which dimerization has been maintained for hundreds of millions of years but in which the mechanism of dimerization has diversified. SRs are hormone-activated transcription factors that contain structurally distinct DNA-binding and ligand-binding domains (DBD and LBD, respectively). There are two major phylogenetic classes of SRs (Fig. 1a, b, Extended Data Fig. 1a). One class, the oestrogen receptors (ERs), homodimerize in solution using a large interface in their LBD[8,9] and bind palindromic repeats of a particular six-base-pair DNA response element (ERE)[10]. The other class, called ketosteroid receptors (kSRs) because of the steroidal ligands that activate them, bind to a different palindromic sequence (steroid response element; SRE) via interactions between DBDs[11,12]. kSR-LBDs are monomeric in solution, and the surface region homologous to the ER dimerization interface binds instead to a C-terminal extension (CTE) on the same LBD, which is absent on ERs (Fig. 1b, c, Extended Data Fig. 1b). Previous work has shown that the ancestral protein from which the two clades arose by gene duplication (AncSR1, more than 500 million years ago) specifically bound oestrogens and non-cooperatively bound EREs; specificity for ketosteroids and SREs, as well as DBD-mediated cooperativity, arose on the branch between AncSR1 and AncSR2, the ancient progenitor of kSRs[12,13]. We reasoned that by identifying the ancestral and derived forms of the LBD interface and characterizing their effects on function and biophysical properties, we could gain insight into the factors that caused the persistence and modification of this interface across deep history.

Evolutionary history of SR interfaces

We first inferred the phylogeny of a large alignment of extant SRs and related proteins (Fig. 1a, Extended Data Fig. 1a). We then inferred the posterior probability distribution of the amino acid sequences of AncSR1 and AncSR2 DBDs and LBDs (Extended Data Fig. 2a, b).

We expressed and purified the maximum a posteriori (MAP) LBD sequences and measured their stoichiometry using size-exclusion chromatography with multi-angle light scattering (SEC–MALS) and native mass spectrometry (nMS) (Fig. 1c, d). AncSR1-LBD was predominantly dimeric at 30 μM and 10 μM, indicating a dissociation constant (K_d) in the high nanomolar range, whereas AncSR2-LBD was

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [2]Department of Chemistry, Texas A&M University, College Station, TX, USA. [3]Department of Chemistry – BMC, Uppsala University, Uppsala, Sweden. [4]Department of Human Genetics, University of Chicago, Chicago, IL, USA. ✉e-mail: joet1@uchicago.edu

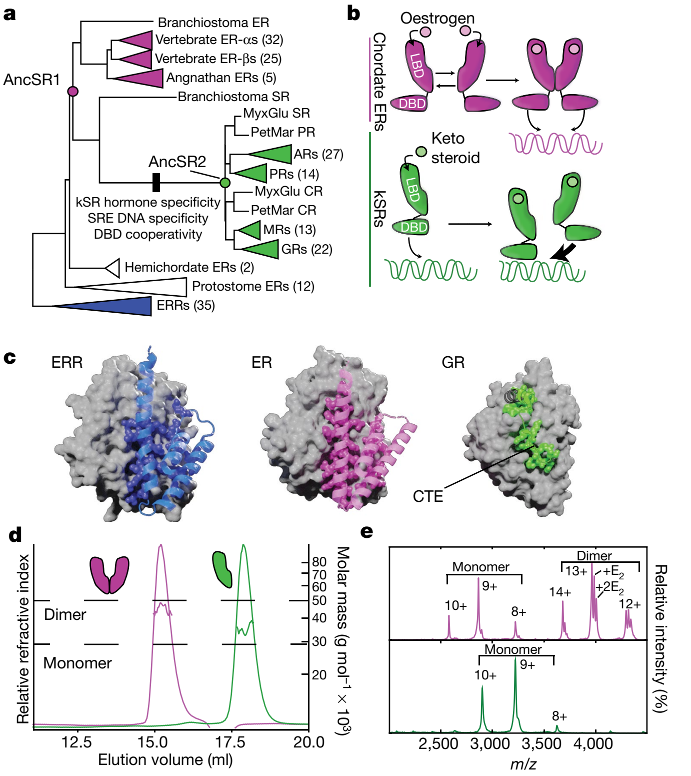


Fig. 1 | Evolution of self-assembly in SRs. a, Reduced phylogeny of steroid and related receptors. Vertebrate ERs (purple), kSRs (green) and ancestral proteins are labelled. Black box, functional changes. Complete phylogeny in Extended Data Fig. 1. AR, androgen receptor; CR, corticosteroid receptor; ERR, oestrogen-related receptor; GR, glucocorticoid receptor; MR, mineralocorticoid receptors; PR, progesterone receptor; MyxGlu, *Myxine glutinosa*; PetMar, *Petromyzon marinus*. Numbers in parentheses denote the number of sequences in each clade. **b**, SR dimerization. ERs dimerize via an interface in the LBD, then bind palindromic EREs. kSR-LBDs are monomeric but cooperatively bind SREs via interactions between DBDs. **c**, LBD interfaces in SRs and closely related receptors. Left, ERR dimer (PDB: 2GP7). Grey surface, one LBD subunit. Blue cartoon and spheres, secondary structural elements and residues that contribute to the interface on the other subunit. Middle, ER-LBD dimer (PDB: 1ERE). Right, GR-LBD monomer (PDB: 4P6X) as grey surface; green spheres, CTE on the same subunit. Cartoon, secondary structure elements connecting CTE to the rest of the LBD. **d**, SEC–MALS of AncSR1 (purple) and AncSR2 (green) at 25 μM. **e**, nMS at 10 μM, with charge series labelled. E_2, dimers bound to 1 or 2 oestradiol molecules.

entirely monomeric at both concentrations. AncSR1 therefore formed LBD-mediated dimers, which were retained in extant vertebrate ERs and lost along the branch leading to AncSR2. Consistent with an ancient origin of LBD dimerization, other members of the nuclear receptor superfamily dimerize through an ER-like interface (Fig. 1c). This inference is robust to statistical uncertainty about the ancestral sequence: alternative versions of the AncSR1- and AncSR2-LBDs, which incorporate the second most likely state at all ambiguously reconstructed sites into a single construct, had the same stoichiometries as the MAP versions (Extended Data Fig. 2c). Moreover, when the LBDs of AncSR1 and AncSR2 were reconstructed using a different plausible phylogeny, AncSR1 remained a dimer and AncSR2 a monomer (Extended Data Fig. 2d–f).

Entrenchment of dimerization

To understand the mechanisms that underlie the long-term persistence of the LBD interface in ERs and its modification in kSRs, we compared the crystallographic structure of AncSR2-LBD[14] to a homology model of AncSR1-LBD. As in modern ERs, the dimer interface in AncSR1 comprises a large patch of hydrophobic residues on helices 10 and 11; the patch on each subunit binds to the corresponding patch on the other to form a tight, water-excluding interface (Fig. 2a). AncSR2 and its descendants retain this patch, but it binds the CTE on the same subunit, which shields the patch from solvent in the monomeric state; this intramolecular interaction is conserved in all descendant kSRs[15–18]. We reasoned that the hydrophobicity of this patch might have entrenched the ancestral interface in AncSR1, because exposure of hydrophobic residues renders many proteins unstable, insoluble, or aggregation-prone[19,20]; acquisition of the CTE, in turn, would have enabled loss of the multimeric state by replacing the intermolecular hydrophobic interaction with a similar intramolecular interaction.

This hypothesis predicts that the LBD interface was already entrenched by the time of AncSR1, and that the CTE interaction that replaced it became quickly entrenched, too. To test the first prediction, we introduced mutations to cause clashes in the AncSR1 dimer interface, thereby preventing dimerization and exposing the interface to solvent (Fig. 2b). We made two mutants: one carrying three historical substitutions from the AncSR1–AncSR2 branch (SR1-LBD(+3)), and another with a non-historical mutation that abolishes dimerization in extant ERs through charge repulsion (SR1-LBD(L184E))[21]. Both mutants formed significantly weaker dimers than AncSR1, with K_d values more than 20-fold higher than that of AncSR1 (Fig. 2c, Extended Data Fig. 3b). In both mutants, exposure of hydrophobic surface area was markedly increased, as shown by binding to 4,4′-bis(1-anilinonaphthalene 8-sulfonate) (bis-ANS), which fluoresces when bound to hydrophobic patches (Fig. 2d). Both mutants had significantly lower melting temperatures than AncSR1-LBD (measured by circular dichroism), although their secondary structures remained largely intact at physiological temperatures (Fig. 2e). Disruption of AncSR1 dimerization without compensating changes would therefore have exposed hydrophobic surface and reduced the stability of the LBD.

Disruption of dimerization severely impairs function: introduction of the LBD dimer-interface mutations into a receptor containing the AncSR1-DBD and -LBD strongly reduced ERE-driven luciferase reporter activation (Fig. 2f). This result could arise for either of two reasons: dimerization might cause the receptor to function better than if it did not have the interface at all by, for example, more effectively occupying DNA response elements or recruiting transcriptional co-activators. Alternatively, disruption of dimerization could be deleterious if exposing the hydrophobic interface simply impairs the stability and function of each monomer.

Four experiments support the latter explanation. First, a chimaera containing AncSR1-DBD and AncSR2-LBD—which is fully monomeric—activates better than AncSR1 from EREs (Fig. 2f), demonstrating that dimerization does not enhance function under our assay conditions, as long as the interface is shielded. Second, to test whether having two active DBDs or LBDs in close proximity is necessary for full activation, we coexpressed the AncSR1-DBD/AncSR1-LBD construct with an excess of a disabled AncSR1 that contains no DBD and an LBD in which the activation function is disabled by a point mutation (SR1-LBD(L126Q))[22]; the resulting heterodimers, which contain a single DBD and a single active LBD but shield the hydrophobic interface from solvent, activated just as well as wild-type AncSR1-DBD–LBD homodimers (Fig. 2f, Extended Data Fig. 3c), indicating that dimerization confers no direct functional benefit. Third, the AncSR1 dimer activates just as well on hybrid response elements containing one ERE half-site and one SRE half-site as it does on ERE palindromes, despite not activating at all from SREs, reinforcing the conclusion that a single effective receptor–half-site complex can drive full activation under our assay conditions (Fig. 2f). Finally, if exposure of the interface explains why interface mutations that reduce the dimerization affinity of AncSR1 impair activation, then driving these mutants to re-occupy the dimeric form by increasing their

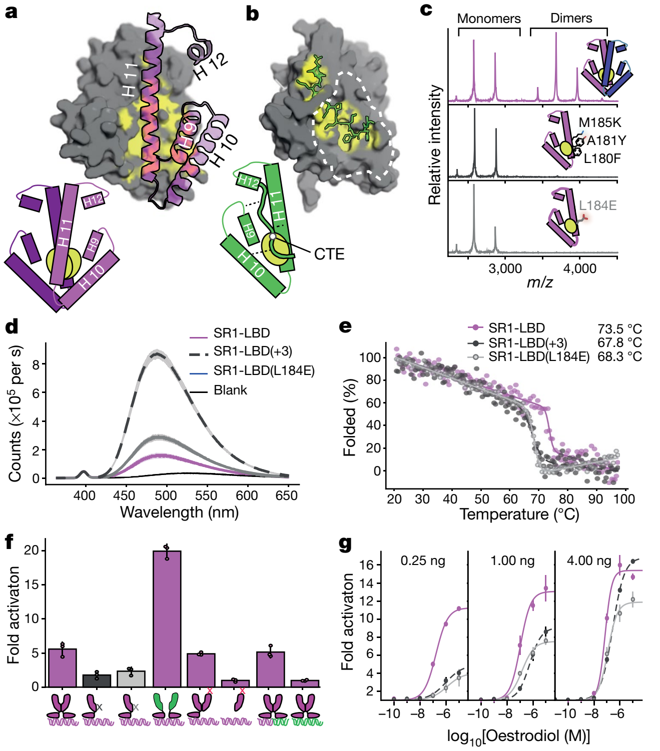


Fig. 2 | The ancestral LBD interface was hydrophobically entrenched. **a**, Homology model of AncSR1-LBD dimer. Grey surface, one subunit, with carbons shielded by the dimer interface in yellow. Purple, helices on the other subunit involved in dimerization. **b**, Atomic structure of AncSR2-LBD monomer (PDB: 4FN9). Grey surface, main body of LBD; yellow, carbons shielded by CTE; green sticks, CTE. Dotted line, surface homologous to the AncSR1 dimer interface. **a**, **b**, Bottom, schematics of AncSR1 (**a**) and AncSR2 (**b**) LBDs. **c**, nMS of wild-type AncSR1-LBD (top) and dimerization mutants SR1-LBD(+3) (middle) and SRI-LBD(L184E) (bottom). Insets, locations of mutations. **d**, bis-ANS fluorescence of AncSR1 and mutant LBDs at 2.5 µM. Line, mean; shading, 95% confidence interval (CI) from three technical replicates. **e**, Circular dichroism melting curves for AncSR1-LBD mutants. Melting points are shown. **f**, Activity of AncSR1 and of chimaeric and mutant receptors in a dual luciferase assay. Purple bars, receptors with shielded hydrophobic interfaces; black and grey bars, dimerization mutants. Cartoons indicate protein constructs (purple, AncSR1; green, AncSR2; black and grey x, dimerization interface mutants as in **d**; red x, activation-function helix mutants). DNA response elements are shown as helices (purple, ERE palindrome; green, SRE palindrome; mixed colours, hybrid containing one ERE and one SRE half-site). Receptor plasmid concentration was 0.25 ng (below saturation for AncSR1; Extended Data Fig. 3a). Transcription was activated with 10^{-6} µM oestradiol or 10^{-7} µM progesterone. Each point shows the average fold change against empty receptor control of three technical replicates. Mean ± 95% CI of three biological replicates. **g**, Activation by AncSR1-LBD and mutants on ERE at variable receptor plasmid and oestradiol concentrations. Mean ± 95% CI of three biological replicates.

concentration should rescue activation. As predicted, increasing plasmid concentration of the mutant receptor by 4- or 16-fold caused them to progressively recover activation. This effect was far greater than that of increasing the concentration of wild-type AncSR1, demonstrating that shielding the interface restores function (Fig. 2g).

These experiments establish that the ancestral dimeric interaction was entrenched because dissociating it into monomers and exposing the interface to solvent impaired the functions of the subunits, not because dimerization caused each subunit to function better than if it had never had the interface. Purifying selection against the deleterious effects of exposing this surface therefore maintains

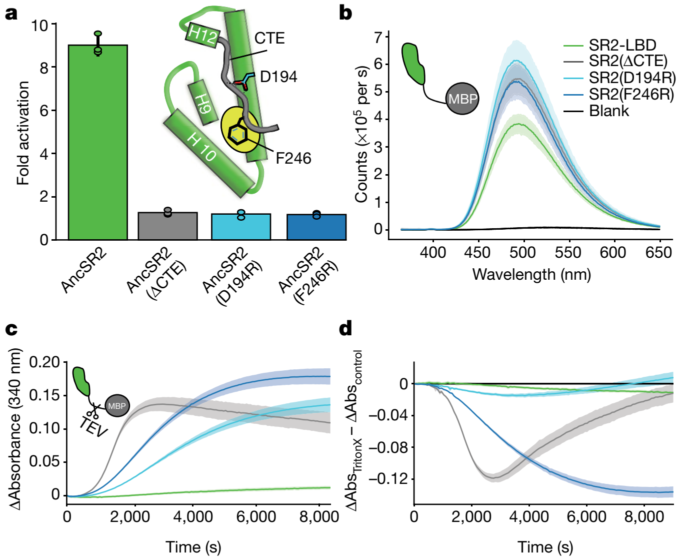


Fig. 3 | AncSR2 traded intermolecular for intramolecular entrenchment. **a**, Fold activation by AncSR2-LBD and mutants on SREs in HEK293T cells using 4 ng receptor plasmid and 10^{-8} µM progesterone. Mean ± 95% CI of three biological replicates (points, each of which shows the mean of three technical replicates). Inset, schematic of AncSR2-LBD. CTE (grey) and sites mutated to disrupt CTE–LBD interaction are indicated. **b**, bis-ANS fluorescence by AncSR2-LBD and mutants fused to MBP. Mean ± 95% CI (dark line and lighter shading) of three biological replicates is shown. **c**, Aggregation (measured as 340 nm absorbance) of AncSR2-LBD and mutants (coloured as in **b**) when MBP tag is removed by TEV cleavage. Mean ± 95% CI from ten technical replicates is shown. **d**, Difference in light scattering between measurements in **c** and the same experiment with 2% Triton X-100. Mean ± 95% CI from ten technical replicates is shown.

the dimeric state. It is possible that dimerization might contribute to function under different assay conditions, but our experiments show that the interface is entrenched even when dimerization does not enhance function per se. This entrenchment has persisted to the present: mutations that interfere with dimerization in human ERs also markedly impair receptor function[21]. Because interface-disrupting mutations compromise function and reduce stability without unfolding secondary structure at physiological temperatures, the likely mechanism is that exposure of the hydrophobic dimerization interface destabilizes the LBD's active conformation relative to its inactive conformations.

Entrenchment of the CTE–LBD interaction

The entrenchment hypothesis implies that the loss in AncSR2 of AncSR1's entrenched dimer interaction was possible only because the interface became shielded by the new AncSR2-CTE, and that this new intramolecular interaction itself became hydrophobically entrenched. To test this prediction, we made AncSR2 mutants in which the CTE was deleted entirely or LBD–CTE interaction was disrupted through charge repulsion (Fig. 3a). All failed to activate in a reporter assay (Fig. 3a), indicating that the interaction was indeed indispensable. To test whether the LBD–CTE interaction is entrenched specifically because exposure of the hydrophobic patch is deleterious, we purified and characterized the AncSR2-LBD–CTE interaction mutants. As predicted, they were very poorly soluble and produced higher bis-ANS fluorescence than did AncSR2 when purified with a maltose binding protein (MBP) tag, confirming that the hydrophobic patch was exposed (Fig. 3b, Extended Data Fig. 4a). When the tag was removed, all CTE mutants aggregated quickly, whereas intact AncSR2 did not (Fig. 3c). Moreover, shielding hydrophobic surfaces in a micelle by adding a mild non-denaturing

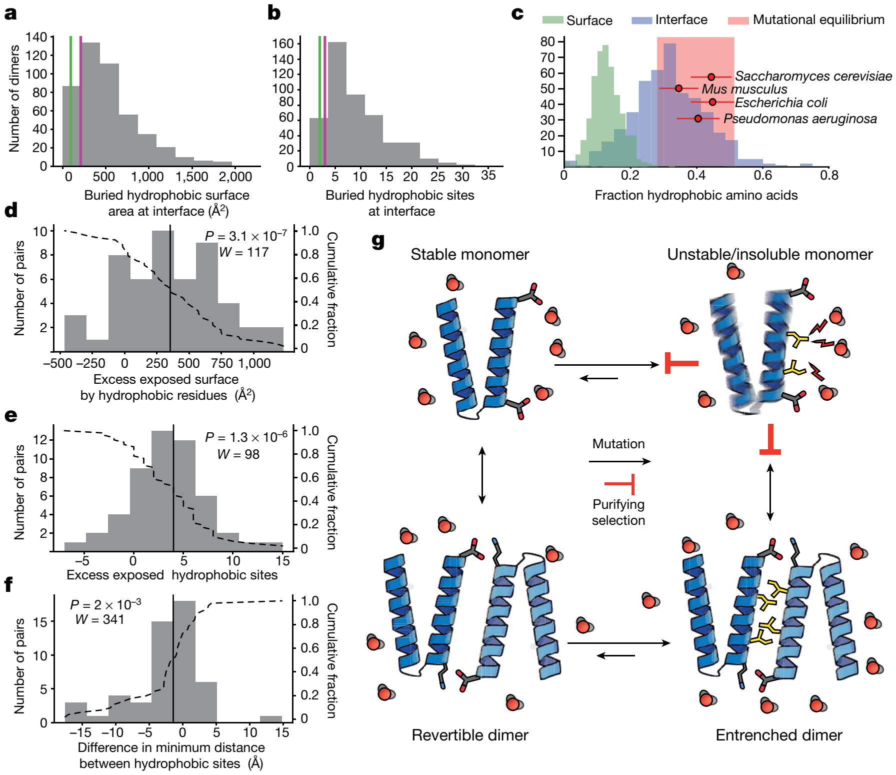


Fig. 4 | Pervasive hydrophobic entrenchment of molecular complexes. **a**, **b**, Surface area (**a**) and count (**b**) of hydrophobic residues (amino acids CFILMVY) buried in dimer interfaces in a database of 466 non-redundant dimer structures. Purple and green lines, AncSR1-LBD and AncSR2-LBD interfaces, respectively. **c**, Dimer interfaces are more hydrophobic than is tolerated at surfaces and are close to the hydrophobicity expected by mutation alone. Histograms, fractions of residues that are hydrophobic on solvent-exposed surfaces (green) or buried in dimer interfaces (blue) on proteins in the database. Red circles, expected fraction of hydrophobic amino acids from mutation alone, based on spectra from mutation accumulation data in four model organisms (Extended Data Fig. 5c). Mean ± s.d. from 200 replicate simulations over long evolutionary timescales is shown. Pink box, ±1 s.d. of the mean across all simulations and spectra. Red dots are offset vertically for visual clarity. **d**–**f**, The extent of hydrophobic entrenchment is shown as the difference in surface properties between dimer subunits (when dissociated into monomers) and their monomeric homologues. Dotted line, cumulative fraction of pairs with greater difference; solid line, median. **d**, Exposed surface area contributed by hydrophobic residues in dissociated dimer subunits minus that on monomeric homologues. **e**, Number of hydrophobic residues on surfaces of dimer subunits minus that on monomers. **f**, Difference in clustering of hydrophobic surface residues between dimer subunits and monomers, calculated as average surface distance from exposed hydrophobic residues to their nearest hydrophobic neighbours. $n = 51$ independent monomer–dimer pairs; two-tailed paired Wilcoxon test. **g**, Mechanism of the hydrophobic ratchet. In monomers (top row), purifying selection (red bar) counteracts mutational pressure (black horizontal arrows, with size representing propensity) towards increased surface hydrophobicity (yellow sticks), which would be deleterious because of increased propensity to aggregate and/or misfold. Once shielded from solvent (red) in dimers (bottom row), hydrophobic mutations are free to accumulate in the buried interface. Purifying selection then preserves the complex.

detergent slowed aggregation of the mutants (Fig. 3d) without affecting tobacco etch virus (TEV) protease cleavage (Extended Data Fig. 4b).

The interaction between the CTE and hydrophobic patch of AncSR2 is therefore entrenched because exposure of the patch causes insoluble aggregates and thereby abolishes AncSR2 function. Using 2-µs molecular dynamics simulations, we found no evidence that the protein unfolds completely (Extended Data Fig. 4c), although denaturation on longer timescales remains possible. Appending the AncSR2-CTE to AncSR1 did not abolish dimerization, indicating that other substitutions during the AncSR1–AncSR2 interval were required to generate a high-affinity interaction with the CTE that outcompetes the dimerization interaction at the same surface (Extended Data Fig. 4d). The extant descendants of AncSR2 have inherited the CTE–patch interaction and require the CTE to function[23,24], indicating that hydrophobic entrenchment continues to preserve this interaction some 450 million years later.

A universal hydrophobic ratchet

Finally, we investigated whether hydrophobic entrenchment is a general evolutionary phenomenon. We compiled a database containing the atomic structures of 466 homodimers[25] and analysed their solvent-exposed surfaces and interfaces. Eighty-three per cent of dimer interfaces in our dataset are more hydrophobic than the AncSR1 interface, and 94% are more hydrophobic than the CTE-shielded patch in AncSR2 (Fig. 4a). Given our experimental finding that the SR interfaces are entrenched, it is likely that most dimers in the database are entrenched, too (Fig. 4a, b).

Inspired by earlier work on evolutionary entrenchment of complexity[4,26–30], we reasoned that entrenchment will arise if two conditions are met: 1) there is a class of substitutions for which the complex state has a higher tolerance than the simple state; and 2) the mutational process alone generates more of these substitutions than can be tolerated in the simple state. Under these conditions, reversion to the simple state will rapidly become unlikely under purifying selection. Specifically, hydrophobic entrenchment will arise if buried interfaces tolerate higher hydrophobicity than exposed sites do, and if mutation tends to produce a higher fraction of hydrophobic residues than surfaces allow.

To evaluate the first condition, we characterized the hydrophobicity of multimeric interfaces and surface-exposed sites in our structural database. Multimeric interfaces are indeed much more hydrophobic than exposed surface sites. Across all multimers in our database, the median fraction of hydrophobic residues at interface sites was 31%, whereas the median at exposed surface sites was 12% (Fig. 4c, Extended Data Fig. 5a, b).

To evaluate the second condition, we investigated whether mutation alone would be expected to generate more hydrophobic amino acids than surfaces tolerate. Hydrophobic amino acids comprise more than 40% of all sense codons; moreover, hydrophobic amino acids are AT-rich, and the mutational process universally favours G/C-to-A/T transitions, irrespective of genomic GC content[31]. To estimate the expected fraction of hydrophobic residues in the absence of purifying selection, we simulated coding sequence evolution over long divergence times using the universal genetic code and empirical mutation spectra observed in mutation accumulation experiments in prokaryotes and eukaryotes with a range of GC contents. We found that the fraction of hydrophobic residues expected from mutation alone is 33 to 45%, far greater than is tolerated on exposed surfaces and much closer to the hydrophobicity of buried interfaces (Fig. 4c). This result holds when sequences are simulated using only the universal genetic code and GC content across a wide empirical range (Extended Data Fig. 5c). Exposed sites are therefore constrained by purifying selection to maintain lower hydrophobicity than would be generated by mutation, and this constraint is greatly relaxed once an interface becomes protected in a multimer. The conditions for hydrophobic entrenchment of interfaces are therefore universally satisfied and arise from general properties of surfaces, interfaces, the mutational process, and the genetic code.

For a multimer to escape entrenchment, its surface would have to return to a level of hydrophobicity that can be tolerated in the unassembled state. To understand the extent of entrenchment, we analysed protein families in our dataset that contain both dimers and monomers. We compared the surface area of all exposed hydrophobic residues on monomers to the total area that would be exposed on the subunits of their homologous dimers if the dimers were dissociated. The degree of apparent entrenchment is large: the hydrophobic surface on dimer subunits was greater than on monomers by a median of 340 Å^2, and the exposed hydrophobic residues were more spatially clustered (Fig. 4d–f). Dimers would have to lose a median of four hydrophobic residues for their surfaces upon dissociation to become similar to their monomeric relatives; in about 20% of cases, dimers are enriched by seven or more hydrophobic residues. For multimerization to be lost, many or all of these excess hydrophobic residues would have to be mutated or compensated for, but mutational propensity towards high hydrophobicity makes this outcome extremely unlikely (Fig. 4g).

Entrenchment of molecular complexity

Our findings suggest that many molecular complexes are likely to be entrenched by a simple biochemical ratchet: mutational propensity drives sites buried in a multimeric interface to accumulate hydrophobic substitutions to a level that then renders reversion to the ancestral monomeric state deleterious. Complexes in which multimerization makes no direct contribution to function or fitness will therefore be preserved by purifying selection. Other biochemical mechanisms may also entrench multimers, deepening or broadening the effect of hydrophobic enrichment of buried interfaces in causing molecular complexes to persist[5].

Hydrophobic entrenchment explains only the persistence of complexes; it neither explains their origin nor is limited to cases in which the multimeric association was initially acquired by drift. A multimer that originated under selection because it enabled a multimer-dependent function would still become entrenched, preserving the association even if multimerization later became functionally dispensable. Entrenchment and functional benefit can coexist: even when multimerization enables beneficial properties such as allostery or cooperativity, hydrophobic entrenchment further reduces the probability of reversion to the monomeric form, because losing the interaction would impair all functions of the protein, not only those that depend on multimerization. The hydrophobic ratchet could even facilitate evolution of assembly-associated functions by preserving interfaces that are initially functionally inconsequential, and mutational pressure towards increasing hydrophobicity could quickly strengthen fortuitous interactions. In some cases, multimerization is undoubtedly functionally important; however, given the universal conditions that cause hydrophobic entrenchment, entrenchment should be the null hypothesis to explain the persistence of any particular complex in the absence of evidence that multimerization enhances its functions.

Entrenchment does not make multimers impossible to lose. Unlikely trajectories that restore solubility to a hydrophobic interface or otherwise shield it from solvent can in rare cases be followed, as apparently occurred in AncSR2. Selection could increase the probability of overcoming entrenchment if the assembled state became deleterious—for instance, if inherited interactions after gene duplication produced interference between paralogues[32,33]. In the absence of such pressures, however, even useless interfaces may persist for long periods of time. The cell may therefore be filled with an ever-accumulating stock of entrenched molecular complexes that never performed a useful function, or long ago ceased to do so.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-3021-2.

1. Marsh, J. A. & Teichmann, S. A. Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem*. **84**, 551–575 (2015).
2. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct*. **29**, 105–153 (2000).
3. Lynch, M. Evolutionary diversification of the multimeric states of proteins. *Proc. Natl Acad. Sci. USA* **110**, E2821–E2828 (2013).
4. Lukeš, J., Archibald, J. M., Keeling, P. J., Doolittle, W. F. & Gray, M. W. How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* **63**, 528–537 (2011).
5. Manhart, M. & Morozov, A. V. Protein folding and binding can emerge as evolutionary spandrels through structural coupling. *Proc. Natl Acad. Sci. USA* **112**, 1797–1802 (2015).
6. Schank, J. C. & Wimsatt, W. C. Generative entrenchment and evolution. *PSA: Proc. Biennial Meeting Philos. Sci. Assoc. 1986*, 33–60 (1986).
7. Muller, H. J. Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics* **3**, 422–499 (1918).
8. Moody, A. D., Miura, M. T., Connaghan, K. D. & Bain, D. L. Thermodynamic dissection of estrogen receptor-promoter interactions reveals that steroid receptors differentially partition their self-association and promoter binding energetics. *Biochemistry* **51**, 739–749 (2012).
9. Tamrazi, A., Carlson, K. E., Daniels, J. R., Hurth, K. M. & Katzenellenbogen, J. A. Estrogen receptor dimerization: ligand binding regulates dimer affinity and dimer dissociation rate. *Mol. Endocrinol*. **16**, 2706–2719 (2002).
10. Robblee, J. P., Miura, M. T. & Bain, D. L. Glucocorticoid receptor-promoter interactions: energetic dissection suggests a framework for the specificity of steroid receptor-mediated gene regulation. *Biochemistry* **51**, 4463–4472 (2012).
11. Alroy, I. & Freedman, L. P. DNA binding analysis of glucocorticoid receptor specificity mutants. *Nucleic Acids Res*. **20**, 1045–1052 (1992).
12. McKeown, A. N. et al. Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* **159**, 58–68 (2014).
13. Harms, M. J. et al. Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors. *Proc. Natl Acad. Sci. USA* **110**, 11475–11480 (2013).
14. Eick, G. N., Colucci, J. K., Harms, M. J., Ortlund, E. A. & Thornton, J. W. Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genet*. **8**, e1003072 (2012).
15. Fagart, J. et al. Crystal structure of a mutant mineralocorticoid receptor responsible for hypertension. *Nat. Struct. Mol. Biol*. **12**, 554–555 (2005).
16. Kauppi, B. et al. The three-dimensional structures of antagonistic and agonistic forms of the glucocorticoid receptor ligand-binding domain: RU-486 induces a transconformation that leads to active antagonism. *J. Biol. Chem*. **278**, 22748–22754 (2003).
17. Sack, J. S. et al. Crystallographic structures of the ligand-binding domains of the androgen receptor and its T877A mutant complexed with the natural agonist dihydrotestosterone. *Proc. Natl Acad. Sci. USA* **98**, 4904–4909 (2001).

18. Williams, S. P. & Sigler, P. B. Atomic structure of progesterone complexed with its receptor. *Nature* **393**, 392–396 (1998).
19. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* **247**, 1306–1310 (1990).
20. Pakula, A. A. & Sauer, R. T. Reverse hydrophobic effects relieved by amino-acid substitutions at a protein surface. *Nature* **344**, 363–364 (1990).
21. Valentine, J. E., Kalkhoven, E., White, R., Hoare, S. & Parker, M. G. Mutations in the estrogen receptor ligand binding domain discriminate between hormone-dependent transactivation and transrepression. *J. Biol. Chem*. **275**, 25322–25329 (2000).
22. Ince, B. A., Zhuang, Y., Wrenn, C. K., Shapiro, D. J. & Katzenellenbogen, B. S. Powerful dominant negative mutants of the human estrogen receptor. *J. Biol. Chem*. **268**, 14026–14032 (1993).
23. Xu, J., Nawaz, Z., Tsai, S. Y., Tsai, M. J. & O'Malley, B. W. The extreme C terminus of progesterone receptor contains a transcriptional repressor domain that functions through a putative corepressor. *Proc. Natl Acad. Sci. USA* **93**, 12195–12199 (1996).
24. Zhang, S., Liang, X. & Danielsen, M. Role of the C terminus of the glucocorticoid receptor in hormone binding and agonist/antagonist discrimination. *Mol. Endocrinol*. **10**, 24–34 (1996).
25. Ahnert, S. E., Marsh, J. A., Hernández, H., Robinson, C. V. & Teichmann, S. A. Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aaa2245 (2015).
26. Finnigan, G. C., Hanson-Smith, V., Stevens, T. H. & Thornton, J. W. Evolution of increased complexity in a molecular machine. *Nature* **481**, 360–364 (2012).
27. Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
28. Gray, M. W., Lukes, J., Archibald, J. M., Keeling, P. J. & Doolittle, W. F. Irremediable complexity? *Science* **330**, 920–921 (2010).
29. Lynch, M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl Acad. Sci. USA* **104** (Suppl. 1), 8597–8604 (2007).
30. Stoltzfus, A. On the possibility of constructive neutral evolution. *J. Mol. Evol*. **49**, 169–181 (1999).
31. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet*. **6**, e1001115 (2010).
32. Hochberg, G. K. A. et al. Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions. *Science* **359**, 930–935 (2018).
33. Kaltenegger, E. & Ober, D. Paralogue interference affects the dynamics after gene duplication. *Trends Plant Sci*. **20**, 814–821 (2015).

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

Methods

Phylogenetics and ancestral reconstruction

Nuclear receptor LBD and DBD amino acid sequences were aligned using Muscle (version 3.8.31)[34]; the alignment was corrected manually, and sites corresponding to lineage-specific insertions were removed. The unalignable hinge region was also removed. We used Protest 3[35] to identify the best-fit model using the Akaike information criterion, which was the Jones-Taylor-Thornton substitution matrix with empirical amino acid frequencies and a four-category gamma distribution of among site rate-variation (JTT+F+G). We used PhyML 3.0[36] to infer the maximum likelihood phylogeny. Two different topologies were used for ancestral reconstruction. First, we imposed on the ML tree particular rearrangements to reflect prior corroborated phylogenetic information from large-scale studies: 1) We moved the two agnathan ER paralogues to form a monophyletic clade sister to all other vertebrate ERs, rather than successive sister clades to the ERb clade, because this duplication is thought to be a lineage-specific duplication[37]; 2) we moved the agnathan CRs to be sister to the gnathostome GR/MR clade, in accord with evidence concerning the timing of vertebrate genome duplications[38] and previous work on SR phylogenetics[14]; and 3) we moved the ERR sequences of *Xenoturbella* and hemichordates to form a monophyletic sister clade to chordate ERRs, instead of forming successive outgroups to bilaterian ERRs, in accordance with evidence at the time[39], although this grouping has now been revised[40]. The phylogenies were rooted between RXRs and SF-1s[41]. Transfer bootstrap values were calculated using the Booster server[42]. Approximate likelihood ratio statistics were calculated using PhyML.

This topology (the 'Bilaterian' topology) (Extended Data Fig. 2d) places the gene duplication that split the chordate ERs from the chordate kSRs deep in the Bilateria, with subsequent losses of kSRs in all protostomes and in all non-chordate deuterostomes. A more parsimonious topology with respect to gene duplications and losses was created by rearranging the two weakly supported branches leading to hemichordate and protostome ERs; this topology places the ER/kSR duplication within the chordates and requires no subsequent gene losses (the 'Chordate' topology) (Fig. 1a, Extended Data Figs. 1, 2d). This topology is only 0.4 lnL units less likely than the Bilaterian topology (Extended Data Fig. 2d). We therefore used the Chordate topology for the primary reconstructions of ancestral proteins, but we also produced and tested alternative reconstructions that used the Bilaterian topology as the underlying phylogeny (AltPhy reconstructions, Extended Data Fig. 2f).

Ancestral sequences were inferred using marginal reconstruction in the codeml module of PAML 4.8[43] and JTT+F+G. Branch lengths and model parameters were inferred separately for the DBD and LBD and the posterior distribution of states at each site then estimated assuming the alignment, tree, and model parameters. The MAP sequence contains the state with the highest posterior probability (PP) at each site. The AltAll sequence contains the MAP state at all sites where only one state has PP > 0.2, and the state with the second highest PP at all other sites.

Protein expression and purification

Codon-optimized sequences coding for SR-LBDs were obtained from Integrated DNA Technologies and cloned into a pET LIC vector containing an N-terminal, TEV-cleavable 6×His MBP tag (Addgene plasmid 27989). Proteins were transformed in BL21(DE3) *E. coli*, and inoculated into 50-ml Luria broth (LB) cultures and grown overnight. For expression, starter cultures were used to inoculate 0.5-l cultures of Terrific Broth, which were grown to an optical density of 0.6–0.8. Hormone dissolved in DMSO (oestradiol for AncSR1-LBDs and progesterone for AncSR2-LBDs) was then added to a final concentration of 50–100 μM. Cultures were induced with 500 μM isopropyl β-D-1-thiogalactopyranoside (IPTG) and incubated with shaking overnight at 22 °C. In the morning cultures were spun down at 5,000*g*, resuspended in PBS, transferred to conical falcon tubes through another 5,000*g* spin and stored at −80 °C until use.

For purification, proteins were re-suspended in buffer A (150 mM NaCl, 20 mM Tris, 20 mM imidazole, 10% (w/v) glycerol, pH 8), supplemented with 20 mM β-mercaptoethanol and one protease inhibitor tablet (Roche) per 0.5 l culture. Cultures were lysed on ice using a sonicator for 30 one-second-on/one-second-off pulses. The lysate was clarified by first spinning at 20,000*g* for 20 min and passed through a 0.45-μm syringe filter. The solution was then loaded at room temperature onto a 5-ml HisTrap nickel column (GE) equilibrated with Buffer A. After washing with at least five column volumes of Buffer A, the protein was eluted with a linear gradient over 12 ml from 0 to 100% Buffer B (150 mM NaCl, 20 mM Tris, 500 mM imidazole, 10% glycerol pH 8). Fractions containing the LBD construct were pooled, approximately 0.5 mg of TEV was added, and the solution was then dialysed overnight at room temperature against 4 l Buffer A containing 50–100 μM oestradiol or progesterone, depending on the construct. The cut product was passed over a HisTrap column equilibrated in Buffer A and the flow-through collected and concentrated. For the last purification step, the sample was injected onto a Superdex 200 10/300 size exclusion column (GE) equilibrated into PBS run at 0.4 ml/min. Fractions containing purified LBD were pooled and concentrated. Glyercol was added to a final concentration of 10%, and then flash frozen in liquid nitrogen and stored at −80 °C until use.

Native mass spectrometry

MS measurements of native protein samples were collected on a Synapt G1 HDMS instrument (Waters Corporation) equipped with a radio frequency generator to isolate higher *m*/*z* species (up to 32k) in the quadrupole, and a temperature-controlled source chamber as previously described[44]. Instrument parameters were tuned to maximize signal intensity while preserving the solution state of the protein complexes. Data were collected in positive ion mode, typically taking 30 s to 60 s per sample. Instrument settings were as follows: source temperature of 25 °C, capillary voltage of 1.7 kV, sampling cone voltage of 100 V, extractor cone voltage of 5 V, trap collision energy of 20 V, argon flow rate in the trap was set to 7 ml/min (5.6×10^{-2} mbar), and transfer collision energy set to 10 V. The T-wave settings were for trap (300 $\mathrm{ms^{-1}}$/1.0 V), IMS (300 $\mathrm{ms^{-1}}$/20 V) and transfer (100 $\mathrm{ms^{-1}}$/10 V), and trap DC bias (30 V). Molar fractions were extracted from spectra analysed using UniDec[45]. Titrations of AncSR1 and its variants were fit to the equation $\frac{2[D]}{[P]_0} = \frac{4[P]_0 + K_\mathrm{d} - \sqrt{(8[P]_0K_\mathrm{d} + {K_\mathrm{d}}^2)}}{4[P]_0}$, using a custom Python script where D is the concentration of dimers, $[P]_0$ is the total concentration of monomers, and K_d is the dissociation constant.

SEC–MALS

SEC–MALS experiments were carried out using a DAWN HELEOS II MALS detector (Wyatt) coupled to an in-line Optilab T-rEX detector for refractive index measurements and a Superdex 200 10/300 size-exclusion column (GE). Prior to the experiment, proteins were dialysed against PBS, and diluted to a final concentration of 0.66 mg/ml. We injected 150 μl protein onto the column for each run. The column was run at 0.5 ml/min at room temperature. Data analysis was carried out using the ASTRA 6.0 software package.

bis-ANS incorporation

For AncSR1 LBD, experiments were carried out at 2.5 μM protein concentration with 40 μM bis-ANS in PBS. For AncSR2, experiments were carried out with 1 μM of protein and 40 μM bis-ANS in PBS. bis-ANS fluorescence was measured on a HORIBA Fluorolog-3 spectrofluorometer, using a 500-μl cuvette. The excitation wavelength was set at 350 nm. Emission was monitored from 350 to 600 nm, with gratings set to 120.500.2. Entrance and exit slit widths were set to 2 and 1 nm, respectively. For the negative control experiment (blank), no protein was added.

Protein stability

CD spectra for thermal melts were recorded on a Jasco J-1500 Circular Dichroism Spectrometer. Proteins were exchanged in 50 mM NaPi, 20 mM NaF, pH 7.4 in a concentrator before the experiment and diluted to a final concentration of 2.5 μM. Spectra were recorded between 260 and 180 nm, at a 1-nm pitch and a scanning speed of 100 nm/min. The temperature was ramped from 20 to 98 °C in 0.5 °C steps at a rate of 3 °C per minute. The data were analysed using the calfitter server[46] using a reversible two state model ($N = D$).

Aggregation assays

AncSR2 and variant LBDs were purified using an Ni column as previously, but in the presence of 100 μM progesterone, and the cleavage and SEC steps were omitted. Their concentrations were determined using a Bradford assay. For the aggregation assay, proteins were diluted to a final concentration of 40 μM, in 150 mM NaCl, 20 mM Tris, pH 7.4, 20 mM imidazole, 10% glycerol supplemented with 5 mM BME, and 0.1 mg/ml TEV protease, either with or without 2% Triton X-100. The solution was transferred into a clear 96-well plate, using 100 μl solution per well and followed at 400 nm on a Perkin Elmer Victor X5 plate reader at room temperature.

Reporter activation assays

Ancestral DBD and LBDs were cloned into pcDNA3, separated by the hinge of the human GR, which neither confers nor abolishes dimerization[10]. Response element plasmids contained four copies of ERE (AGGTCAGAGTGACCT), SRE (AGAACAGAGTGTTCT), or a hybrid ERE–SRE element (AGGTCAGAGTGTTCT), upstream of a luciferase reporter gene. HEK293T cells were obtained from ATCC. Cells were verified by ATCC, using short tandem repeat profiling, cellular morphology, karyotyping and cytochrome oxidase C I assays. We did not check for mycoplasma contamination. Cells were grown at 37 °C and 5% CO_2 in Dulbecco's modified Eagle's medium (DMEM, Gibco) supplemented with 5 mM sodium pyruvate, 10% fetal bovine serum (Gibco), and penicillin streptomycin solution to a final concentration of 1%. For transcriptional activation assays, cells were transfected with a variable amount of receptor plasmid, 40 ng of response element plasmid, 1 ng of a renilla luciferase plasmid for normalization and pUC19 up to a total amount of 100 ng per well. Each well contained 0.05 μl lipofectamine and 0.5 μl plus reagent, to which Optimem (Gibco) was added to bring the total to 65 μl per well. To this, 135 μl of cell suspension was added per well. After 18 h incubation, medium was replaced with 50 μl of DMEM with stripped fetal BSA, supplemented with 1% ETOH and variable concentrations of hormones (see Figures for details). The cells were incubated for 6 h with hormone. Ten microlitres of well solution was then aspirated from each well, and 30 μl luciferase dual-GLO mixture added per well. The mixture was incubated for 2 min at room temperature, and 60 μl per well was then transferred into a white 96-well plate, incubated for 8 more minutes, followed by reading of FFL luminescence on a Perkin Elmer Victor X5 plate reader. Thirty microlitres of Stop and Glo (Promega) mixture was then added to each well and the plate incubated at room temperature for 10 min before recording Renilla luminescence. FFL luminescence was normalized by Renilla luminescence for each well; fold activation is the ratio of the normalized luminescence observed for any treatment divided by normalized luminescence for an empty vector control treated with 1% ETOH.

Homology modelling and molecular dynamics

The AncSR1-LBD structure was modelled using the SWISS-MODEL server using default parameters and specifying the human ER (PDB 1ERE) as the template. For MD simulations, the AncSR2-LBD X-ray crystal structure (PDB 4FN9), and a modified version with all CTE residues removed, were used as starting points using Gromacs software[47]. Each LBD structure was encased in a rhombic-dodecahedral box with a minimal protein–box-edge distance of 1.5 nm. Water and NaCl were added corresponding to a 0.154 M saline solution. Proteins and ions were modelled using the Amber99SB-ildn force field[48] together with the Tip3p water model[49]. The hormone was modelled using GAFF/BCC force field parameters[50]. Virtual sites[51] for the hormone were constructed using the MkVsites tool[52], all bonds were constrained with LINCS[53], and we used SETTLE[54] to keep water molecules rigid, enabling a 4-fs time step in all subsequent simulations. Steepest descent energy minimization was carried out for both systems. Each system was replicated fivefold at this stage, and each replicate was subjected to a 100-ps NVT simulation (using different random seeds used for velocity generation for the different replicas) with position restraints applied to all heavy atoms in the protein and the hormone in order to remove internal strain from the structures. Each replica was simulated for 1 ns under NVT conditions and then for 10 ns under NPT conditions for equilibration, using a Berendsen barostat[55]. Production simulations were run for 2 μs per replica under NPT conditions using the Parrinello-Rahman barostat[56]. The v-rescale thermostat[57] was used for all simulations with temperature coupling. The first half of each simulation was excluded from all analysis to allow for structural relaxation. The backbone root mean square deviation (r.m.s.d.) with respect to the starting structure of the production runs was calculated for each system to assess overall convergence of the simulations. The r.m.s.d. was also calculated between all frames in all trajectories.

Structural bioinformatics

A curated set of structures was downloaded from the PDB based on the database in ref. [25]. Structures were downloaded as biological assemblies. We retained monomers and dimers that were annotated as part of a non-redundant set (filtered at <30% sequence identity) and whose quaternary structure was annotated as correct in the database. To find protein families containing both monomers and dimers, we built a BLAST database from the sequences of the monomers and used the dimer sequences as a query using a 20% sequence identity cutoff. We excluded hits that shared more than 70% sequence identity over the aligned portion, to exclude proteins that can populate both stoichoimetries but that were crystallized independently as dimers and monomers in closely related species.

Structures were stripped of atoms labelled as HETEROATOM in the PDB file using a custom Python script to remove ligands. We created a separate PDB file containing only the first subunit of each dimer. We calculated the exposed hydrophobic surface area of dimers, dissociated dimer subunits, and homologous monomers using the Areaimol program[58] with default parameters. Exposed residues were defined as amino acids for which the solvent-accessible surface area (SASA) was greater than 20% of the maximum theoretical surface area obtained from Gly-X-Gly peptides, where X is the residue of interest[59,60]. Sites buried at the interface were identified as sites at which the difference between SASA in the dissociated monomer and SASA in the dimer was greater than 10% of value in the dissociated monomer[59]. Hydrophobic residues were defined as amino acids CFILMVW. The number of sites with hydrophobic or non-hydrophobic residues was recorded for both exposed and interfacial sites. The total hydrophobic exposed surface area was calculated as the sum of the solvent-exposed area of all exposed hydrophobic residues. This method is conservative, because hydrophobic portions of amino acids not classified as hydrophobic can contribute to hydrophobic surface area. The buried surface area of hydrophobic sites in dimer interfaces was calculated as $A_{\text{interface =}} A_{\text{monomer}} - ½A_{\text{dimer}}$, where $A_{\text{interface}}$ is the surface area buried by hydrophobic sites at the interface, A_{monomer} is the SASA of exposed hydrophobic sites in a dissociated single chain of the dimer, and A_{dimer} is the SASA of exposed hydrophobic sites in the dimer. The number of hydrophobic sites buried at the interface was calculated in the same way.

To compare the exposed hydrophobic surface area of dissociated dimers to their monomeric homologues, we only used dimers for which we found monomeric homologues that differed in the length of the aligned portion of their sequence by no more than nine residues. Each monomer was used only once, so that additional dimers that are homologous to only a previously used monomer were excluded from the calculation. Exposed hydrophobic surface area and the number of exposed hydrophobic sites were calculated using only the aligned portion of the proteins. The degree of clustering among exposed hydrophobic sites was calculated using the DynamXL program[61], which calculates the shortest path along the protein surface between two points on that surface. We calculated all pairwise Cα-to-Cα distances between all exposed hydrophobic sites, with exposure being defined as above. For each site, we recorded the distance to its closest hydrophobic neighbour. Finally, we averaged these distances for all exposed hydrophobic sites within one protein and then calculated the pairwise difference between the averages for dissociated dimers and their monomeric homologues.

Expected hydrophobic content

The G/C content of source organisms in our database was obtained from the NCBI genome database[62]. To produce the expected hydrophobic content of a protein sequence given some specified GC content, we drew nucleotides randomly based on the expected A/T and G/C frequency. The length of the sequence to be drawn was determined by randomly drawing a length from the length distribution in our database of dimers. The sequence was translated using the standard genetic code, and the fraction of hydrophobic amino acids CFILMVW was calculated, with stop codons excluded. This procedure was repeated 200 times to obtain a mean and standard deviation.

To calculate the expected hydrophobic fraction using empirical mutational spectra, we used mutation accumulation experimental data from *S. cerevisiae*[63], *M. musculus*[64], *E. coli*[65] and *P. aeruginosa*[66]. We first constructed an instantaneous DNA mutation rate matrix $\mathbf{Q}$ (Supplementary Table 2a–d) for each species by entering the relative frequencies of observed mutations from each wild-type nucleotide to each other possible nucleotide. The matrix has six free parameters, because each mutation changes the complementary nucleotide on both DNA strands; for example, every A-to-C mutation is associated with a T-to-G mutation, so the rates of these two kinds of mutation are constrained to be equal. Diagonals were filled so each row added to zero, and the matrix was scaled so the sum of diagonals = −1. We then calculated the probability matrix $\mathbf{P}$ of final nucleotide states given each possible starting state across a branch length of 100 expected substitutions per site as $\mathbf{P} = e^{100\mathbf{Q}}$. We simulated a starting DNA sequence given each species' GC content as previously described and assigned a final state at each site given the starting state and $\mathbf{P}$. We translated the resulting DNA sequence using the universal genetic code, excluded stop codons, and calculated the fraction of hydrophobic amino acids. This procedure was repeated 200 times for each species to calculate an average and standard deviation of the expected near-equilibrium fraction of hydrophobic amino acids.

Statistical software

All statistical tests were carried out using scipy 1.2.1. All plots were produced using matplotlib 2.0.0 in Python 2.7.11.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data have been deposited in the Open Science Framework (https://osf.io/) under accession GTJ86, including alignment, phylogeny, sequences and posterior probability of ancestral reconstructions; list of PDB identifiers for coordinates of dimers and monomers in our structural database; and molecular dynamics trajectories.

Code availability

Scripts and code for structural bioinformatics analysis have been deposited at github (https://github.com/JoeThorntonLab).

34. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
35. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
36. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
37. Katsu, Y. et al. A second estrogen receptor from Japanese lamprey (*Lethenteron japonicum*) does not have activities for estrogen binding and transcription. *Gen. Comp. Endocrinol.* **236**, 105–114 (2016).
38. Simakov, O. et al. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat. Ecol. Evol.* **4**, 820–830 (2020).
39. Philippe, H. et al. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature* **470**, 255–258 (2011).
40. Cannon, J. T. et al. Xenacoelomorpha is the sister group to Nephrozoa. *Nature* **530**, 89–93 (2016).
41. Bridgham, J. T. et al. Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol.* **8**, e1000497 (2010).
42. Lemoine, F. et al. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).
43. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
44. Cong, X. et al. Determining membrane protein-lipid binding thermodynamics using native mass spectrometry. *J. Am. Chem. Soc.* **138**, 4346–4349 (2016).
45. Marty, M. T. et al. Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Anal. Chem.* **87**, 4370–4376 (2015).
46. Mazurenko, S. et al. CalFitter: a web server for analysis of protein thermal denaturation data. *Nucleic Acids Res.* **46** (W1), W344–W349 (2018).
47. Abraham, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
48. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).
49. Jorgensen, W. L. et al. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
50. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
51. Feenstra, K. A. et al. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **20**, 786–798 (1999).
52. Larsson, P., Kneiszl, R. C. & Marklund, E. G. MkVsites: A tool for creating GROMACS virtual sites parameters to increase performance in all-atom molecular dynamics simulations. *J. Comput. Chem.* **41**, 1564–1569 (2020).
53. Hess, B. et al. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
54. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).
55. Berendsen, H. J. C. et al. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
56. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
57. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
58. Winn, M. D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).
59. Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**, 53–64 (1997).
60. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. *PLoS One* **8**, e80635 (2013).
61. Degiacomi, M. T., Schmidt, C., Baldwin, A. J. & Benesch, J. L. P. Accommodating protein dynamics in the modeling of chemical crosslinks. *Structure* **25**, 1751–1757.e5 (2017).
62. Wang, D. GCevobase: an evolution-based database for GC content in eukaryotic genomes. *Bioinformatics* **34**, 2129–2131 (2018).
63. Zhu, Y. O., Siegal, M. L., Hall, D. W. & Petrov, D. A. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl Acad. Sci. USA* **111**, E2310–E2318 (2014).
64. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc. Natl Acad. Sci. USA* **109**, E2774–E2783 (2012).
65. Dumont, B. L. Significant strain variation in the mutation spectra of inbred laboratory mice. *Mol. Biol. Evol.* **36**, 865–874 (2019).
66. Dettman, J. R., Sztepanacz, J. L. & Kassen, R. The properties of spontaneous mutations in the opportunistic pathogen *Pseudomonas aeruginosa*. *BMC Genomics* **17**, 27 (2016).

Acknowledgements We thank J. Bridgham for cell culture training and advice, A. Pillai for assistance with experiments, and members of the Thornton Laboratory for comments.

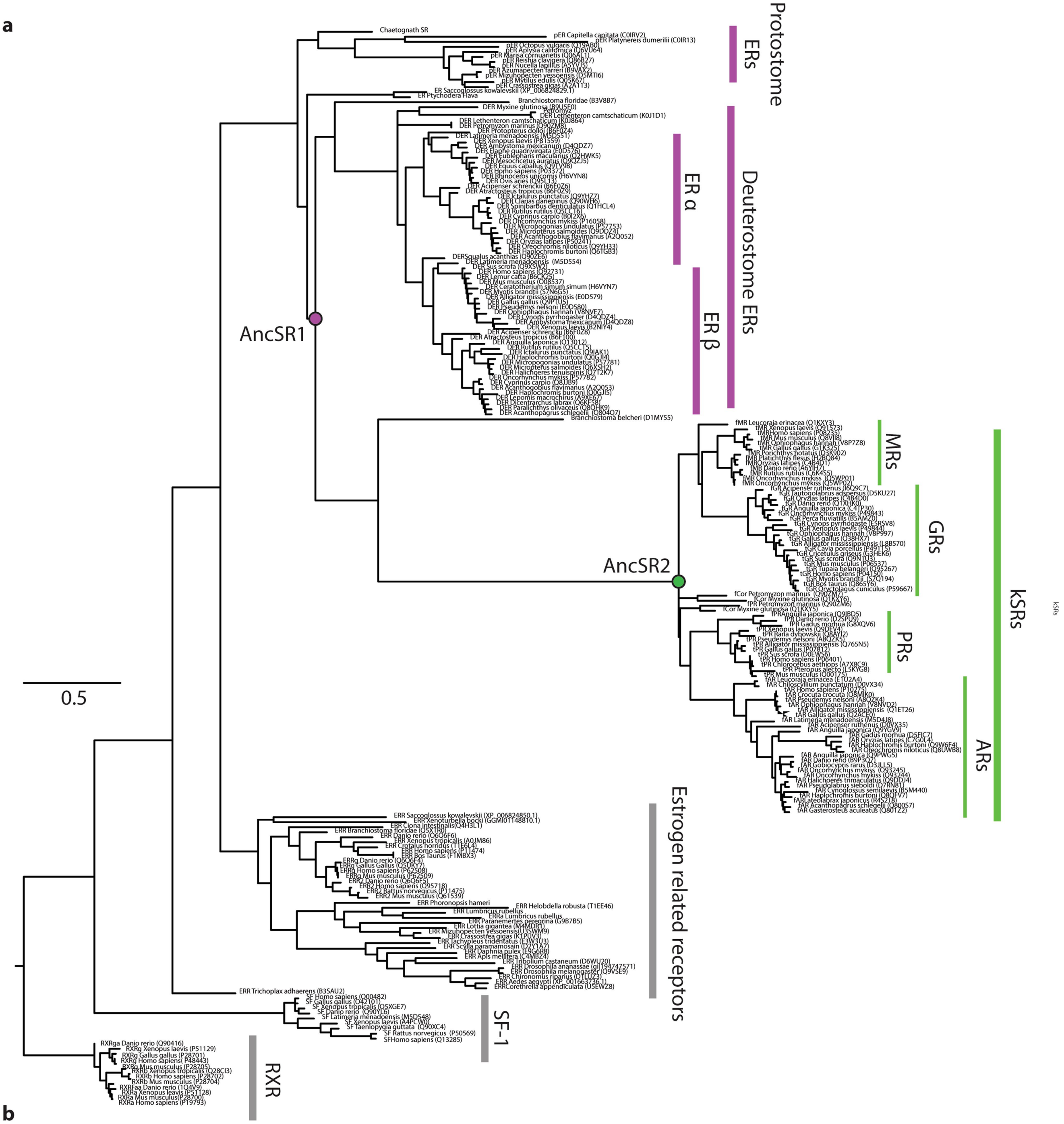
Molecular dynamics computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Projects SNIC 2019/8-36 and SNIC 2019/3-189. Supported by a Chicago Fellowship (G.K.A.H.), NIH R01GM131128 (J.W.T.) and R01GM121931 (J.W.T.).

Author contributions G.K.A.H. and J.W.T. conceived the project and oversaw the manuscript writing. G.K.A.H. performed phylogenetics, ancestral sequence reconstruction, protein purification, cell culture, and biophysical experiments. Y.L. and A.L. performed and interpreted native MS experiments. E.G.M. performed and analysed molecular dynamics simulations. G.K.A.H. and B.P.H.M. designed bioinformatic analyses, which G.K.A.H. performed. G.K.A.H. and J.W.T. interpreted all data. All authors contributed to manuscript writing.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-3021-2.
Correspondence and requests for materials should be addressed to J.W.T.
Peer review information *Nature* thanks Douglas Theobald, Claus Wilke and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

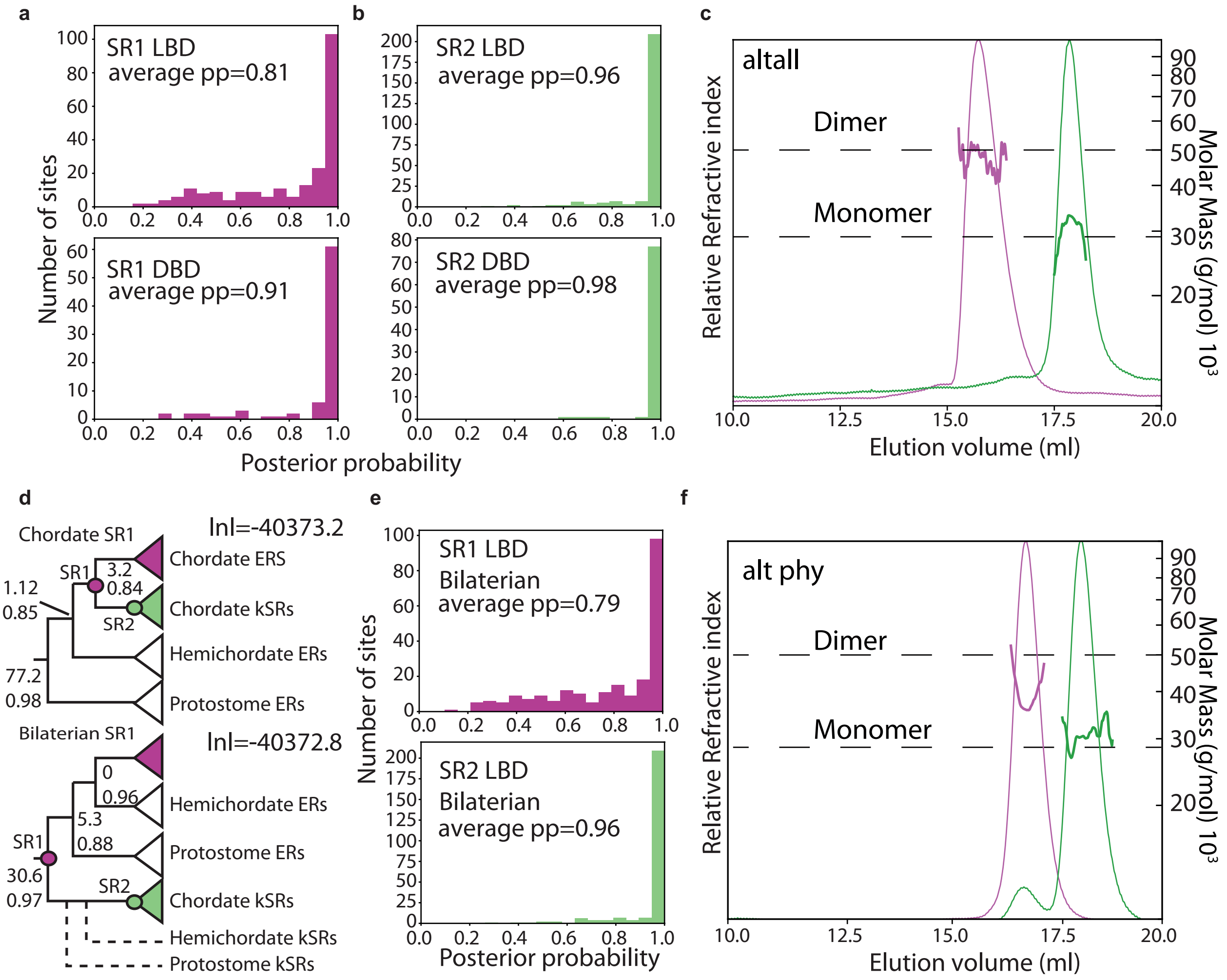
a


Protostome ERs
Deuterostome ERs
ER α
ER β
AncSR1
AncSR2
MRs
GRs
PRs
ARs
kSRs
Estrogen related receptors
SF-1
RXR
0.5
Chaetognath SR
Branchiostoma floridae (B3V8B7)
Branchiostoma belcheri (D1MY55)
ERR Trichoplax adhaerens (B3SAU2)

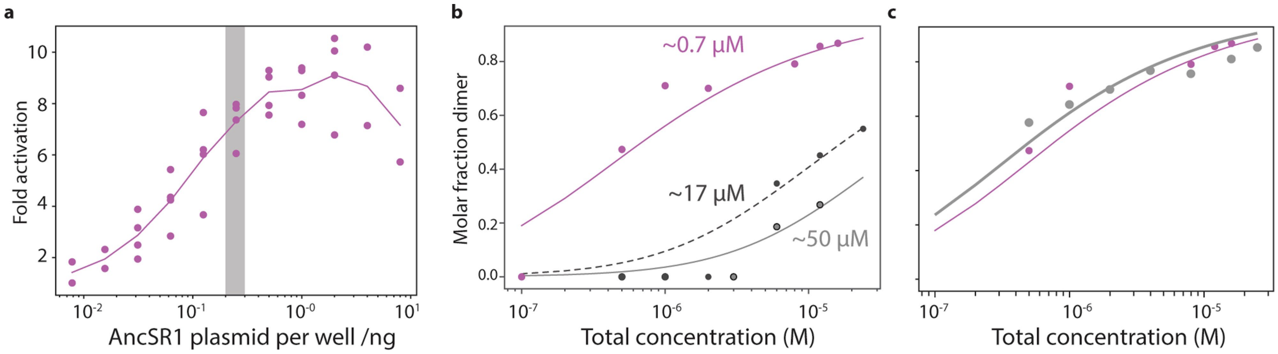
b

```
Human ER   QMVSALLDAEPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKRVPGFVDLTLHDQVHLLECAWLEILMIGLVWRSMEH-PG-
AncSR1     PIISALLEAEPPPLYSNHDPNQPPSEAYLMTSLTNLADRELVHVINWAKQIPGYTDLSLNDQVHLLESSWLEILIIGLVWRSMQH-RGE
Human GR   TLVSLLEVIEPEVLYAGYDSSVPDSTWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDDQMTLLQYSWMFLMAFALGWRSYRQSSAN
AncSR2     SLISILQAIEPEVVYAGYDNTQPDTTNYLLSSLNRLCERQLVSVVKWAKALPGFRNLHIDDQMTLIQYSWMGLMAFAMGWRSYKHTNGQ

Human ER   KLLFAPNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIHL
AncSR1     KLVFAPDLIMDRDDSKAAG-ISEICDHILAIASKFAELNLTKEEYVCLKALVLVNS--------SSKRLESHDQIQKLQDDLTDALADA
Human GR   LLCFAPDLIINEQRMTLPC-MYDQCKHMLYVSSELHRLQVSYEEYLCMKTLLLLSS-------VPKDGLKSQELFDEIRMTYIKELGKA
AncSR2     MLYFAPDLIFNEQRMQQS-AMYDLCLGMRQISQEFVRLQVSQEEFLCMKALLLLST-------VPQEGLKSQAYFDEMRMNYIRELNRA

                                                                                      C -terminal
                                                                                       extension
Human ER   MAKAG--LTLQQQH--QRLAQLLLILSHIRHMSNKGMEHLYSM---KCKNV--VPLYDLLLEMLDAHRLH
AncSR1     VAKSH--PSEP-----RRLAKLLMLLSHVRHISSKGIEHLFDV---KSEGT--VPLYDLLLEMLDAQTSQ
Human GR   IVKREGNSSQN----WQRFYQLTKLLDSMHEVVENLLNYCFQTFL-DKTMS--IEFPEMLAEIITNQIPKYSNGNIKKLLFHQK
AncSR2     IAKRENNSAQS----WQRFYQLTKLLDSMHDLVGGLLQFCFHTFVQSQALS--VEFPEMMSEIISAQLPKVLAGMAKPLLFHKK
```

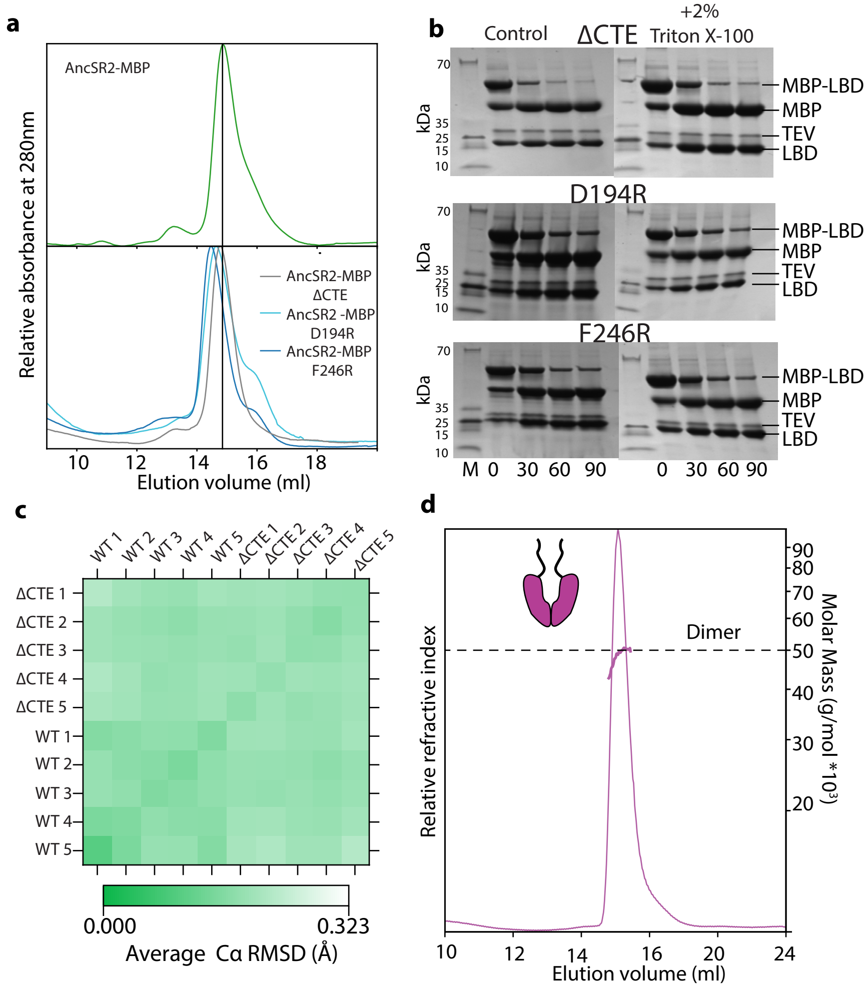
Extended Data Fig. 1 | Phylogeny and alignment of steroid and related receptors. a, Phylogeny of steroid receptors and related nuclear receptor family members. AR, androgen receptors, PR, progestorone receptors, GR, gluccocortociod receptors, MR, mineralocortocoid receptors. Sequence identifiers are in brackets. This topology corresponds to the 'Chordate tree' in Extended Data Fig. 2. Scale bar, expected substitutions per site. **b**, Sequence alignment of the human ER and GR LBDs, with the MAP sequences of AncSR1 and AncSR2. Green, C-terminal extension. Most ERs contain additional sequence on the C terminus that is unalignable, even among ERs.



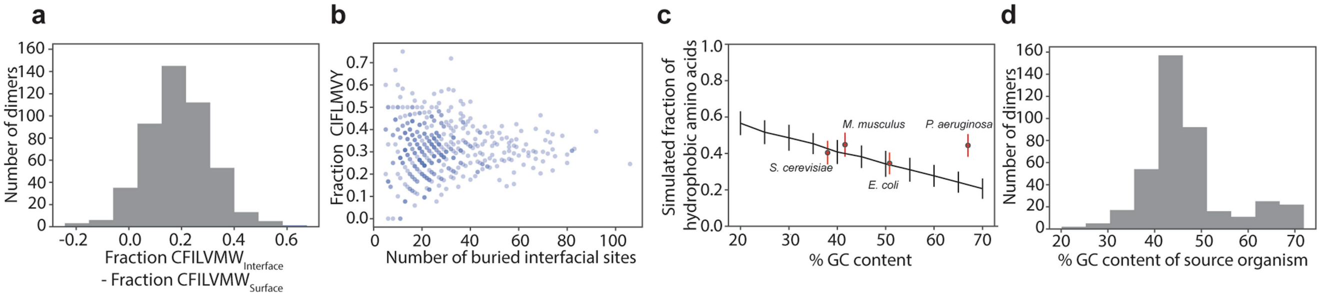
Extended Data Fig. 2 | Robustness of ancestral reconstructions. **a**,**b**, Distribution of posterior probabilities (PP) of the maximum a posteriori (MAP) state at each site in reconstructed LBDs (top) and DBDs (bottom) of AncSR1 (**a**) and AncSR2 (**b**). **c**, Stoichiometry of purified alternative LBD reconstructions (AltAll) of AncSR1 (pink) and AncSR2 (green), as measured by SEC-MALS. AncSR1 is a dimer, AncSR2 a monomer. AltAll reconstructions contain the MAP state at unambiguously reconstructed sites and the state with the next highest PP at all ambiguously reconstructed wites. **d**, The 'chordate' phylogeny (top) was used for primary ancestral reconstructions; it places the gene duplication yielding ERs and kSRs within the chordates. An alternative less parsimonious tree ('Bilaterian' because it places the duplication deep in the Bilateria, bottom), has very slightly higher likelihood but requires two additional gene losses (dashed lines). The Bilaterian topology was used for alternative reconstructions (AltPhy). Node labels, approximate likelihood ratio test statistic and transfer bootstrap value. lnl, log-likelihood. **e**, Distribution of per-site posterior probabilities for reconstructed LBDs on the Bilaterian topology for AncSR1 (top) and AncSR2 (bottom). **f**, Stoichiometry of purified AltPhy versions of AncSR1 (pink) and AncSR2 (green) LBDs, as measured by SEC-MALS. The average molar mass and elution time of AltPhy-AncSR1-LBD are between that of a dimer and a monomer, indicating that it is a fast-exchanging, weaker dimer than other AncSR1-LBD versions.



Extended Data Fig. 3 | Concentration-dependence of activation and dimerization by AncSR1-LBD and mutants. a, Activation of AncSR1 from 40 ng ERE response element plasmid as a function of the AncSR1 plasmid concentration. Grey bar, concentration at which assays in Fig. 2f were performed. **b**, Molar fraction in the dimeric form measured by nMS as a function of LBD concentration for AncSR1-LBD (purple) and dimerization-interface mutants SR1-LBD(+3) (black) and SR1-LBD(L184E) (grey). Dissociation constant (K_d) estimated by nonlinear regression is indicated next to each curve. **c**, Dimeric fraction as a function of LBD concentration for AncSR1-LBD (purple) and activation-helix mutant SR1-LBD(L126Q) (grey), which affects activation but not dimerization.



Extended Data Fig. 4 | Entrenchment of the CTE in AncSR2. a, SEC of AncSR2 LBD (top) and mutants that delete the CTE (ΔCTE) or contain point mutations that impair CTE-LBD interactions (bottom), when fused to MBP. The mutants elute in the same fraction as AncSR2, demonstrating that they are monomeric and that re-exposing the patch does not re-establish dimerization. **b**, TEV cleavage of AncSR2 mutants in the absence (left) and presence (right) of 2% Triton X-100. The positions of bands corresponding to the uncleaved construct, cleaved MBP, cleaved LBD, and TEV protease are indicated. This experiment was performed twice, with similar results. See Supplementary Fig. 1 for uncropped gels. **c**, Average root mean square deviation (r.m.s.d.) from replicate 2-μs molecular dynamics simulations of AncSR2-LBD (WT) and ΔCTE mutant. The average Cα r.m.s.d. in pairwise comparisons of all simulations is shown as a heatmap. **d**, SEC-MALS trace of AncSR1-LBD fused to the CTE of AncSR2-LBD. The LBD is still dimeric.



Extended Data Fig. 5 | Observed hydrophobicity of interfaces compared to expected hydrophobicity from mutation. **a**, Difference between the fraction of residues that are hydrophobic in dimer interfaces versus that on solvent-exposed surfaces of the same proteins. The histogram shows the distribution of this difference across every protein in our structural database. **b**, Fraction of hydrophobic residues in dimer interfaces as a function of the number of interface residues. The variation in the fraction is caused mostly by very small interfaces. **c**, Expected equilibrium fraction of hydrophobic amino acids from mutation alone. Black: expectation based on GC content and the genetic code. Red dots and lines: mean and standard deviation of the hydrophobic fraction of residues observed in 200 replicate simulations using mutational spectra from mutation accumulation experiments (Fig. 4b), plotted against GC content of the organism tested. **d**, GC content of organisms represented by proteins in our database.

nature research

Corresponding author(s): Joseph Thornton

Last updated by author(s): 25 Sept 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We did not use software for data collection.
Data analysis	We used Muscle (version 3.8.31) for alignments, PhyML (version 3.0 for MacOS) for phylogenetc trees, and PAML 4.8 to infer ancestral sequences. Areaimol was used as part of the CCP4 software suite (version 7.0). We downloaded the DynamXL from (http://dynamxl.chem.ox.ac.uk/) on 25.05.2019 (no version number is provided with the software). All statistical tests were carried out in scipy 1.2.1 and plots were created using matplotlib 2.0..0 both using Canopy Python 2.7.11. Molecular dynamics simulations were carried out using Gromacs version 2019.1. We downloaded the Mkvsites tool from https://github.com/ErikMarklund/mkvsites. This is the first release version, and no version number is supplied. ASTRA 6.0 was used for analysis of SEC-MALS data. Code used in this study is deposited at github here https://github.com/JoeThorntonLab.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The alignment, tree, and ancestral sequences generated in this study, as well as the database of structures,
processed PDB files, and calculations of surface exposed sites are deposited in the publicly accessible Open Science Framework repository (DOI 10.17605/OSF.IO/

GTJ86). PDB codes of dimers used for analyses in Figure 4a-c are: 13GS, 1A0F, 1A27, 1A4X, 1A59, 1AC6, 1AD3, 1AFW, 1AH8, 1AJ8, 1AJS, 1AN8, 1ANG, 1AOJ, 1AOZ, 1ASU, 1AX0, 1AXD, 1AY2, 1AZY, 1B0X, 1B49, 1B5O, 1B78, 1B9I, 1BBH, 1BBU, 1BD9, 1BET, 1BG5, 1BHY, 1BJF, 1BK5, 1BO4, 1BSF, 1BT4, 1BUO, 1C1L, 1C6O, 1C80, 1CBG, 1CC5, 1CD8, 1CDC, 1CI7, 1CIV, 1CM9, 1CZ3, 1D1L, 1D2F, 1D6F, 1D6S, 1D7Y, 1DAK, 1DAP, 1DCF, 1DHR, 1DOE, 1DOS, 1DQP, 1DTO, 1DTY, 1DVB, 1DWV, 1DXX, 1DXY, 1DZ3, 1E3I, 1E3V, 1E5D, 1E87, 1EEJ, 1EEM, 1EKF, 1EMD, 1EOV, 1EQT, 1ERT, 1EW3, 1EYV, 1F08, 1F13, 1F1C, 1F36, 1F3B, 1F3H, 1F5M, 1FA9, 1FC4, 1FD9, 1FE4, 1FF5, 1FG3, 1FI4, 1FOC, 1FP1, 1FP2, 1FP5, 1FUX, 1FW1, 1FXD, 1FXR, 1G3M, 1G3S, 1G57, 1G60, 1G6Y, 1G85, 1G8S, 1GAN, 1GDH, 1GFL, 1GFS, 1GNW, 1GPE, 1GPR, 1GQ1, 1GQA, 1GRT, 1GU7, 1GV3, 1GVJ, 1GY8, 1H0C, 1H0X, 1H1M, 1H1Z, 1H4V, 1H7G, 1H8X, 1H8Y, 1H9R, 1HA4, 1HGX, 1HLC, 1HLM, 1HN4, 1HND, 1HQV, 1HT9, 1HUW, 1HW1, 1I0L, 1I1H, 1I2C, 1I2L, 1I3L, 1I52, 1I58, 1IAM, 1ICW, 1IE0, 1IG0, 1IHK, 1II2, 1II5, 1IMA, 1IOM, 1IPE, 1IS3, 1IUG, 1IUJ, 1IUO, 1IVY, 1IYZ, 1IZ3, 1J1I, 1J32, 1J55, 1J5P, 1J6W, 1J6X, 1J93, 1JA3, 1JD0, 1JEZ, 1JF9, 1JHD, 1JHZ, 1JL9, 1JLD, 1JLW, 1JPK, 1JQX, 1JSG, 1JUO, 1JXI, 1JYQ, 1K2E, 1K38, 1K51, 1K66, 1K96, 1K9U, 1KC3, 1KC7, 1KCM, 1KEP, 1KEU, 1KJI, 1KO5, 1KP0, 1KSO, 1KTN, 1L1Q, 1L2U, 1L5B, 1L5Z, 1LBD, 1LC7, 1LCA, 1LCL, 1LED, 1LKZ, 1LPF, 1LVL, 1LXE, 1LYN, 1M0U, 1M6J, 1M7Y, 1MAP, 1MB4, 1MCI, 1MER, 1MG5, 1MH9, 1MJF, 1MK4, 1MKA, 1MKH, 1MO9, 1MOE, 1MP9, 1MPU, 1MQQ, 1MQW, 1MR8, 1MSC, 1MYR, 1MZJ, 1MZV, 1N0H, 1N1A, 1N1Z, 1N2L, 1N31, 1N57, 1N5I, 1N5S, 1NBQ, 1NC3, 1NHJ, 1NL3, 1NMX, 1NNI, 1NPD, 1NRV, 1NSJ, 1NVD, 1NVT, 1NX2, 1O4S, 1O4T, 1O7X, 1O80, 1O89, 1O9E, 1OAH, 1OAN, 1OAT, 1ODB, 1OKI, 1ORF, 1OV3, 1OV4, 1OYB, 1P3W, 1P9E, 1P9O, 1PB1, 1PCZ, 1PD2, 1PIN, 1PIW, 1PRH, 1PRX, 1PSA, 1PSQ, 1PTM, 1PUC, 1PV9, 1PY9, 1PZS, 1Q3O, 1Q4J, 1Q50, 1Q6U, 1Q8Q, 1QI9, 1QIN, 1QKT, 1QLS, 1QMA, 1QMR, 1QO8, 1QOR, 1QP8, 1QPP, 1QRD, 1QUP, 1QVZ, 1QXH, 1QYA, 1QYC, 1QZ9, 1QZR, 1R1D, 1R46, 1R5A, 1R7A, 1R7H, 1R8B, 1R8W, 1RFT, 1RIY, 1RK4, 1RKD, 1RMR, 1RPO, 1RPY, 1RQG, 1RQI, 1RRM, 1RTR, 1S5K, 1S96, 1SB8, 1SE8, 1SEI, 1SEP, 1SEZ, 1SG0, 1SJ1, 1SLA, 1SO6, 1SQD, 1SQE, 1SQI, 1SU2, 1T0I, 1T3C, 1T3I, 1T47, 1T5D, 1TBP, 1TD2, 1TDF, 1TFC, 1TIK, 1TIS, 1TKL, 1TLG, 1TLK, 1TRE, 1TVD, 1TVZ, 1TW3, 1TYT, 1U08, 1U0K, 1U0M, 1U7U, 1UAM, 1UBY, 1UCF, 1UDV, 1UIS, 1UJN, 1ULT, 1UMO, 1UOU, 1UPI, 1UQ5, 1UU2, 1UUF, 1UUJ, 1UWH, 1V08, 1V1O, 1V2F, 1V2Z, 1V3V, 1V47, 1V59, 1V5X, 1V9C, 1VC1, 1VDC, 1VDW, 1VGZ, 1VHD, 1VHM, 1VHZ, 1VJO, 1VLJ, 1VM7, 1VPI, 1W7N, 1X2A, 1X77, 1XKJ, 1Z3A, 1ZOP, 2A4N, 2AG8, 2AJ9, 2AK7, 2APS, 2AY8, 2BK3, 2BLG, 2BT3, 2CAI, 2CCY, 2DAB, 2DLD, 2DPG, 2DTS, 2F48, 2F5C, 2FBZ, 2FF2, 2G6W, 2GSQ, 2H4X, 2HGS, 2HR5, 2ITG, 2J6H, 2LIG, 2NAD, 2NQN, 2OPY, 2OZ9, 2PHK, 2QJD, 2SNW, 2TMK, 2Z9A, 3GAR, 3GSB, 3KZ9, 3LJR, 3ORQ, 3SDH, 4CPV, 4FIV, 4MDH, 4VHB, 4VUB, 5ENL, 6CSC, 6GSY, 8CHO, 9RUB.
PDB codes of monomers in analyses in Figure 4d-f are 2SRC, 1HLB, 2FAM, 1IAL, 1WDN, 1Q20, 1CQW, 1NA5, 1POH, 1E7Z, 5PAL, 1YAT, 1NAT, 1AGI, 1GH2, 1CPR, 1LLN, 1NW8, 1PY5, 1DBP, 2BDA, 1SDZ, 1HQP, 1AE7, 1K2A, 1V9I, 1ZNY, 1MRQ, 1VJW, 1Y0P, 1DP5, 1QYO, 1NCX, 1J2L, 1MIL, 1NRF, 1BSQ, 1JWO, 1G8A, 1IE9, 1G0D, 1E6K, 1UOK, 4ICB, 1ESO, 1ICX, 1H75, 1R6N, 1MEO, 1IJT, 1MB8

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   □ Behavioural & social sciences   □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No populations were sampled for this study. For cell based experiments we chose to perform three biological replicates on three different days, consiting of three technical replicates each. This represented a compromise between adequate throughput and reasnoable quatification of the variability of experiments.
The complete referenced database of protein structures was utilized, then subject to filtering to
eliminate redundancy (see below). Sampling of taxa for phylogeny and ancestral reconstruction was performed to minimize redundancy and maximize breaking up of long branches. Tree inference and ancestral reconstruction are most senstive to long branches near the nodes of interest, which in our case were the divergences at the base of vertebrates. We sampled all genomes available from JGI and NCBI at the time the study was conducted to ensure we covered all available basal branching vertebrates and chordates.

Data exclusions

Data exclusion was only performed for experiments in Figure 4. Protein structures were filtered for redundancy using percent identity cutoffs. We predefined two exclusion criteria: Monomers that are mutant versions of dimeric proteins were to be excluded. Proteins that are misannotated as dimers were also to be excluded. These were defined as higher order oligomers present when symmetry mates are calculated in the structures. Both criteria were assesed manually, by inspection of PDB files.

Replication

Mass spectra were and CD measurements were not replicated SEC-MALS experiments were replicated once with similar results. Cell culture experiments were each replicated on three separate days. On each day we performed three technical replicates for each experiment. Computational measurements of protein surfaces (in Figure 4) were not replicated, because there is only one set of structures for them to be performed on. Simulations of codon evolution were replicated 100 times.

Randomization

There was no group allocation in our experiments.

Blinding

Consistent with accepted practice in biochemistry and molecular biology, protein purifications and cell culture experiments were not conducted blindly.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
□	☒ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T, acquired from ATCC
Authentication	ATCC verify all commercial cell lines using short tandem repeat profiling, cellular morphology, karyotyping and cytochrome oxidase C I assays.
Mycoplasma contamination	We did not test for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used