Evidence that nuclear receptors evolved from terpene synthases

Douglas R. Houston[a], Jane G. Hanna[a,e], J. Constance Lathe[a,f], Stephen G. Hillier[c,1], and Richard Lathe[d,1]

[a]Institute of Quantitative Biology, Biochemistry, and Biotechnology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK; [b]Program in Neuroscience, University of Glasgow, Glasgow, UK; [c]Medical Research Council Centre for Reproductive Health, University of Edinburgh, Edinburgh, UK; [d]Division of Infection Medicine, University of Edinburgh Medical School, Edinburgh, UK; [e]Present address: Arab Academy for Science, Technology, and Maritime Transport (AASTMT), Cairo Campus, Cairo, Egypt; [f]Present address: Ashfield MedComms, Glasgow, UK

[1]To whom correspondence may be addressed. Email: s.hillier@ed.ac.uk or richard.lathe@ed.ac.uk

Ligand-activated nuclear receptors (NRs) including steroid receptors orchestrate development, growth, and reproduction across all animal lifeforms – the Metazoa – but how NRs evolved remains mysterious. Given the universality of terpenoids – including steroids and retinoids – as activating NR ligands, we asked if NRs might have evolved from enzymes that catalyze terpene synthesis and metabolism. We provide evidence suggesting that NRs are a sub-branch of the terpene synthase (TS) enzyme superfamily. Based on over ten thousand 3D structural comparisons, backed up by multiple primary sequence alignments and mapping of ligand-contacting residues, we report that the NR ligand-binding domain and TS enzymes share a conserved core of seven α-helical segments. Primary sequence comparisons reveal potential amino acid sequence similarities between NRs and the subfamily of *cis*-isoprene transferases, in particular dehydrodolichyl pyrophosphate synthase (DHDPPS) and its obligate partner, NUS1/NOGOB receptor. Our results suggest that a ligand-gated receptor may have arisen from an enzyme antecedent, and thus resolve the long-standing debate about whether the ancestral NR was unliganded. This would also explain aspects of NR ligand 'promiscuity', with implications for the development of pharmaceuticals targeting NRs and TS enzymes.

The evolutionary origins of NRs are unknown. NRs are present in Metazoa including sponges, insects, and vertebrates, but not in Archaea, fungi, bacteria, or plants. However, few new protein domains were acquired at the transition from pre-Metazoa to Metazoa, and the majority of innovations were generated through rearrangement of pre-existing components (1). A precursor to the characteristic structure of NRs was probably present before the first multicellular animal species emerged.

NRs are ligand-activated transcription factors that broadly comprise two functional domains, the N-terminal DNA-binding domain (DBD) and a C-terminal LBD. Ligand binding to the LBD results in NR nuclear translocation and DBD binding to response elements in target genes. Of note, the LBD may have arisen independently from the DBD because some NRs lack the DBD or contain an unrelated sequence (2-4).

In the search for potential antecedents to the NR LBD we considered that the inferred earliest NRs are most similar to the NR2 group that includes HNF4, COUP-TF, and retinoic acid receptor (RXR) (5, 6) that heterodimerize with RXR and bind to response elements for the C30 terpenoid retinoic acid in target genes and modulate their transcription (7), leading to an evolutionary 'RXR big bang' (8). Indeed, natural NR ligands including sterols, steroids, retinoic acid, vitamin D, and bile acids are predominantly terpenoids (9), suggesting that terpenoids may have been the earliest NR ligands.

Terpenoids are a vast superfamily of organic chemicals with unified structures based on the C5 isoprene repeat unit. Their biological properties underpin all of life’s sustainability and communication processes from the formation of the first cell membranes through to steroid signaling and beyond (10-12). Terpenoid NR ligands including steroids and retinoids (Figure 1A) are built from C5 isoprene units in a conserved sequence of steps (Figure 1B) generally involving pyrophosphate as the leaving group (13). These include (i) generation of a pyrophosphate-activated C15 trimer, farnesyl pyrophosphate (FPP); (ii) head-to-head linkage of two FPP to generate C30 squalene; followed by (iii) cyclization to generate basic C30 sterols and related molecules (14). The key enzymes involved – FPP synthase (FPPS), squalene synthase (SQS), and squalene cyclase (SQC) – belong to the diverse TS clan of enzymes (https://pfam.xfam.org/clan/Terp_synthase) whose representatives are already present in Archaea and that are conserved between bacteria, yeast, plants, insects, and vertebrates (12, 15, 16).

Given the ubiquity of terpene biosynthesis via conserved TSs that generate NR ligands such as retinoids and steroids (14), we postulated that the LBD of NRs might have evolved from a TS enzyme that bound a structurally similar polyisoprene (steroid-like) substrate (or product). We report here structural and sequence similarities between TSs and the LBDs of NRs that point to terpenoid forerunners of NR signaling at the Metazoan dawn.

Results

A molecular precursor to the signature LBD of NRs is likely to have been present before the Metazoa emerged. In seeking a potential antecedent we considered that protein structure is far more conserved than the primary sequence (17, 18). We therefore performed 3D structure comparisons between NR LBDs and a range of potential candidate terpenoid-binding enzymes that could have provided the framework for the NR LBD. These included enzymes such as retinoid- and steroid-metabolizing hydroxysteroid dehydrogenases and cytochromes P450. All such groups examined, with the exception of TS enzymes, failed to reveal significant structural similarities; we therefore focused on the clan of TS enzymes.

3D Structure Overlaps Suggest that NRs Are an Evolutionary Sub-Branch of TSs. Primary/secondary sequence alignments between TSs and NR LBDs revealed a series of conserved α-helical segments in both TS enzymes and NR LBDs. To address this systematically, we assembled a 64 × 64 matrix of representative TS enzymes and NR LBDs (Table S1 in the supplementary material online), and performed pairwise comparisons using three different 3D structure comparison programs. This revealed that the overall structures of NR LBDs are similar to those of TS enzymes (pairwise *P* values for NR versus TS comparisons were in the range 0.02 to <0.001 for closest matches; Data S1). A typical overlap between the 3D structures of a TS and an NR generated by the 3D comparison program FATCATflexible is given in Figure 1C. The 12,288 comparisons not only emphasized the structural diversity of the TS clan but also revealed that some TS structures are more closely related to NR LBDs (higher similarity score) than to other TS enzymes (distance scores are given in Table S2).

To validate our 3D comparison approach, we separately compared the trees generated by structure/structure versus sequence/sequence comparisons for TS enzymes and NRs. The two types of analysis gave very similar and often identical trees (not presented), confirming that 3D structure comparisons are a valuable adjunct to sequence-based comparisons, particularly when distant protein families are being compared.

Phylogenetic tree drawing based on structure placed NRs as a sub-branch of the TS clan most closely related to the polyisoprene synthases and lipid phosphatases (Figure 1D). The same finding was reiterated in all three 3D comparison methods and with all phylogenetic tree-drawing programs (Figures S1–S5), notably regarding the branch point between TS enzymes and NRs. Figure 1E presents a midpoint-rooted radial tree with branch lengths, again indicating that NRs are most closely related to polyisoprene synthases and lipid phosphatases. These data argue that NRs could have arisen as a sub-branch of TSs.

A Conserved Seven-Helix Core. Because TS enzymes can contain over 30 α-helical segments, to identify a conserved core of α-helical segments that characterizes both NRs and TSs we made pairwise 3D comparisons after stripping away poorly matching segments both between and within TS enzymes and NRs. This analysis revealed that the crucial ligand-binding domain of both TSs and NRs is constituted by a minimal core of seven α-helices that overlap in their primary through tertiary structures (Figure 2). Because standard numbering differs between TSs and NRs, we number these helices in both groups core (c) helices c1–c7 (Figure 2 legend for correspondence with conventional NR helix numbering).

This allowed us unambiguously to align the core α-helical segments of representative NRs to their TS counterparts (Figure 2 and Figure S6), based on the positions of the seven core α-helices, 3D structure overlaps, and primary sequence comparisons. Because residue numbering differs between species, isoforms within a single species, and even between different crystal structures of the same protein, in all cases we provide the primary sequence of human ESR1 isoform A as a reference point (Figure 2).

Ligand-Binding Site Overlap between NR LBDs and TS Enzymes. Recognizing that similar structures could have arisen fortuitously, we asked whether ligand contact sites are conserved between NR LBDs and TS enzymes. To map contact sites we (i) built on known ligand-contacting residues in both NR LBDs and TS enzymes (RCSB PDB; Methods), and also (ii) performed reciprocal docking studies (computer-based docking simulations; Methods) of a key TS enzyme ligand (FPP, both TS ligand and product, and also a known NR ligand) and representative NR ligands (estradiol, phosphatidylglycerol, and dafachronic acid; Figure 1A) into key TS enzyme and NR structures. Docking was performed on the complete native structures rather than on the core helices.

In all cases where crystal structures were available containing the corresponding ligand, docking accurately reiterated the crystal ligand pose (Table S3). This, combined with further docking studies, revealed that the locations of contact sites in both TS enzymes and NRs are conserved, as mapped to primary, secondary, and tertiary structures of the proteins (Figure 2, also Table S3 and Figure S6). Conservation of contact sites adds weight to the conclusion that TS enzymes and NRs may be evolutionarily related.

Interestingly, this provided evidence for an ancestral internal duplication in TS enzymes. The primary ligand-binding site in TS enzymes is defined by the catalytic DDxxD motif [e.g., (13)] at the c2/c3 junction. However, many TS enzymes (e.g., chicken FPPS) contain a second DDxxD motif at the c6/c7 junction. Analysis indicates that the second site is the relic of an ancestral

duplication (Data S2) that is present even in Archaeal enzymes. In the 3D protein structures the two sites are in close proximity. In some TS enzymes the second site contributes to catalysis, in others it represents an allosteric site that modulates enzyme activity (Data S2). Although the evidence for NRs is much weaker, the same duplication may also be present (Data S2). In NR LBDs site 1 is the principal ligand-binding site, but some large ligands extend beyond site 1 into site 2, with implications for NR pharmacology (Data S2).

In further experiments we docked the FPPS inhibitor, zoledronic acid (ZA), into ESR1 (and also the estrogen-related receptor ERRG). This revealed that ZA binds into the same ligand-binding cavity that is occupied by estradiol, and probably also binds to the estrogen-related receptor ERRG (Data S3).

NRs Are Most Similar to the TS Dehydrodolichyl Pyrophosphate Synthase (DHDPPS) Subfamily. Structural and docking analysis implicates polyisoprene synthases as relatives of NR LBDs. However, structural homologies alone may be misleading, and might be generated by convergent evolution of protein structures adapted to binding structurally similar ligands. We therefore sought confirmation based primary sequence homologies. We argued that NRs at the base of Metazoan radiation would be most informative. In addition to the two receptors aqNR1 and aqNR2 from the sponge *Amphimedon queenslandica* (5), and four receptors from the free-living Placozoan *Trichoplax adherens* (19), we retrieved NR LBD sequences for the stony coral *Orbicella faveolata* (20) (five NR sequences) and the free-living marine Orthonectid *Intoshia linei* (21) (seven sequences) (Methods).

Sequence comparisons using this collection of 18 'early' NRs suggested that NR LBDs are most similar to the *cis*-isoprene transferases (Figure 3A, an extended phylogenetic tree is given in Figure S7). These atypical TS enzymes include DHDPPS and its evolutionarily related obligate partner NUS1, also known as NOGOB receptor (NOGOBR) [(22), see also (23-25)]; indeed, there are potential primary structure similarities between NUS1 and NR LBDs (Figure 3B and Figure S8). Of note, this subgroup is also built around a seven-helix core (26), and the regions of similarity overlap accurately with the conserved seven-helix core identified in NRs (compare Figures S6 and S8); NR LBDs may therefore be evolutionarily related to this specific TS subfamily.

Structural Changes between a TS Enzyme and an NR LBD. Given suggestive evidence that NR LBDs might be related to TS enzymes, we addressed the degree of deformation required to overlap the structures of the two protein groups. Because few detailed structures are available for the 'early' NRs or for the NUS1/NOGOBR/DHDPPS subgroup of TS enzymes, we compared the archetypical TS, chicken FPPS, to human ESR1. This revealed that, although the core helices c4–c6

adopt a similar geometry in the two molecules, the positioning of helices c1–c3 is somewhat different (Figure 4), and rotation of the N-terminal c1–c3 block by 115° around the *z* axis (as defined by the coordinate system) relative to the rest of the protein was necessary to maximize the 3D similarities between the two proteins without other distortion. Similar rotations were necessary for other TS/NR pairs examined (not presented), indicating that this is a general feature of the TS–NR transition.

We then compared primary sequence motifs within key structures. All TS enzymes contain a deep hydrophobic pocket generated by the cluster of α-helices that accommodates the (C5–C30 or more) hydrophobic terpene chain (13, 27) where the catalytic site at the 'mouth' of this pocket comprises the aspartate-rich (DDxxD) motif at the end of helix c2, followed by a flexible loop containing paired arginine (RR) residues (termed here the 'catalytic loop'). These charged residues together generate the primary metal-binding active center of the enzyme.

Comparison of NR LBD and TS 3D structures, assisted by primary/secondary sequence alignment, confirmed that the same primary pocket is also present in NRs, but the DDxxD motif at the mouth has been replaced by a one or more basic residues (Arg, Lys, or Gln) corresponding to Arg397 within the WRS motif in human ESR1 (Figure 2 and Figure S6) – a major contact point for both steroids (e.g., estradiol) and TS ligands (e.g., FPP), as well as for charged residues in other NR ligands such as dafachronic acid (Figure 2 and Figure S6). In the DHDPPS subgroup the DDxxD motif has also been replaced by a basic residue (compare Figures 2 and S8).

We speculate that one of the key primary sequence alterations between typical TS enzymes and NRs is replacement of the DDxxD catalytic motif at the end of helix c2 in TSs by a basic residue in NRs. This may tend to lock the ligand in the pocket. This could have taken place in two steps because some TS-related molecules, notably DHDPPS/NUS1/NOGOBR, contain a basic residue at this position (like NRs), and NUS1/NOGOBR has little enzymatic acticity. Of note, the sponge *Amphimedon queenslandica* has only two NRs (aqNR1 and aqNR2) (5); aqNR2 might represent a hybrid form because it contains both the key arginine and an adjacent DD motif (5) that is reminiscent of the adjacent aspartates in the DDxxD motif of typical TS enzymes (Figure 2 and Figure S6).

Discussion

These data provide an important new perspective on the evolution of nuclear receptors. It was previously conjectured that NR LBDs at the base of the Metazoa had no ligand (28, 29), or that terpenoids (9) or fatty acids (5) were the ancestral ligands. However, our new evidence suggests that the prototypic NR could have been a modified TS enzyme that subsequently acquired a DNA-binding domain, with terpenoid substrate or metabolite assuming the role of activating ligand.

Although we argue that NRs descended from the TS clan of enzymes, a legitimate counterargument could be that any protein with an appropriate number of α-helices would give a positive match using 3D comparison programs (FATCAT and jCE) that allow helices to be rearranged. However, this may not be valid. First, the scores generated take into account the degree of rearrangement required: although structure searching on the RCSB Protein Databank using FATCAT did detect some 'similar' structures, these were rated far lower than TS versus NR comparisons. Furthermore, some TS enzymes were rated to be closer to NR LBDs than to other members of the TS clan. Other enzyme groups built around a cluster of α-helices (e.g., cytochrome P450 enzymes and hydroxysteroid dehydrogenases) also scored low in comparisons with NR LBDs. Indeed, 3D comparisons are often employed to establish relatedness in cases where primary sequence similarities are insufficient [e.g., (30-32)], and we found that separate 3D comparisons for TS enzymes and NR LBDs generated the same phylogenetic trees as those based on primary sequences.

Second, primary through tertiary structure comparisons, combined with crystal data and docking, point to close conservation of contact site locations, which would be unexpected in unrelated proteins. Third, the observation that TS ligands such as FPP have robust interactions with NRs argues in favor of our hypothesis. Fourth, potential primary sequence similarities between 'early' NRs and the DHDPPS group of *cis*-isoprene transferases could argue for an evolutionary relationship, although convergent evolution remains a potential confounding factor in all such comparisons.

Finally, biological plausibility – steroids (the classic NR ligands) are polyterpenes, and it is reasonable to suspect that the protein framework that first evolved to metabolize terpenes could have given rise to a receptor for the same molecules. A different interpretation would require that NRs assembled from an entirely different protein structure, perhaps by chance. However, this would require simultaneous (instead of stepwise) acquisition of a DNA-binding domain; Ockham's razor militates against this alternative hypothesis.

It remains the case that our results do not formally exclude convergent evolution. In addition, we cannot dismiss the possibility that a so far uncharacterized protein family exists in pre-

Metazoa that more closely resembles NRs in form and function. However, in the absence of any supportive evidence, this is a purely theoretical proposition.

As potential confirmation of the TS–NR transition, the question arises of whether some NRs might have residual catalytic activity. We have been unable to address this issue, but it is possible that LBDs of some extant NRs, notably aqNR2, may retain TS-like enzymatic activity. Further experiments will be necessary to address this question.

Interestingly, many NRs, including so-called orphan NRs, are reported to respond to the terpenoid precursor and TS substrate FPP (33). Human ESR1 maintains the key Arg397 binding residue for FPP in the LBD, even though mutation of this residue to inhibit FPP binding did not abolish receptor activation by estradiol (34). Why has it not been lost? We speculate that basal low-level estrogen receptor activation is maintained by FPP in states of physiological estrogen deficiency (e.g., in neonates).

Irrespective of evolutionary implications, the structural similarities we have identified between the terpenoid binding sites in TSs and NRs have pharmacological implications because molecules that target terpene biosynthesis potentially have collateral effects on NRs, and vice versa. A prime example of this is the bone-sparing TS inhibitor ZA that targets FPPS (35-37) and, as shown here, crossreacts with ESR1, potentially explaining a secondary inhibitory effect of ZA on metastatic breast cancer (38, 39). This is an important factor to be considered in developing the next generation of terpenoid-based clinical drugs.

From a broader perspective, this work supports the concept that an enzyme group may have evolved and diversified to become a ligand-modulated transcriptional regulator. There are several examples in bacteria, yeast, and vertebrates where metabolic enzymes evolved to regulate transcription (40-42), but possible gating by substrate/product was not addressed. Indeed, the TS to NR transition could represent the tip of an evolutionary iceberg in which enzymes more generally provided the primary ligand-binding site that subsequently evolved to become a ligand-regulated receptor. Previous analysis suggested that steroid-activated NRs coevolved *with* enzymes involved in ligand synthesis/metabolism [e.g., (43)], whereas our analysis suggests that NRs could have evolved *from* such an enzyme. The distinction is more than semantic, given the centrality of terpene biosynthesis to NR biology.

In conclusion, we report suggestive evidence that TSs and NRs share an evolutionary origin. Our results, based on >10,000 3D structure comparisons, backed up by primary/secondary sequence mapping, reciprocal docking, and primary sequence homologies, argue that NR LBDs are related to a specific TS subfamily of *cis*-isoprene transferases. If confirmed, the emergence of NRs from a subclass of TS enzymes at the base of the Metazoan radiation would reframe the involvement of ancestral terpenoid molecules in morphogenesis, early development, and vertebrate evolution.

Materials and Methods

3D Comparisons. A 64 × 64 matrix of TS enzymes (n = 52) and nuclear receptors (n = 12) was assembled; TS enzymes were selected from the Protein Family (PFAM) Database to represent the diversity of the TS clan of enzymes (https://pfam.xfam.org/clan/Terp_synthase); the enzyme and receptor structures and their access codes are presented in Table S1. Structures were accessed from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Databank (PDB; https://www.rcsb.org/). Visualization of protein structures and 3D overlaps employed PyMol (https://pymol.org/2/). Pairwise comparisons (4068 in each case) were performed in triplicate using the downloaded Linux command-line versions of FATCAT2 (flexible structure alignment by chaining aligned fragment pairs allowing twists) flexible version (FATCATflexible) and FATCATrigid (44), and also Java rigid-body superposition combinatorial extension (jCE) (45). The FATCAT tool is available online at RCSB PDB (http://www.rcsb.org/pdb/workbench/workbench.do) and at the University of California (http://fatcat.godziklab.org). The output consists of (i) a raw score, (iii) a P value, and (iii) RMSD differential between each two structures. Because the raw score takes into account both the degree and the extent of similarity, comparisons were based on this metric. The correspondence between the similarity score and the P value of each alignment is given in supplementary Data S1. In all three cases the results of each comparison were expressed as a similarity score that was typically in the range 0–1000 (96.9% of values; FATCATflexible). The mean scores for NR versus NR (FATCATflexible) was 486, and for TS versus TS was 431 (some TS/TS values scores fell to under 100; Data S1); because outlier (>1000) values were likely to bias the comparison, values >1000 (3.1% of all comparisons) were converted to 1000 before further analysis (Data S1). To determine the extent of distortion required to superimpose a terpene synthase (FPPS) onto a nuclear receptor (ESR1), the CE structural alignment algorithm (45) was used to align the first three helices from the FPPS core onto the first three helices of the ESR core, and the rotation involved in this movement was measured within PyMol.

Tree Drawing. Similarity scores (0–1000) were first converted to distance values (1.0–0.0) by division by 10^3, subtraction of 1.0, and rectification. The pairwise values (a full spreadsheet is given in Table S2) were separately submitted as matrices to PHYLIP (phylogeny inference package) version 3.57c by Joseph Felsenstein (University of Toronto; http://bar.utoronto.ca/webphylip/) under phylogeny methods/distance matrix for analysis by (i) the Fitch–Margoliash method (46); (ii) neighbor joining (47); (iii) UPGMA (unweighted pair group method with arithmetic mean) (48). Trees were drawn using http://bar.utoronto.ca/webphylip/ (plot trees/draw cladograms and phenograms) with output style = phenogram; tree grows = vertically; use branch lengths = yes;

angle = 90°; ancestral nodes = weighted average. The same trees were generated following randomization of the input sequences. All three comparison methods and all three tree-drawing programs generated comparable results (Figures S1–S5), notably with regard to the TS–NR branchpoint, and a 'consensus' tree was assembled that amalgamates all comparisons. Because TS enzymes precede NRs in the evolutionary timeline (TS enzymes are present in Archaea and bacteria, but both taxa lack NRs), the tree is rooted in the weighted average that falls within the TS clan of enzymes. For the radial tree, the mean similarity score was calculated for the three different comparison programs, the tree was computed using Fitch–Margoliash at the University of Toronto, and drawn using iTOL (Interactive Tree of Life) at htpps://itol.embl.de/tree.

Molecular Docking. Ligands dafachronic acid (DAFA), estradiol (E2), farnesyl pyrophosphate (FPP), and phosphatidylglycerol (PG) were docked into the crystal structures of the following protein receptors, excluding the physiological ligand: human estrogen receptor (ESR1; PDB: 1QKU), human liver receptor homolog-1 (LRH1; PDB: 1YOK), *Caenorhabditis elegans* nuclear receptor (DAF-12; PDB: 3GYT), and chicken (*Gallus gallus*) farnesyl pyrophosphosphate synthase (FPPS; PDB: 1FPS), using the program PSOVina2 Ref. (49). Structures were downloaded from PDB, water molecules and other heteroatoms were removed, and the program PDB2PQR 2.1.1 Ref. (50) was used to assign position-optimized hydrogen atoms utilizing the additional PropKa2 algorithm (51) with a pH of 7.4 to predict protonation states. The MGLTools 1.5.6 Ref. (52) utility prepare_receptor4.py was used to assign Gasteiger charges to atoms. Hydrogen atoms were assigned to compound structures using OpenBabel 2.4.1 Ref. (53), utilizing the -p option to predict the protonation states of functional groups at pH 7.4. The MGLTools utility prepare_ligand4.py was used to assign Gasteiger charges and rotatable bonds. PSOVina2 was used to automatically dock the compounds into the crystal structures, and calculate a predicted binding pose and free energy. A grid box that encompassed the maximum dimensions of the ligand plus 12 Å in each direction was used; all other parameters were set to default. PyMol (PyMOL Molecular Graphics System, Version 2.0 Schrödinger LLC) was used to visualize the results.

Primary/Secondary Sequence Alignments. Primary sequences for NR LBDs and TS enzymes were downloaded from the RCSB PDB. The positions of α-helices were manually registered from the PDB structures. For *Amphimedon queenslandica* NR2 no crystal structure is available, and the extents of the α-helices were predicted using three prediction programs: AGADIR (Centro de Regulacióò Genòmica, Barcelona, Spain; http://agadir.crg.es) (54), Jpred4 (University of Dundee, UK; http://www.compbio.dundee.ac.uk/jpred4) (55), and PredictProtein (Technical University of Munich, Germany; https://open.predictprotein.org/) (56). All three programs gave

essentially the same result. NR sequences were aligned using Clustal Omega (57) at the European Bioinformatics Institute (EBI, Cambridge, UK; https://www.ebi.ac.uk/Tools/msa/clustalo/) and COBALT (constraint-based multiple alignment tool) (58) (NCBI; https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi). TS sequences were aligned using the same tools, guided by the detailed alignments presented in the NCBI Conserved Domain Database (CDD; https://www.ncbi.nlm.nih.gov/Structure/cdd/) and in the EMBL-EBI Protein Family (PFAM) database (https://pfam.xfam.org/clan/Terp_synthase). For more distant primary sequence relationships we employed PRRN (progressive pairwise alignment with iterative refinement) (59, 60) (Kyoto University; https://www.genome.jp/tools-bin/prrn); all alignments were checked by secondary structure matching and by 3D matching using the sequence display option of FATCATflexible to reach a consensus.

Primary Sequence Homologies: The DHDPPS Subfamily. We used a large compendium of NR LBD sequences to screen the genomes of species that mark the pre-Metazoa to Metazoa transition, including Chanoflagellida, Ctenophora, Porifera, and Mesozoa (Orthonectida and Rhombozoa). No significant matches were found in Choanoflagellida or in the Ctenophora, but sequences were detected as expected for the Porifora *Amphimedon queenslandica* and the Placozoa *Trichoplax aderens*. This analysis confirmed that aqNR1 is most similar to RXR, whereas aqNR2 is most similar to HNF4 (not presented). In addition, matches were found in the Cnidaria *Orbicalla faveolata* and in the Orthonectida (Rhopaluridae) *Intoshia linei* (not presented). These generated a set of 18 'early' NR sequences that were used to search (tBLASTn at NCBI) for primary sequence homologies to diverse TS enzymes; this identified the DHDPPS subfamily, including its obligate partner proteins NUS1 and NOGOBR, as being particularly closely related. A selection of representative DHDPPS, NUS1, and NOGOBR protein sequences were retrieved from NCBI and compared against the primary sequences of the 18 early NRs using the program PRRN (default settings) as implemented on the GenomeNet website (Tokyo, Japan; https://www.genome.jp/tools-bind/prrn), a center-rooted tree was generated using the UPGMA algorithm, and drawn using iTOL at htpps://itol.embl.de/tree (Figure 3A). To refine the tree, alignment and phylogenetic reconstructions were performed using the function 'build' of ETE3 v3.1.1 (61) implemented on the GenomeNet website (https://www.genome.jp/tools/ete), and the ML tree was inferred using RAxML v8.1.20 using the model PROTGAMMAJTT at the same site and default parameters (62). Branch supports were computed with 100 bootstrap trees (Figure S7). For closer inspection of primary sequence alignments, we first used TCoffee (63, 64) to separately align (i) the sequences of the 'early' NR LBDs, and (ii) the sequences of a selection of NUS1 sequences (NOGOBR was omitted because, although NUS1 and NOGOBR are held to be orthologs, Figure 3A revealed that

they may fall into two distinct subgroups). The separate alignments are presented in Figure S8 (left). The same approach was then used to jointly align the NR LBDs with NUS1 sequences (Figure S8, right). As shown in the figure, both NRs and NUS1 sequences comprise three conserved regions (numbered 1–3), and these same three domains match in the joint NR/NUS1 alignment. An enlargement of domain 1 for both NRs and NUS1 is presented in Figure 3B; correspondence with the conserved α-helical core of NRs is given in Figure S3.

ACKNOWLEDGMENTS. We would like to thank Alan Bateman (Cambridge, UK) for advice on the PFAM clan of terpene synthases, Adam Godzik (California, USA) for making his software widely available, and Cameron Mura (Virginia, USA) for advice on interpreting 3D comparisons.

Competing Interests Statement

The authors declare no competing interests

Additional Information

Supplementary Information is available for this paper.

Data Availability. All data generated or analyzed during this study are included in the article and in the supplementary information files.

References

1. D. López-Escardó *et al.*, Reconstruction of protein domain evolution using single-cell amplified genomes of uncultured choanoflagellates sheds light on the origin of animals. *Philos. Trans. R. Soc. Lond B Biol. Sci.* **374**, 20190088 (2019)

2. E. Zanaria *et al.*, An unusual member of the nuclear hormone receptor superfamily responsible for X-linked adrenal hypoplasia congenita. *Nature* **372**, 635-641 (1994)

3. W. Seol, H. S. Choi, D. D. Moore, An orphan nuclear hormone receptor that lacks a DNA binding domain and heterodimerizes with other receptors. *Science* **272**, 1336-1339 (1996)

4. A. M. Reitzel *et al.*, Nuclear receptors from the ctenophore *Mnemiopsis leidyi* lack a zinc-finger DNA-binding domain: lineage-specific loss or ancestral condition in the emergence of the nuclear receptor superfamily? *Evodevo.* **2**, 3 (2011)

5. J. T. Bridgham *et al.*, Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol.* **8**, e1000497 (2010)

6. G. Holzer, G. V. Markov, V. Laudet, Evolution of nuclear receptors and ligand signaling: toward a soft key-lock model? *Curr. Top. Dev. Biol.* **125**, 1-38 (2017)

7. H. Nakshatri, P. Chambon, The directly repeated RG(G/T)TCA motifs of the rat and mouse cellular retinol-binding protein II genes are promiscuous binding sites for RAR, RXR, HNF-4, and ARP-1 homo- and heterodimers. *J. Biol. Chem.* **269**, 890-902 (1994)

8. R. M. Evans, D. J. Mangelsdorf, Nuclear receptors, RXR, and the big bang. *Cell* **157**, 255-266 (2014)

9. D. D. Moore, Diversity and unity in the nuclear hormone receptors: a terpenoid receptor superfamily. *New. Biol.* **2**, 100-105 (1990)

10. R. E. Summons, A. S. Bradley, L. L. Jahnke, J. R. Waldbauer, Steroids, triterpenoids and molecular oxygen. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **361**, 951-968 (2006)

11. W. D. Nes, Biosynthesis of cholesterol and other sterols. *Chem. Rev.* **111**, 6423-6451 (2011)

12. S. Y. Jiang, J. Jin, R. Sarojam, S. Ramachandran, A comprehensive survey on the terpene synthase gene family provides new insight into its evolutionary patterns. *Genome. Biol. Evol.* **11**, 2078-2098 (2019)

13. D. W. Christianson, Structural and chemical biology of terpenoid cyclases. *Chem. Rev.* **117**, 11570-11648 (2017)

14. S. G. Hillier, R. Lathe, Terpenes, hormones, and life: isoprene rule revisited. *J. Endocrinol.* **242**, R9-R22 (2019)

15. G. Ourisson, Y. Nakatani, The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol. *Chem. Biol.* **1**, 11-23 (1994)

16. J. D. Rudolf, C. Y. Chang, Terpene synthases in disguise: enzymology, structure, and opportunities of non-canonical terpene synthases. *Nat. Prod. Rep.* **37**, 425-463 (2020)

17. A. S. Yang, B. Honig, An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J. Mol. Biol.* **301**, 691-711 (2000)

18. K. Illergård, D. H. Ardell, A. Elofsson, Structure is three to ten times more conserved than sequence - a study of structural response in protein cores. *Proteins* **77**, 499-508 (2009)

19. M. E. Baker, Trichoplax, the simplest known animal, contains an estrogen-related receptor but no estrogen receptor: Implications for estrogen receptor evolution. *Biochem. Biophys. Res. Commun.* **375**, 623-627 (2008)

20. C. Prada *et al.*, Empty niches after extinctions increase population sizes of modern corals. *Curr. Biol.* **26**, 3190-3194 (2016)

21. K. V. Mikhailov *et al.*, The genome of *Intoshia linei* affirms orthonectids as highly simplified spiralians. *Curr. Biol.* **26**, 1768-1774 (2016)

22. K. D. Harrison *et al.*, Nogo-B receptor is necessary for cellular dolichol biosynthesis and protein N-glycosylation. *EMBO J.* **30**, 2490-2500 (2011)

23. K. A. Grabinska, B. H. Edani, E. J. Park, J. R. Kraehling, W. C. Sessa, A conserved C-terminal RXG motif in the NgBR subunit of cis-prenyltransferase is critical for prenyltransferase activity. *J. Biol. Chem.* **292**, 17351-17361 (2017)

24. J. Ma *et al.*, Structural insights to heterodimeric cis-prenyltransferases through yeast dehydrodolichyl diphosphate synthase subunit Nus1. *Biochem. Biophys. Res. Commun.* **515**, 621-626 (2019)

25. B. H. Edani *et al.*, Structural elucidation of the cis-prenyltransferase NgBR/DHDDS complex reveals insights in regulation of protein glycosylation. *Proc. Natl. Acad. Sci. U. S. A* **117**, 20794-20802 (2020)

26. M. L. Bar-El *et al.*, Structural basis of heterotetrameric assembly and disease mutations in the human cis-prenyltransferase complex. *Nat. Commun.* **11**, 5273 (2020)

27. M. Fujihashi *et al.*, Crystal structure and functional analysis of large-terpene synthases belonging to a newly found subclass. *Chem. Sci.* **9**, 3754-3758 (2018)

28. H. Escriva *et al.*, Ligand binding was acquired during evolution of nuclear receptors. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 6803-6808 (1997)

29. H. Escriva, F. Delaunay, V. Laudet, Ligand binding and nuclear receptor evolution. *Bioessays* **22**, 717-727 (2000)

30. H. Jiang, C. Blouin, Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. *BMC Bioinformatics* **8**, 444 (2007)

31. V. Modi, R. L. Dunbrack, Jr., A structurally-validated multiple sequence alignment of 497 human protein kinase domains. *Sci. Rep.* **9**, 19790 (2019)

32. A. L. Leitão, F. J. Enguita, Structural insights into carboxylic polyester-degrading enzymes and their functional depolymerizing neighbors. *Int. J. Mol. Sci.* **22**, (2021)

33. S. Das *et al.*, Farnesyl pyrophosphate is a novel transcriptional activator for a subset of nuclear hormone receptors. *Mol. Endocrinol.* **21**, 2672-2686 (2007)

34. R. Goyanka, S. Das, H. H. Samuels, T. Cardozo, Nuclear receptor engineering based on novel structure activity relationships revealed by farnesyl pyrophosphate. *Protein. Eng. Des. Sel.* **23**, 809-815 (2010)

35. J. R. Center, K. W. Lyles, D. Bliuc, Bisphosphonates and lifespan. *Bone.* **141**, 115566 (2020)

36. R. T. Guo *et al.*, Bisphosphonates target multiple sites in both cis- and trans-prenyltransferases. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 10022-10027 (2007)

37. M. K. Tsoumpra *et al.*, The inhibition of human farnesyl pyrophosphate synthase by nitrogen-containing bisphosphonates. Elucidating the role of active site threonine 201 and tyrosine 204 residues using enzyme mutants. *Bone.* **81**, 478-486 (2015)

38. M. C. Winter, I. Holen, R. E. Coleman, Exploring the anti-tumour activity of bisphosphonates in early breast cancer. *Cancer Treat. Rev.* **34**, 453-475 (2008)

39. C. N. George *et al.*, Oestrogen and zoledronic acid driven changes to the bone and immune environments: potential mechanisms underlying the differential anti-tumour effects of zoledronic acid in pre- and post-menopausal conditions. *J. Bone Oncol.* **25**, 100317 (2020)

40. B. A. Citron *et al.*, Identity of 4a-carbinolamine dehydratase, a component of the phenylalanine hydroxylation system, and DCoH, a transregulator of homeodomain proteins. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 11891-11894 (1992)

41. D. A. Hall *et al.*, Regulation of gene expression by a metabolic enzyme. *Science* **306**, 482-484 (2004)

42. E. Levati, S. Sartini, A. Bolchi, S. Ottonello, B. Montanini, Moonlighting transcriptional activation function of a fungal sulfur metabolism enzyme. *Sci. Rep.* **6**, 25165 (2016)

43. M. E. Baker, Origin and diversification of steroids: co-evolution of enzymes and nuclear receptors. *Mol. Cell Endocrinol.* **334**, 14-20 (2011)

44. Y. Ye, A. Godzik, Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics.* **19 Suppl 2**, ii246-ii255 (2003)

45. I. N. Shindyalov, P. E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein. Eng.* **11**, 739-747 (1998)

46. W. M. Fitch, E. Margoliash, Construction of phylogenetic trees. *Science* **155**, 279-284 (1967)

47. N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987)

48. R. R. Sokal, S. D. Michener, A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **38**, 1409-1438 (1958)

49. H. K. Tai, S. A. Jusoh, S. W. I. Siu, Chaos-embedded particle swarm optimization approach for protein-ligand docking and virtual screening. *J. Cheminform.* **10**, 62 (2018)

50. T. J. Dolinsky *et al.*, PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **35**, W522-W525 (2007)

51. H. Li, A. D. Robertson, J. H. Jensen, Very fast empirical prediction and rationalization of protein pKa values. *Proteins.* **61**, 704-721 (2005)

52. G. M. Morris *et al.*, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785-2791 (2009)

53. N. M. O'Boyle *et al.*, Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 33 (2011)

54. V. Muñoz, L. Serrano, Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers*. **41**, 495-509 (1997)

55. A. Drozdetskiy, C. Cole, J. Procter, G. J. Barton, JPred4: a protein secondary structure prediction server. *Nucleic. Acids. Res.* **43**, W389-W394 (2015)

56. G. Yachdav *et al.*, PredictProtein - an open resource for online prediction of protein structural and functional features. *Nucleic. Acids. Res.* **42**, W337-W343 (2014)

57. F. Sievers *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011)

58. J. S. Papadopoulos, R. Agarwala, COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics.* **23**, 1073-1079 (2007)

59. O. Gotoh, An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705-708 (1982)

60. O. Gotoh, Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **264**, 823-838 (1996)

61. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635-1638 (2016)

62. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014)

63. C. Notredame, D. G. Higgins, J. Heringa, T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205-217 (2000)

64. E. Garriga *et al.*, Multiple sequence alignment computation Using the T-Coffee regressive algorithm implementation. *Methods Mol. Biol.* **2231**, 89-97 (2021)

65. G. V. Markov *et al.*, Origin of an ancient hormone/receptor couple revealed by resurrection of an ancestral estrogen. *Sci. Adv.* **3**, e1601778 (2017)

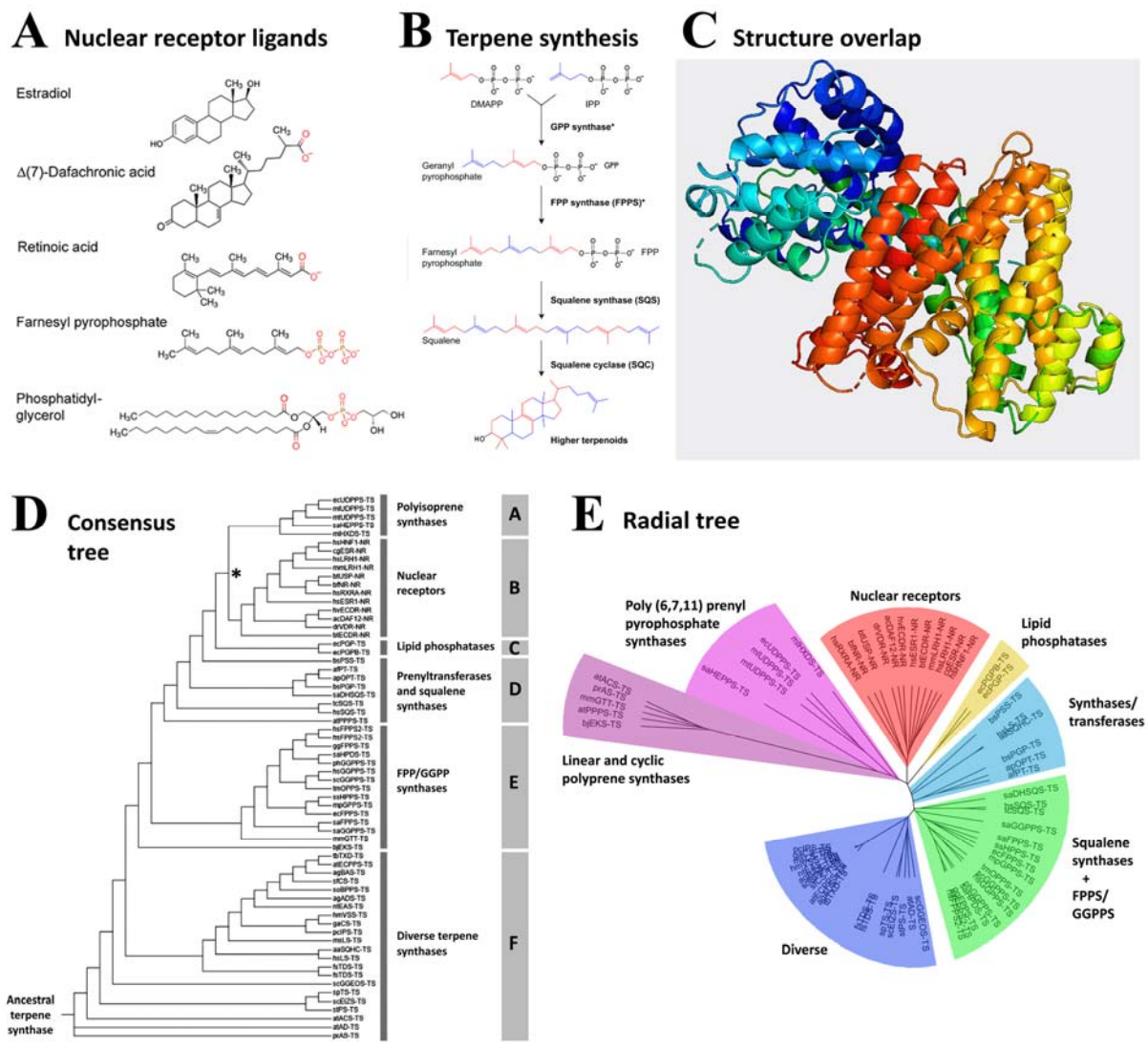

Fig. 1. The ligand-binding domains (LBDs) of nuclear receptors (NRs) are similar in 3D structure to terpene synthase (TS) enzymes. (A) Representative ligands for NRs; terminal charged groups are shown in red. (B) The terpene synthase pathway; the asterisk indicates that a single enzyme can catalyze the synthesis of both geranyl pyrophosphate (GPP) and farnesyl pyrophosphate (FPP) from DMAP (dimethylallyl diphosphate) and its isomer, IPP (isopentenyl pyrophosphate). (C) Example 3D overlap of TS monoterpene synthase from the Greek sage plant, *Salvia fruticosa* (PDB 2J5C, ~600 residues) with the ligand-binding domain of *Mus musculans* NR LRH1 (PDB 1PK5, ~240 residues). The overlap was performed using FATCATflexible (Methods), and the combined 3D overlap was imaged using the 'chainbow' option of PyMol in which similar helices are colored sequentially in both proteins according to the visual spectrum. (D) Consensus phylogenetic tree for nuclear receptors (NRs) and terpene synthase (TS) enzymes established from a 64 × 64 matrix analyzed using three different 3D structure comparison programs. The NR and TS structures analyzed are listed in Table S1. The tree is midpoint-rooted, that falls within the TS group; branch lengths are for illustration only. The different polypeptide groups are designated A–F to facilitate comparison with trees drawn using different programs (Figures S1–S5). (E) Radial tree with branch lengths. This midpoint-rooted tree is based on the mean similarity scores obtained with three comparison programs but differs from the consensus tree in (D) because it was computed using a single algorithm (Fitch–Margolaish). Both (D) and (E) locate NRs between polyisoprene synthases and lipid phosphatases.

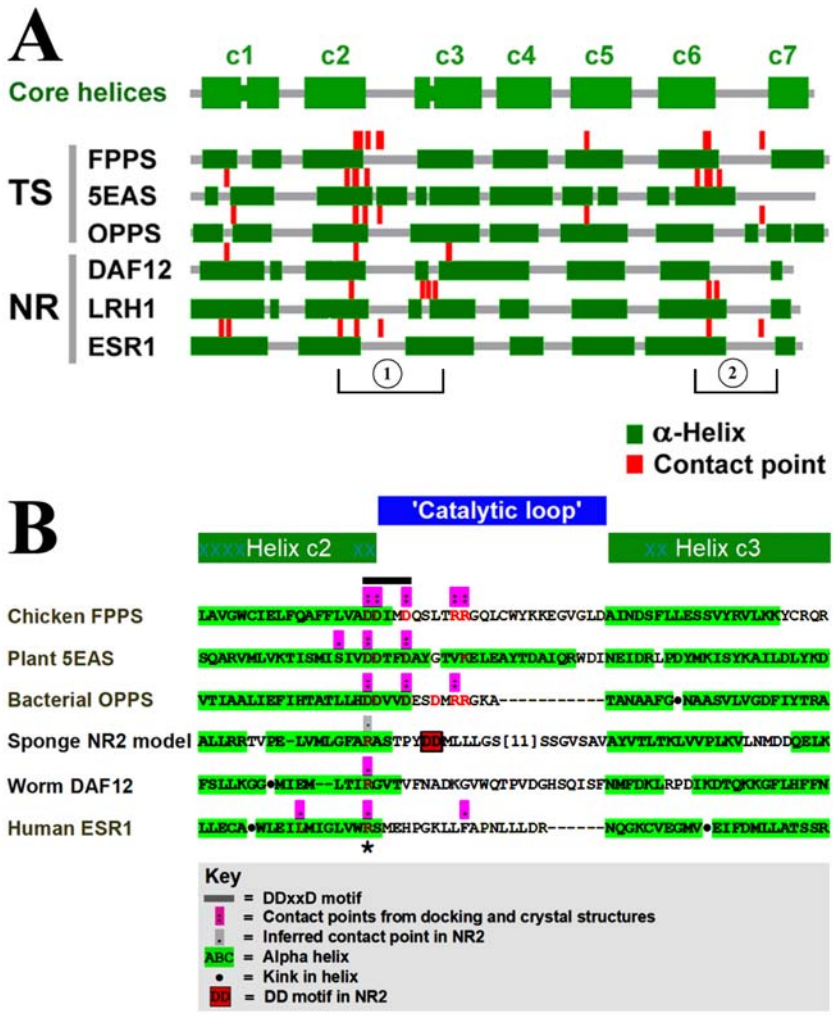

Fig. 2. Alignment of terpene synthases (TSs) and nuclear receptors (NRs) showing conservation of α-helices and clustering of ligand contact points in two subregions (1 and 2). (A) Overall alignment of core α-helices c1–c7 between NRs and TSs (the full version of the alignment is given in Figure S6), showing conservation of the α-helical framework and ligand contact points between the two groups. Core helices c1–c7 correspond to NR helices H3, H4/5, H6/H7, H8, H9, H10/11, and H12 as defined by Markov *et al.* (65), noting that the precise extents of the α-helical segments differ between different crystal structures of the same protein. (B) Detailed map of the c2–c3 junction (encompassing region 1 in A) highlighting replacement of the DDxxD motif in TS enzymes by a single arginine residue (asterisk) in NRs. Protein structures depicted are DAF12, *Strongyloides stercoralis* (nematode) DAF-12 nuclear receptor (PDB 3GYT); 5EAS, *Nicotiniana tabacum* (tobacco) 5-epi-aristolochene synthase (PDB 5EAS and 5EAT); ESR1, human estrogen receptor α (PDB 1QKU, 2OCF, other); FPPS, *Gallus gallus* (chicken) farnesyl pyrophosphate synthase (PDB 1FPS); NR2, *Amphimedon queenslandica* (sponge) NR2 model (see Methods); LRH1, human liver receptor homolog 1 (PDB 1YOK); OPPS, *Escherichia coli* octaprenyl pyrophosphate synthase (PDB 3WJN).

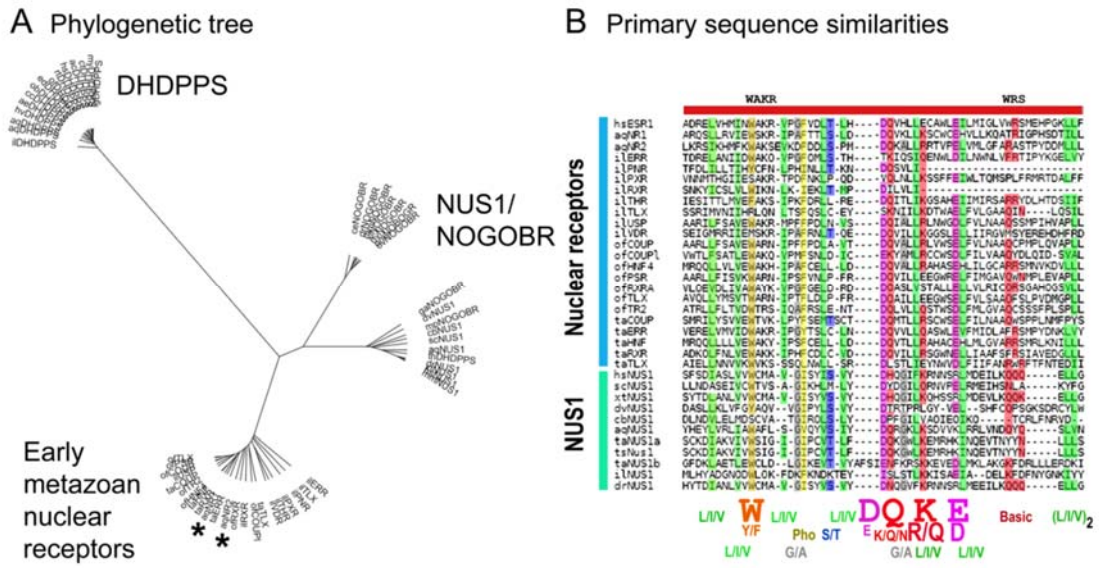

Fig. 3. Ligand-binding domains (LBDs) of 'early' nuclear receptors (NRs) are most similar to the dehydrodolichyl pyrophosphate synthase (DHDPPS)/NUS1/NOGOB receptor (NOGOBR) subfamily of terpene synthase enzymes. A group of early NRs was compiled from the sponge *Amphimedon queenslandica* (aq), Placozoan *Trichoplax adherens* (ta), stony coral *Orbicella faveolata* (of), and the marine Orthonectid *Intoshia linei* (of). Homology searching revealed that these are similar to the DHDPPS/NUS1/NOGOB receptor (NOGOBR) subfamily; panel (A) shows a midpoint-rooted phylogenetic tree constructed using PRRN (Kyoto University Bioinformatics Center, Kyoto, Japan) and the UPGMA algorithm. Branch lengths are to scale, indicating that the early NRs are similar to both NUS1/NOGOBR and their binding partner DHDPPS that are known to be evolutionarily related. Asterisks indicate the positions of aqNR1 and aqNR2. A detailed tree is presented in Figure S7. (B) Primary sequence homologies detected by TCoffee multiple sequence alignment (MSI) program M-Coffee (Center for Genomic Regulation, Barcelona, Spain) that combines the results of different alignment programs. The human estrogen receptor hsESR1 has been added for reference; WAKR and WRS represent conserved motifs in ESR1, where WRS is a primary ligand contact site ligand. An extended alignment is given in Figure S8.

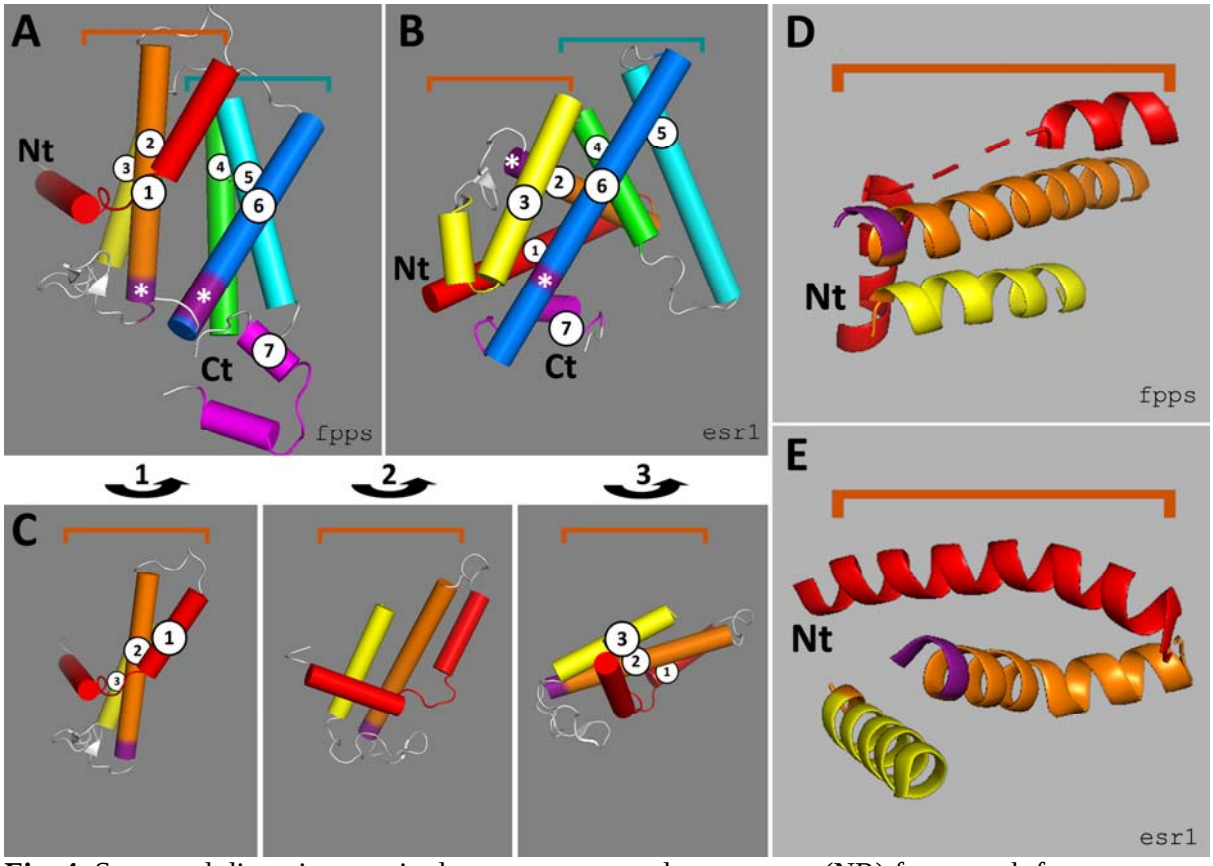

Fig. 4. Structural distortion required to generate a nuclear receptor (NR) framework from a terpene synthase (TS) framework. (A–C) The native core helices c1–c7 of *Gallus gallus* (chicken) farnesyl pyrophosphosphate synthase (FPPS, panel A) and human estrogen receptor alpha (ESR1, panel B) (structure details are given in Table S1) colored sequentially from the N-terminus (Nt) to the C-terminus (Ct) red, orange, yellow, green, cyan, blue, magenta. Major contact points (asterisks) are colored deep violet. Helices c4–6 (green/blue) adopt a similar configuration in FPPS and ESR1 (A and B), whereas TS helices c1–c3 (red/orange/yellow) require (i) rotation through 115° (C) to generate a 3D structure similar to that of NRs (D and E); in addition, (ii) the c1 helix that is 'broken' in FPPS (D) is contiguous in ESR1 (E). Rotation of the TS N-terminus may be essential to accommodate (or may have been generated by) fusion to the DNA-binding domain (that is held to have been independently acquired).