

Phylogenomic Analysis Reveals Extensive Phylogenetic Mosaicism in the Human GPCR Superfamily

Robin G. Allaby¹ and Mathew Woodwark²

¹Warwick HRI, University of Warwick, Wellesbourne, CV35 9EF, UK. ²Cambridge Antibody Technology Ltd., Milstein Building, Granta Park, Cambridge CB1 6GH, UK.

Abstract: A novel high throughput phylogenomic analysis (HTP) was applied to the rhodopsin G-protein coupled receptor (GPCR) family. Instances of phylogenetic mosaicism between receptors were found to be frequent, often as instances of correlated mosaicism and repeated mosaicism. A null data set was constructed with the same phylogenetic topology as the rhodopsin GPCRs. Comparison of the two data sets revealed that mosaicism was found in GPCRs in a higher frequency than would be expected by homoplasy or the effects of topology alone. Various evolutionary models of differential conservation, recombination and homoplasy are explored which could result in the patterns observed in this analysis. We find that the results are most consistent with frequent recombination events. A complex evolutionary history is illustrated in which it is likely frequent recombination has endowed GPCRs with new functions. The pattern of mosaicism is shown to be informative for functional prediction for orphan receptors. HTP analysis is complementary to conventional phylogenomic analyses revealing mosaicism that would not otherwise have been detectable through conventional phylogenetics.

Keywords: evolution/GPCRs/mosaicism/phylogenomics

Introduction

The G-protein coupled receptors (GPCRs) are a remarkably diverse superfamily of seven-transmembrane proteins that are responsible for intracellular G-protein activation on recognition of an extracellular ligand. A wide array of stimuli are recognised by GPCRs including light, odorants, calcium ions, nucleotides, lipids and peptides. The superfamily can be separated into at least five families that recognise diverse ligands (Bockaert and Pin, 1999; Fredriksson et al. 2003). Fredriksson et al. 2003 define these family groups as the Rhodopsin, Adhesion, Secretin, Glutamate and Frizzled receptor groups. These major divisions of the superfamily show little or no sequence similarity making phylogenetic inference about the origin of the GPCR superfamily problematic. The nature of the functionality of GPCRs makes them attractive targets for drug design; about half of all modern drugs act on GPCRs (Flower, 1999; Howard et al. 2001). Frequently, receptors that have not had their ligands characterised, so called orphan receptors, have their functionality inferred through phylogenetics. Consequently, an understanding of how GPCRs are related is highly desirable in the course of drug design. Each major family of GPCRs is itself made up of a number of groups. The phylogenetic relationships within these groups appear to be relatively clear, but it remains a challenge to ascertain how the different GPCR groups relate to each other within a family (Joost and Methner, 2002). Even within the rhodopsin family sequence conservation is too poor for good multiple sequence alignments across the whole family. To achieve robust phylogenetic analyses it is necessary to separate out the GPCR groups as relatively small but high quality alignments (Sjölander, 2004).

It is evident from the GPCR superfamily as a whole that sequence conservation is low, but a core tertiary structure is maintained. Consequently, any functional sequence conservation that does occur between sequences is likely to be phylogenetically local and possibly differential over the sequence length. It is possible then that sequence similarity high enough between GPCR groups to be identified as homology may only be fragmentary. The fragmentary nature of homology between sequences is exacerbated by the possibility that many GPCRs may be the result of recombination processes (Shields, 2000). This feature echoes findings that the majority of transmembrane and extracellular proteins are thought to be modular (Patthy, 1999), and the human genome appears to have a high number of modular

Correspondence: Robin G. Allaby, Fax: +44 (0) 2476574500; Email: r.g.allaby@warwick.ac.uk

Please note that this article may not be used for commercial purposes. For further information please refer to the copyright statement at <http://www.la-press.com/copyright.htm>

genes (Li et al. 2001). Consequently, a biologically realistic strategy to reconstruct the phylogeny of the GPCRs should begin with the reconstruction of the phylogenies of those sequence fragments within GPCRs that can be identified as being homologous to other sequences.

It was the aim of this study to establish homology and reconstruct the phylogenies of sequence fragments within GPCR sequences across the whole rhodopsin family, and compare those to the phylogenies of the whole sequences within a GPCR group. In this way we hoped to extend the homologous relationships between GPCR groups on a fine scale and identify instances of potential recombination events through changes to expected phylogenetic topologies. In order to be able to identify relationships between the different GPCRs we elected to study sequence data from a single organism, humans. In this way more distant relationships will not be masked by close orthologous relationships. The experimental strategy was to construct phylogenetic trees for each GPCR group alignment, and then search for instances in which fragments of sequences were similar between GPCR groups and could be utilised in phylogenetic reconstruction. However, this approach rapidly leads to a high number of phylogenetic analyses all of which require careful interpretation of results. Consequently, we developed a novel high throughput phylogenomic (HTP) approach in which both phylogenetic tree construction and subsequent interpretation are automated, outlined in Figure 1. The analysis is phylogenomic because it uses all the known rhodopsin type GPCRs in the human genome. The robustness of phylogenetic interpretation is incumbent on underlying assumptions and validity of the input data. While the former is dependent upon the phylogenetic algorithm applied and was not further vetted in this study, the latter was subject to an automated quality control procedure that enabled a minimum quality of data input to be met for automated phylogenetic analyses. The stringent screening of data input and removal of data that is too distant made the novel large-scale automation of phylogenetic interpretation possible. Automated phylogenetic interpretation was limited to the identification of nearest neighbors. This automated approach enabled a powerful and comprehensive analysis of the rhodopsin GPCRs giving a novel perspective of the reticulate relationships across the family.

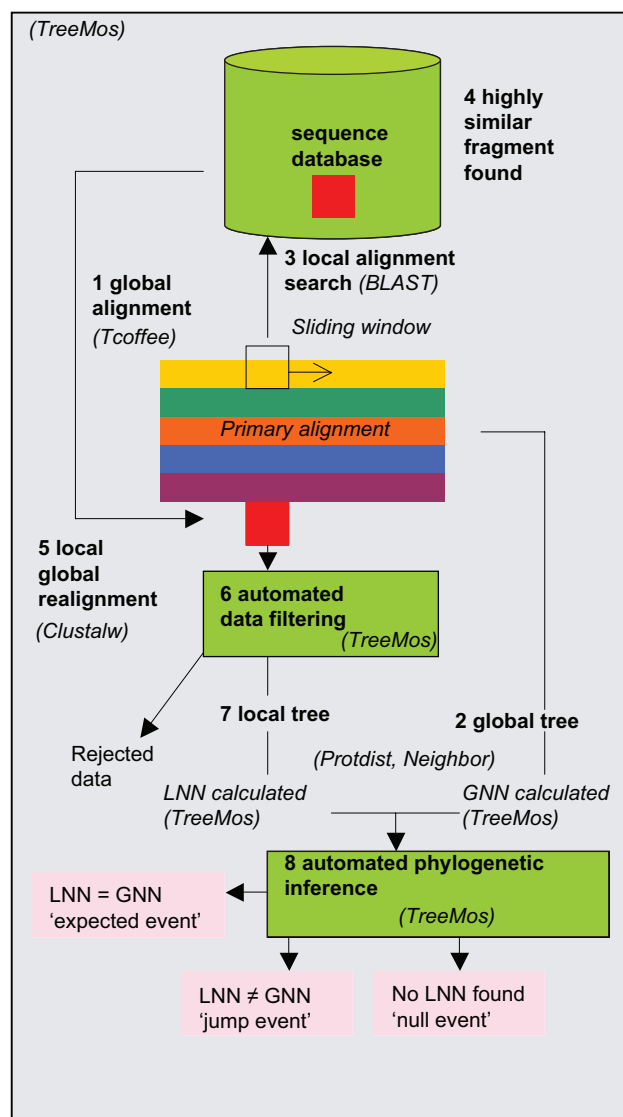


Figure 1. Schematic overview of HTP analysis. The analysis begins with a global alignment (1) from which a global tree is calculated (2). A sliding window is then moved down the alignment. The sequence within each window is subject to a BLAST search against a sequence database (3). If a sequence fragment is found which with a very low genetic distance (4), then this sequence fragment is globally aligned to the window (5). The window alignment is screened and filtered for high distances (6), and a local tree calculated based on the remaining data (7). The nearest neighbour to each taxon is then compared between the local tree and the global tree (8). The possible event outcomes of the analysis are shown in pink boxes. The programs employed at each stage of the analysis are shown in brackets.

Results and Discussion

We assembled 257 sequences of human rhodopsin type GPCRs into 19 alignments based on the groupings of Joost and Methner (2002), termed group 1–group 19. In order to improve alignments to meet our own minimum quality requirement criteria, we further subdivided the 19 alignments into 33 alignments such that each group did not

Table 1. GPCR groups of the rhodopsin family used in this study. Adapted from Joost and Methner (2002).

GPCR group	Receptor ligand types
group 1	CC chemokines
group 2a	CXC chemokines
group 2b	duffy
group 3	angiotensin, apelin, bradykinin
group 4	opioids, neuropeptide B, nociceptin, somatostatin
group 5	galanin, melanin-concentrating hormone, kisspeptin, leukotriene-B4, somatostatin, urotensin
group 6a	cholecystokinin, gastrin, neuropeptide FF, orexin
group 6b	gonadotrophin-releasing hormone, vasopressin
group 6c	orphan
group 7a	bombesin, endolethin, gastrin-releasing peptide, neuromedin B
group 7b	growth hormone secretagogue, neuromedin U, neurotensin, motilin, thyrotropin-releasing hormone
group 8a	anaphylatoxin, lipoxin A4, N-formyl peptide
group 8b	mas
group 9a	substance-P, substance K, neuromedin K
group 9b	prolactin-releasing peptide, melatonin, neuropeptide Y, prokineticin
group 10	folliotropin, thyrotropin, thyrotropin-releasing hormone, lutropin-choriogonadotropic
group 11	lipids, proprionate, short chain fatty acids, nicotinic acid, ATP
group 12	UDP-glucose, ADP, platelet activating factor
group 13	adrenocorticotrophic hormone, cannabinoid, sphingosine 1-phosphate, lysophosphatidic acid, sphingolipid, melanotropin
group 14	prostaglandin, prostacyclin, thromboxane
group 15	lysophosphatidylcholine, psychosine, thrombin, proteinase
group 16	opsins, peropsin
group 17	5-hydroxytryptamine, adrenergic, dopamine, histamine H2
group 18a	muscarinic acetylcholine, histamine, H2
group 18b	adenosine
groups18c-i	orphan
group 19	5-hydroxytryptamine

yield a distance between any two taxa greater than the parameter κ (see methods), Table 1 (see also supplementary alignments). Alignment groups that had become subdivided to accommodate κ kept the same group names with a single letter suffix to denote the subdivided group. Consequently, each GPCR alignment contains closely related sequences for which a robust phylogenetic tree could be calculated. To interpret and test the significance of our HTP analysis we required a rhodopsin GPCR tree. A phylogenetic methodology of ancestor reconstruction was used to reconstruct the rhodopsin tree resulting in six major clades being identified, Figure 2 (see supplementary data for the complete treefile). The six major GPCR clades identified in this study represent the distance down the tree from the tips the ancestor reconstruction analysis could be carried out before the distance between ancestors exceeded our value for κ used in the HTP analysis. The topology of the tree is in

broad agreement with that of Fredriksson et al. (2003). Both our tree and that of Fredriksson differ from the Joost and Methner (2002) organization by considering group 18 paraphyletic.

HTP analyses of real GPCRs and the null data set

The HTP analysis of GPCRs produced 6041 phylogenetic outcomes representing a maximum of 18,123 phylogenetic analyses. Phylogenetic mosaicism occurred frequently, data indicating 46% of all phylogenetic outcomes found jump events for taxa in all the GPCR groups. Furthermore, 18% of all events were jump events between different GPCR groups, Figure 3, Supplementary Figure 1. As an example, the opioid receptor is shown in more detail in Figure 4.

To help identify the underlying reasons for the phylogenetic mosaicism we compared the real data

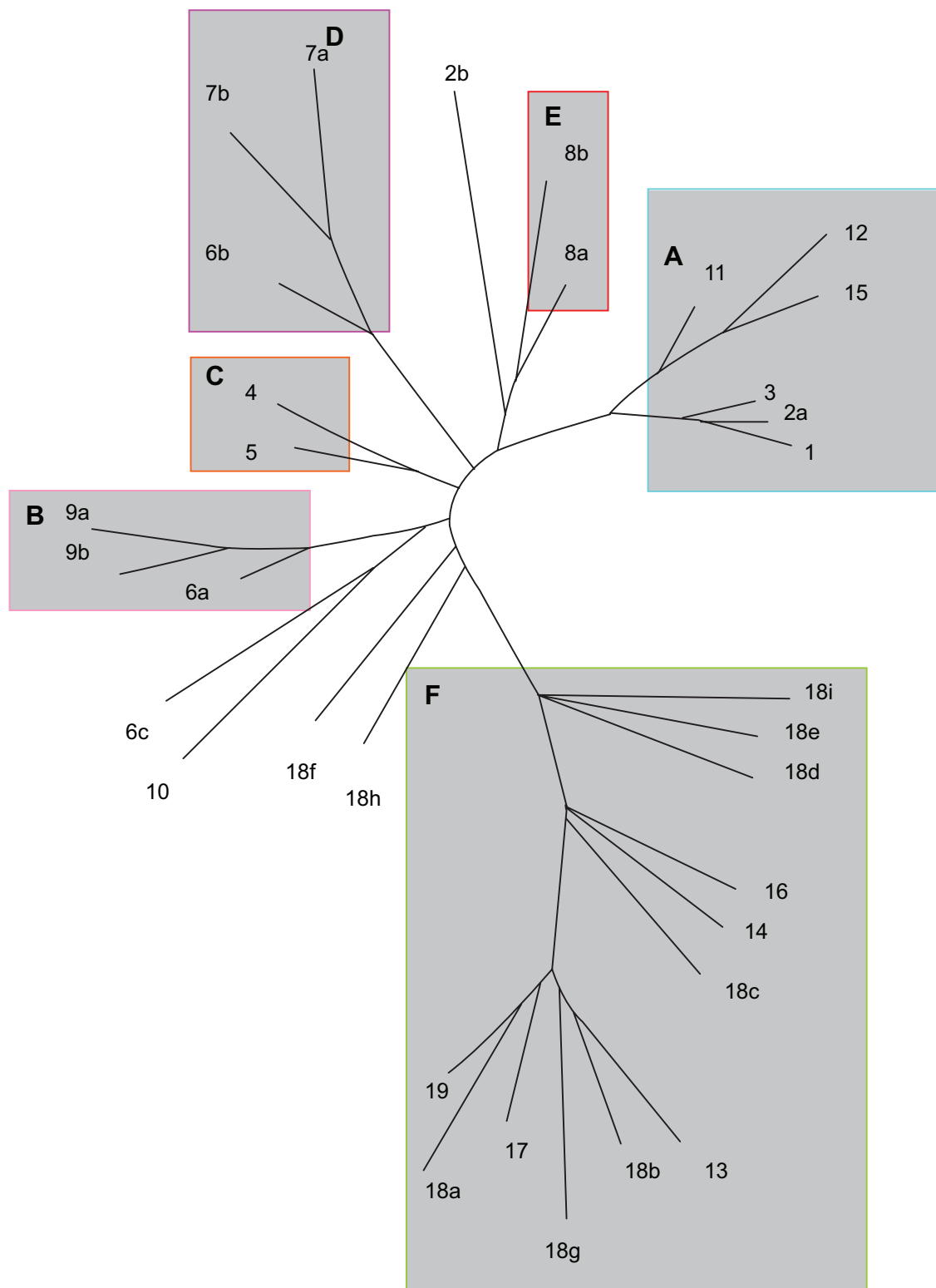


Figure 2. Phylogeny of rhodopsin GPCR groups. GPCR phylogeny based on reconstructed ancestors. Six principal clades are resolved within which genetic distances between reconstructed ancestors and extant sequences do not exceed k . The principal clades are labeled **A–F** as follows. **A.** Nucleotide receptor group: includes peptide, nucleotides and lipid receptors. **B.** The endocrine hormone receptor group: includes peptide receptors involved with endocrine hormones. **C.** The opioid receptor group: includes opioid, somatostatin and galanin related peptide receptors. **D.** The neuropeptide receptor group: includes peptide receptors involved with endocrine hormones and neuropeptides. **E.** The mas receptor group: includes nociceptive mas and mas related receptors. **F.** The amine receptor group: includes biogenic amine, rhodopsin, arachidonic and peptide and lipid receptors. The remaining ancestors to groups are not included in major families because the genetic distance between them and all other taxa is aberrantly high.

with a simulated null hypothesis GPCR data set in which no homogenisation through recombination had occurred. The null data set was generated by simulating mutations to a random seeded sequence to represent the evolutionary pattern of divergence illustrated in the phylogenetic tree for GPCRs in Figure 2, producing 257 new sequences. A HTP analysis was carried out on the null data set resulting in 4590 phylogenetic outcomes reflecting a maximum of 13,770 phylogenetic analyses (data not shown). A level of phylogenetic mosaicism was apparent in this second analysis which is similar to that found with the real data set, with 52% percent of all phylogenetic outcomes being jump events, and 16% of all outcomes being jump events between groups. However, only 2% of phylogenetic outcomes were null events, where no neighbour could be identified, considerably less than the 12% observed with the real data. Differences between the two data sets are further highlighted by paired T-tests for the relative frequencies for each of the nineteen possible phylogenetic outcomes that showed a significant difference in all but two cases (Supplementary Table 1). Mosaicism between families, represented by only one outcome type, did not differ significantly between the real and simulated data sets. However this apparent similarity is strongly reduced by a lurking variable regarding the phylogenetic distance of jumps. The phylogenetic distance represented by jump events did differ significantly between the two data sets (Supplementary Table 2). There is a threefold increase in the frequency of jumps over just 1–4 nodes in the real data, significant at the 0.01 level. Consequently, there are many more jumps between closely related GPCR groups than would be expected from homoplasy. Conversely, results obtained with the real and simulated data sets for jumps of larger phylogenetic distance were similar. This indicates that the resemblance of sequence fragments between distantly related GPCR groups is frequently attributable to homoplasy. It also implies that functional constraint on sequence change is negligible across large distances in the family because differentially conserved sequences in different groups would be expected to lead to a greater number of instances of mosaicism at this level.

Patterns of mosaicism and possible underlying evolutionary models

Several interesting patterns are discernible in the phylogenetic mosaicism between GPCR groups

shown in Figure 3. Firstly, there is a notable tendency for a higher number of mosaic events between groups of the same major groupings, as defined in Figure 2. Secondly, the similarity between fragments associated in an instance of phylogenetic mosaicism can be extremely high. Thirdly, mosaicism appears to be frequently correlated for members of the same family. Fourthly, similar mosaicism is frequently repeated in individual sequences.

The high similarity that can occur between fragments is shown in the example of phylogenetic mosaicism between the gastrin receptor group and the biogenic amine receptors, Figure 5. In this case the HTP analysis detected a region of phylogenetic inconsistency in the first cytosolic region between transmembranes 1 and 2 of the orexin receptors and the alpha-1B adrenergic receptor. This region is generally considered to be among the most conserved in GPCRs, but within both the biogenic amine and gastrin receptors group the fragment is highly variable in sequence composition. The similarity between the subset of sequences from these two distant groups is striking and highly significant statistically—a local alignment search would only expect to find 0.0008 sequences this length and similarity (E value = 0.0008). This level of probability makes homoplasy as an explanation of the sequence similarity in this case highly doubtful. One explanation might be that this segment was strongly conserved in these two disparate lineages from antiquity because of an important function. Such an explanation would predict that both the alpha 1 adrenergic and orexin receptors were placed at a basal position in their respective phylogenetic groups, and one would also expect to see the conserved sequence in the respective ancestors. The orexin receptor ancestor does indeed have a sequence very similar to that shared by the two disparate lineages, although they do still share a character state to the exclusion of the ancestor. This supports the idea that this particular region of sequence has been under considerable functional constraint in the gastrin receptors. However, the alpha 1 adrenergic receptors are not basal to their phylogenetic group (see GPCR tree, supplementary data), and the reconstructed ancestor for this group does not resemble the sequence shared by the two disparate lineages. The phylogenetic data are incompatible with the idea that the segment could have been conserved in the form seen in the alpha 1

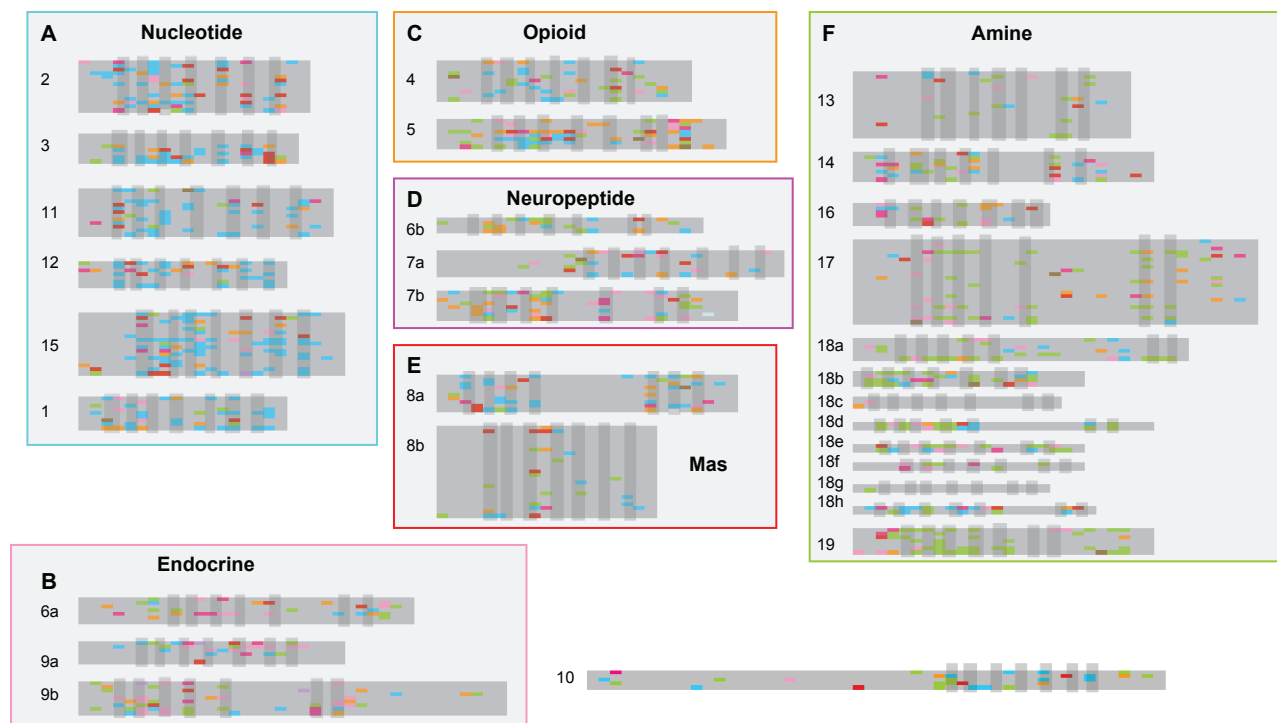


Figure 3. Mosaic relationships of the rhodopsin GPCR groups. Instances of phylogenetic mosaicism between GPCR groups are displayed. Colour signifies the major group from which the mosaic fragment derives: blue, nucleotide receptor group (A); pink, endocrine receptor group (B); orange, opioid receptor group (C); purple, neuropeptide receptor group (D); red, mas receptor group (E); green amine receptor group (F). Letters A–F correspond to major groups in Fig 2. GPCR groups not associated with a major group are colour coded as follows: group 1, white; group 2b pale blue; group 6c, violet; group 10 brown. Transmembrane domains are indicated in grey over the alignments. For details of specific receptor affiliations see Supplementary Figure 1.

adrenergic receptor lineage. A more likely explanation is that the functionally conserved segment from the orexin receptor or its ancestor was donated to the alpha 1 adrenergic lineage some time after the gene duplication event that produced the alpha 1 and alpha 2 lineages, but before the diversification of the alpha 1 lineages. Immediately after the duplication event that split the alpha 1 and 2 lineages, it is likely that functional constraint would have been relaxed on either one of the lineages while the other continued the original function. Therefore the evolutionary opportunity occurred in which the alpha 1 lineage could receive new sequence through recombination without a costly loss of function to the host organism. It is particularly interesting then that the sequence fragment in question appears to be one that is functionally important implying that this may have been an important mechanism in GPCR evolution to produce GPCRs with new function.

Another pattern is that of correlated mosaicism. Correlated mosaicism is the condition in which multiple members of a group have the same or

similar instances of mosaicism. Striking examples of this occur for group 17 in which many members of the group display phylogenetic mosaicism relating to the endocrine receptor group, in particular the orexin receptor, but also receptors of group 6a. In several other members of the same group there is phylogenetic mosaicism with somatostatin receptor 3 (group 4) in the latter part of the alignment. Other instances of correlated mosaicism occur throughout the rhodopsin GPCRs, such as a similarity of multiple Mas type receptors fragments to opioid receptor D in the latter half of the alignment; several members of the angiotensin receptor group to the Mas group in the latter part of the alignment. The opioid receptor group in Figure 4 shows correlated mosaicism to the Mas receptors indicating that there is a reciprocal relationship between the Mas type receptors of group 8a and the opioid receptor D. In the same region of the opioid group there is correlated mosaicism between somatostatin receptors three and five, and the beta-2 adrenergic receptor of group 17. A model in which segments of GPCR receptors are differentially conserved in different families could not

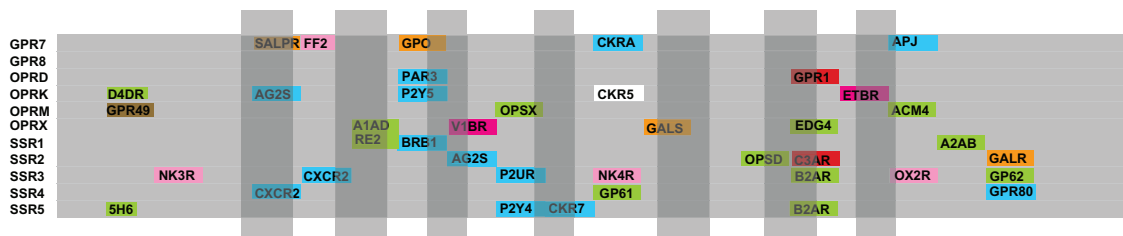


Figure 4. Mosaicism in opioid receptor group (group 4). Segmental phylogenetic affiliations resulting from HTP. Segments that have phylogenetic affiliations to receptors in GPCR groups other than group 4 are shown as colour blocks with the affiliated receptor name. Block colour corresponds to major GPCR group, shown in Figure 3. Transmembrane domains are indicated in grey over the alignment. For further details of affiliated fragment lengths see Supplementary Figure 1.

explain correlated mosaicism. Members within a family would be expected to appear more closely related to each other than outside taxa because they would be derived from a conserved common ancestor. Four phylogenetic models could be forwarded to explain the non-random pattern of correlated mosaicism, outlined in Figure 6. In the first model (Fig. 6A), two or more group members that display mosaicism to an ‘outside’ member from a different group may have had high rates of change in the past resulting in long branches. The result is that the genetic distance between the two group members is actually higher than that between either and the outside member. This explanation requires that the branch leading to the outside member be correspondingly short, which becomes increasingly problematic over very long periods of time, but as observed in the orexin receptor example above has occurred to some extent. This model also requires that the taxa at positions A and B in Figure 6 become very divergent to each other and relative to the family, which is not the case in the examples cited above. Another model requirement is that there is some considerable distance between X and both A and B, which as outline from the example above is not the case. In the second phylogenetic model (Fig. 6B) the ancestor to taxa A and B was very similar to the outsider taxon X due to homoplasy. Again in this model there is a higher rate of change in the lineages of A and B relative to X resulting in both A and B appearing to be more like X than each other. This model is problematic because in order for it to work the homoplasy must have been so extensive as to be an exact match between the lineages leading to A/B and X in order for A and B not to resemble each other more than X. The statistics of the example shown in Figure 4 demonstrate that such extensive homoplasy is unlikely in this data set. In

the third phylogenetic model (Fig. 6C) the same pattern of genetic distances as in model 2 could result from a recombination event between the outside member and the common ancestor of the two members displaying correlated mosaicism. This model requires the point of ancestral similarity to be an instance of identity, and so works more credibly than the second model. Alternatively, model 4 shown in Figure 6D attributes each instance of mosaicism to a separate recombination event. This is the only model that does not require a lower rate of change in X from a factor such as functional constraint. This model would be consistent with one which results from the underlying genomic architecture, in which a tendency for recombination between certain regions of receptors of very different groups due to genomic position or processes at the time of recombination. This explanation would predict that the pattern of correlated mosaicism would be due to multiple unrelated events. Alternatively, one may not put the restriction on that the two taxa involved in mosaic correlation are nearest-neighbors. Under this condition, a fifth model also becomes possible, Figure 7. In this case the branches leading to the correlated taxa (A and C) are short representing conservation of sequence. This model appears the most plausible one in which recombination does not occur. However, it requires the taxa concerned to occur at a basal position in the phylogeny, and that branch lengths vary widely within the GPCR group. It is generally not true that branch lengths vary so widely in the GPCR phylogeny, and rarely the case that correlated taxa are basal, as in the case outlined in Figure 5. Of these five phylogenetic models, those that do not involve recombination have serious phylogenetic problems. This leaves the recombination-based models as the most likely explanations for correlated mosaicism.

