Comparative sequence analyses of rhodopsin and RPE65 reveal patterns of selective constraint across hereditary retinal disease mutations

Frances E. Hauser, Ryan K. Schott, Gianni M. Castiglione, Alex Van Nynatten, Alexander Kosyakov, Portia Tang, Daniel Gow, and Belinda S.W. Chang

Version  Post-print/accepted manuscript


How to cite TSpace items

Always cite the published version, so the author(s) will receive recognition through services that track citation counts, e.g. Scopus. If you need to cite the page number of the author manuscript from TSpace because you cannot access the published version, then cite the TSpace version in addition to the published version using the permanent URI (handle) found on the record page.

This article was made openly accessible by U of T Faculty. Please tell us how this access benefits you. Your story matters.
Comparative sequence analyses of rhodopsin and RPE65 reveal patterns of selective constraint across hereditary retinal disease mutations

Frances E. Hauser\textsuperscript{1}, Ryan K. Schott\textsuperscript{1}, Gianni M. Castiglione\textsuperscript{2}, Alex Van Nynatten\textsuperscript{2}, Alexander Kosyakov\textsuperscript{1,2}, Portia Tang\textsuperscript{1,2}, Daniel Gow\textsuperscript{1,2}, and Belinda S.W. Chang\textsuperscript{1,2,3}\textsuperscript{*}

\textsuperscript{1}Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada
\textsuperscript{2}Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada
\textsuperscript{3}Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada

*Correspondence:
Dr. Belinda Chang
Department of Cell & Systems Biology
Department of Ecology & Evolutionary Biology
University of Toronto
25 Harbord St
Toronto ON Canada
M5S 3G5
belinda.chang@utoronto.ca

Short title: Comparative analyses of retinal disease mutations

Number of pages: 51
Number of tables: 3
Number of figures: 5
Abstract

Retinitis pigmentosa (RP) comprises several heritable diseases that involve photoreceptor, and ultimately retinal, degeneration. Currently, mutations in over 50 genes have known links to RP. Despite advances in clinical characterization, molecular characterization of RP remains challenging due to the heterogeneous nature of causal genes, mutations, and clinical phenotypes. In this study, we compiled large datasets of two important visual genes associated with RP: rhodopsin (RHO), which initiates the phototransduction cascade, and the retinoid isomerase RPE65, which regenerates the visual cycle. In this study, we use a comparative evolutionary approach to investigate the relationship between interspecific sequence variation and pathogenic mutations that lead to degenerative retinal disease. Using codon-based likelihood methods, we estimated evolutionary rates ($d_N/d_S$) across both genes in a phylogenetic context to investigate differences between pathogenic and non-pathogenic amino acid sites. In both genes, disease-associated sites showed significantly lower evolutionary rates compared to non-disease sites, and were more likely to occur in functionally critical areas of the proteins. The nature of the dataset (e.g., vertebrate or mammalian sequences), as well as selection of pathogenic sites, affected the differences observed between pathogenic and non-pathogenic sites. Our results illustrate that these methods can serve as an intermediate step in understanding protein structure and function in a clinical context, particularly in predicting the relative pathogenicity (e.g., functional impact) of point mutations and their downstream phenotypic effects. Extensions of this approach may also contribute to current methods for predicting the deleterious effects of candidate mutations, and to
identification of protein regions under strong constraint where we expect pathogenic mutations to occur.

**Keywords:** molecular evolution, retinitis pigmentosa, molecular phylogenetics, rhodopsin, RPE65, phylomedicine, likelihood-based codon models of evolution
**Introduction**

Retinitis pigmentosa (RP) encompasses several heritable diseases that involve the degeneration of photoreceptor cells. Individuals with RP exhibit symptoms such as night blindness, followed by decreasing visual fields, and ultimately progressive visual impairment that can result in legal or complete blindness (Hartong et al., 2006). On the basis of inheritance, non-syndromic RP can be divided into autosomal dominant, recessive, and X-linked forms. Multiple factors contribute to the heterogeneous nature of RP (reviewed in Rivolta et al., 2002; Anasagasti et al., 2012). Clinical RP phenotypes vary considerably with respect to severity, age of onset, and causal genes, and are thought to reflect the effects of particular mutations of RP-associated proteins (Mendes et al., 2005; Hartong et al. 2006; Anasagasti et al., 2012). Currently, mutations in over 50 genes have known links to RP, and recent advances in high-throughput sequencing are frequently uncovering novel causative genes and mutations (e.g., Bowne et al., 2011; Nishiguchi et al., 2013; Yang et al., 2014). Despite these advances in diagnosis and clinical characterization, molecular characterization of RP remains challenging due to the diversity of causal genes and their numerous disease-associated mutations. These disease genes are involved in a variety of processes both within and outside of the visual cycle, such as vitamin A metabolism, cell-cell interactions, cell structure, and signaling (reviewed in Hartong et al., 2006).

One of the most prominent and best characterized RP disease genes investigated to date is rhodopsin (RHO), the visual pigment found in rod photoreceptor cells that initiates the visual transduction cascade. Mutations in RHO account for roughly 30% to 40% of dominant RP cases (Ferrari et al., 2011). Of the variety of RP-linked genes
studied to date, RHO is particularly amenable to studies exploring the link between variation in amino acid residues and protein function. First, it has been established that disease-causing mutations in RHO often interfere with proper protein folding or transportation, resulting in the aggregation of defective proteins and subsequent photoreceptor cell death (Breikers et al., 2002; Mendes et al., 2005). Second, the crystal structure of rhodopsin has been resolved for both the inactive and active conformations (Palczewski et al., 2000; Choe et al., 2011). Third, numerous mutant rhodopsin proteins (comprising both naturally occurring and pathogenic mutants) have been characterized with respect to structure and function (e.g., Janz et al., 2003; Piechnick et al., 2012; McKeone et al., 2014; Morrow & Chang, 2015). Due to its importance in visual transduction, natural variation in rhodopsin sequence, structure, and function has been extensively investigated across a diversity of animals (Zhao et al., 2009; Sugawara et al., 2010; Porter et al., 2011; Schott et al., 2014).

Outside of the visual transduction cascade, the visual cycle gene RPE65 has also been experimentally investigated with respect to clinically relevant mutations: over 60 point mutations have been linked to both RP and an early-onset form of retinal degeneration, Leber’s congenital amaurosis. RPE65 is a large membrane-bound retinoid isomerase located in the retinal pigment epithelium, and is responsible for the conversion of activated all-trans-retinal to 11-cis-retinal. While less extensively studied relative to RHO, recent structural and clinical studies have begun to elucidate the effect of point mutations on RPE65 function in greater detail (Li et al., 2014; Takahashi et al., 2014; Li et al., 2015; Moiseyev et al., 2005; Redmond et al., 2005; Philpa et al., 2009; Cideciyan, 2010), and a crystal structure has also been resolved (Kiser & Palczewski, 2010).
Despite numerous studies examining RP mutations in RHO and other causal RP genes in human populations, molecular characterization of this disease remains challenging, and accurately validating disease mutations is not always straightforward, particularly in genes for which there is little functional information. Comparative analyses combined with studies of protein structure and function can lend insight into the molecular determinants of disease, and help to inform which mutations are more likely to result in a disease phenotype. Intuitively, mutations at highly conserved amino acid sites are more likely to be deleterious, often resulting in the disease phenotypes observed in RP (e.g., Briscoe et al., 2004; Iannaccone et al., 2006). However, despite the high conservation of RHO and RPE65 and their critical role in phototransduction and visual regeneration, there is abundant natural sequence variation in both genes across vertebrate groups. This variation presents an opportunity to conduct an evolutionarily informed (‘phylomedicine’) study of retinal protein function in the context of hereditary disease.

The genetic complexity of RP combined with recent advances in sequencing technology means that accurate and reliable molecular diagnosis of disease-causing RP mutations is particularly important (reviewed in Anasagasti et al., 2012). Through the analysis of natural sequence variation, we can make inferences about the evolutionary constraints operating on amino acid residues critical for proper protein function. In addition to identifying the location of conserved residues, codon-based methods of sequence analysis can estimate the degree of selective constraint imposed on a given site, as indicated by the evolutionary rate ratio of non-synonymous to synonymous amino acid substitutions, or $d_{NS}/d_S$. This approach assumes that functionally important residues will be subject to greater selective constraint, and therefore lower $d_{NS}/d_S$, and that mutations at
these sites are more likely to be deleterious and potentially disease-causing. Numerous studies have harnessed interspecific sequence analysis to examine known disease mutations, encompassing a variety of diseases and proteins (Greenblatt et al., 2003; Gaucher et al., 2006; Stover & Verrelli, 2010; Rishishwar et al., 2012; Kirwan et al., 2013). Beyond an explicitly disease-focused context, these comparative methods have also addressed areas of medical relevance, such as longevity (Morgan et al., 2013) and immune response (Webb et al., 2015), using estimates of positive selection to interpret and identify potential changes in protein function. In addition to known disease sites, cross-species data may aid predictions of other conserved amino acid residues critical for proper protein function that would likely generate a defective product if mutated (Miller & Kumar, 2001; Mooney & Klein, 2002; Greenblatt et al., 2003). To take advantage of the diversity of sequences available, an appropriate phylogenetic scope and for coding sequence sampling must be established. For analyses of human disease genes, although it may be desirable to use only mammalian sequences, if these sequences have high similarity there may be insufficient variation to detect conserved versus non-conserved sites (but see Cooper, 2003).

In this study, we assessed the evolutionary rate ratio ($d_N/d_S$) in vertebrate and mammalian datasets of the visual transduction gene RHO and the visual (retinoid) cycle gene RPE65, and compared the evolutionary conservation of codon sites harboring disease-associated (pathogenic) mutations with sites that have not been implicated in disease. Our aim was to assess the extent to which codon-based evolutionary analyses can inform studies of human diseases caused by defective proteins. Metrics of evolutionary conservation using orthologous sequences are implemented in several disease prediction
programs, such as PolyPhen (Adzhubei et al., 2010) and SIFT (Ng, 2003), but the
appropriate method of homologous sequence analysis varies depending on the nature of
the alignment, disease, and protein in question (Flanagan et al., 2010; Hicks et al., 2011).
Incorporating measures of $d_N/d_S$ as estimated by the methods outlined in this study may
provide a useful complement to programs designed to specifically identify disease-
causing protein defects. Using large sequence datasets, we investigate how varying the
level of taxonomic divergence, the selection of pathogenic sites, and the molecular
evolutionary method can impact analyses of human visual disease genes. Overall,
pathogenic sites had significantly lower $d_N/d_S$ values in comparison to non-pathogenic
sites. Many of the RHO mutations known to cause retinitis pigmentosa have been
examined through mutagenesis experiments; however, new disease-associated RHO
mutations are frequently discovered (Liu et al., 2013; Hollingsworth & Gross, 2013;
Opefi et al., 2013; Yang et al., 2014; Pierrottet et al., 2014). In genes such as RPE65, for
which there are fewer studies of structure and function, computational analyses of this
nature may provide a crucial first step in understanding novel mutations in a clinical
context.

Methods

Dataset acquisition

RHO and RPE65 genes were selected for comparative analyses of coding sequence
evolution in vertebrates. These genes were selected based on an abundance of
nonsyndromic disease-associated missense or nonsense mutations in humans, expression
in the retina, association with hereditary retinal degeneration, and a large number of
available sequences for robust molecular evolutionary analyses. Since there are fewer RPE65 sequences available compared to RHO, an exhaustive search was performed to identify and acquire all vertebrate RPE65 sequences available on GenBank. We then acquired RHO sequences from these same species, when possible, to maintain as similar taxon sampling as possible, which ensured the datasets were comparable. We also created a second RHO dataset that contained only mammals, but with a similar number of sequences as the vertebrate datasets. As we already included all available RPE65 sequences, our mammal dataset for RPE65 is simply the subset containing only mammals. Species list and accession numbers for all sequences used in this study are provided in Supplementary table 1.

Phylogenetic analyses

Sequences from each of the four datasets were aligned separately using PRANK codon alignment, which has been found to reduce coding sequence alignment errors for evolutionary analysis by incorporating phylogenetic information into gap placement (Löytynoja & Goldman, 2005; Löytynoja & Goldman, 2008). Gene trees were estimated in MrBayes 3 (Ronquist & Huelsenbeck, 2003) and by maximum likelihood (ML) using PhyML 3 (Guindon et al., 2010) (Supplemental figs. 1-8). The Bayesian analyses were run for five million generations with a 25% burn-in. Convergence was determined by verifying that the standard deviations of split frequencies approached zero and that there was no obvious trend in the log likelihood plot. The ML analyses were run under the GTR+G+I model with a BioNJ starting tree, the best of NNI and SPR tree improvement, and aLRT SH-like branch support (Anisimova & Gascuel, 2006).
Molecular evolutionary analyses

To estimate the strength of selection acting across RHO and RPE65, as well as the strength of selection acting on individual amino acid sites, the alignment and gene trees were analyzed in the HYPHY software package (Kosakovsky Pond et al., 2005), as implemented on the Datamonkey webserver (Delport et al., 2010) and in the codeml package of PAML 4 (Yang, 2007). Specifically, we used the PAML random sites models (M0, M1a, M2a, M3, M7, M8), and the FEL and FUBAR methods in HYPHY (Scheffler et al., 2006; Yang, 2007; Murrell et al., 2013). The PAML random sites models estimate \( d_N/d_S \) as a single value for a prespecified number of site classes. The posterior probability distribution of the assignment of individual sites to these site classes is then calculated using a Bayes empirical Bayes analysis and \( d_N/d_S \) values calculated as weighted averages from this distribution. For comparison to the HYPHY models we chose M8, as this was most often the best-fitting model and was most comparable in terms of the numbers of site classes. For M8 site categories are calculated from a \( \beta \) distribution ranging from zero to one, discretized into 10 categories, with an additional site category with \( d_N/d_S \geq 1 \).

Unlike the PAML models, FEL and FUBAR estimate \( d_N \) and \( d_S \) separately. FEL estimates \( d_N \) and \( d_S \) independently for each site, while FUBAR estimates \( d_N \) and \( d_S \) using alignment-wide information and 400 rate categories. All analyses were carried out with both the Bayesian and ML trees.

Analyses of disease associated mutations
Disease associated mutations were identified for both RHO and RPE65. For the purposes of this study we defined pathogenic mutations as being any mutation reported in the Human Gene Mutation Database (Stenson et al., 2013), dbSNP (Sherry et al., 2001), UniProt (Magrane & Consortium, 2011) or Online Mendelian Inheritance in Man (Hamosh, 2004) disease databases (referred to as “reported” pathogenic mutations in our RHO dataset). Sites linked to congenital stationary night blindness were included in our list of RHO pathogenic sites, as well as sites linked to Leber congenital amaurosis in RPE65. As substantial verification of disease-associated mutations has been performed for RHO, we also created a second subset of disease sites (referred to hereafter as ‘confirmed pathogenic sites’) that included only those sites with either experimental validation or that were clinically verified with familial co-segregation studies (Supplemental table 2). We used these criteria to categorize sites as either pathogenic or non-pathogenic (and for RHO, confirmed pathogenic and non-pathogenic) and compared $d_{N}/d_{S}$ values between these two groups. Because the HYPHY models estimate $d_{N}$ and $d_{S}$ separately, we were also able to compare a second metric, $d_{N} - d_{S}$, as well as $d_{N}$ and $d_{S}$ separately. To identify statistically significant differences between estimates at pathogenic versus non-pathogenic sites we used the non-parametric Mann-Whitney U test, due to substitution rates being highly non-normal. Comparisons were made using the M8 (PAML), FUBAR (HYPHY), and FEL (HYPHY) models for the vertebrate and mammals datasets of RHO and RPE65 using both the Bayesian and ML tree topologies. Values of $d_{N}/d_{S}$ for RHO and RPE65 from vertebrate FUBAR analyses were scaled and converted into RGB color gradient space (where white represents the highest $d_{N}/d_{S}$, and dark blue the lowest $d_{N}/d_{S}$) and mapped onto the 3D crystal structures of RHO (IU19;
Okada et al., 2004) and RPE65 (3FSN; (Kiser et al., 2009)) using Chimera (Pettersen et al., 2004). This created a ‘heat map’ of conservation values ($d_N/d_S$) that were visualized in the context of the 3D structure. To infer how structural features correspond to conservation and pathogenicity, pathogenic sites in RHO and RPE65 were highlighted in both structures.

Results

Alignments and trees

The alignment length and number of sequences for the vertebrate and mammalian datasets of RHO and RPE65 are summarized in Table 1, as well as the number of disease-associated sites analyzed in each gene. Results from the phylogenetic analyses of the vertebrate and mammalian RHO and RPE65 alignments were generally congruent with species relationships (Meredith et al., 2011; Crottini et al., 2012; Fong et al., 2012). Bayesian and Maximum likelihood phylogenies for RHO and RPE65 were similar, and both phylogenies were used in subsequent molecular evolutionary analyses. The results of the analyses were consistent regardless of the phylogenetic reconstruction method used (Supplemental tables 3-9).

Molecular evolutionary analyses

To assess selective pressure across vertebrate and mammalian datasets of RHO and RPE65 we used a variety of molecular evolutionary approaches: the random sites models in the PAML software package (Yang, 2007), and FEL and FUBAR models in the HYPHY software package (Kosakovsky Pond et al., 2005). The random sites PAML
models confirmed that there is variation in $d_N/d_S$ in the vertebrate and mammalian datasets of both RHO and RPE65 (the ratio of nonsynonymous to synonymous substitution rate; M3 vs. M0, P <0.001; Tables 2 and 3), as expected for protein-coding genes. These models did not find significant evidence for a positively selected class of sites (M2a vs. M1a, M7 vs. M8, M8 vs. M8a where applicable; Tables 2 and 3) for either gene, reflecting high evolutionary conservation across both vertebrates and mammals. No positively selected (i.e., $d_N/d_S > 1$) sites were found using any of the molecular evolutionary analyses implemented in this study.

Evolutionary constraint on disease-associated amino acid sites

In both RHO and REP65, we compared estimates of $d_N/d_S$ across individual amino acid sites generated by three different approaches: the M8 model in PAML, and the FEL and FUBAR methods in HYPHY (Yang, 2005; Kosakovsky Pond et al., 2005) (Murrell et al., 2013). Figure 1 shows a series of histograms illustrating the distribution of pathogenic and non-pathogenic sites across $d_N/d_S$ value categories, as estimated by FUBAR, FEL, and M8 in vertebrate RHO (all reported sites) and vertebrate RPE65. Regardless of method used, pathogenic sites had significantly lower $d_N/d_S$ when compared to non-pathogenic sites (Supplemental table 6); however, Figure 1 illustrates how the distribution of estimated $d_N/d_S$ values in RHO and RPE65 differs depending on the method used. FUBAR estimates the highest proportion of disease sites with $d_N/d_S$ values less than 0.01 in both RHO and RPE65 (Fig. 1A, B), with an increasingly lower proportion estimated at higher $d_N/d_S$ values. FUBAR does not estimate any pathogenic sites with
$d_S/d_S$ values greater than 0.5 in RPE65 (Fig. 1B). Non-pathogenic sites in both genes show a more even distribution across the $d_S/d_S$ categories, with 6–10% of sites approaching neutral selection ($d_S/d_S = 1$), and a small proportion exceeding $d_S/d_S$ of 1. FEL estimates of $d_S/d_S$ reveal a similar trend to the FUBAR analyses (Fig. 1C,D). In contrast, for any given site in RHO or RPE65, M8 does not estimate a $d_S/d_S$ value less than 0.05 (Fig. 1E,F). The $d_S/d_S$ estimates overall show a narrower distribution in both pathogenic and non-pathogenic sites, with the majority of both types of sites having an estimated $d_S/d_S$ of 0.05.

To determine if pathogenic sites had statistically significantly lower $d_S/d_S$ values than non-pathogenic sites we compared values using the non-parametric Mann-Whitney test. For $d_S/d_S$ values estimated with FUBAR at each amino acid site, vertebrate and mammalian RHO sequences showed significantly lower $d_S/d_S$ values at all reported (84) pathogenic sites (Fig. 2A; Supplemental table 7). The vertebrate datasets exhibited greater overall constraint (lower $d_S/d_S$) on pathogenic sites relative to the mammalian datasets, and showed more significant differences when compared to the $d_S/d_S$ of non-pathogenic sites (Fig. 2A, Supplemental table 7). This difference was even larger when the confirmed pathogenic subset of sites was compared to non-pathogenic sites, with these 53 sites having lower $d_S/d_S$ relative to the 84 sites analyzed, and these $d_S/d_S$ values were more significantly different from non-pathogenic sites (Fig. 2A, Supplemental table 7). This pattern was found in both vertebrates and mammals. In RPE65, a similar finding was recovered using the FUBAR-estimated $d_S/d_S$ values. Vertebrates exhibited significant differences in $d_S/d_S$ at pathogenic vs. non-pathogenic sites, while the
differences detected in mammals were not significant, likely due to the lower available sample size (Fig. 2B; Supplemental table 7).

We found that differences between pathogenic and non-pathogenic sites remained significant using both the M8 and FEL methods, and supported our FUBAR results. In general, we found that out of the three methods the FUBAR method performed best (Supplemental tables 6 and 7). At certain sites, FEL estimates either undefined or infinite values of $d_N/d_S$ due to extremely low $d_S$ values, so $d_N-d_S$ is also used to infer selection. When $d_N-d_S$ was compared between pathogenic and non-pathogenic sites, differences were not significant with the exception of the confirmed subset of pathogenic sites in mammalian and vertebrate RHO, and vertebrate RPE65 (Supplemental table 8). Although the $d_N/d_S$ estimates generated through M8 in PAML spanned a much narrower range of values relative to FUBAR and FEL, we still detected differences between pathogenic and non-pathogenic sites.

To explore the distribution of pathogenic vs nonpathogenic sites in more detail, we plotted each amino acid site with respect to the number of different amino acid residues found in the alignment at that site, against its assigned $d_N/d_S$ value as estimated using FUBAR, for RHO and RPE65 (Fig. 3A,B). This shows that the majority of pathogenic sites are consistently found at invariant and highly conserved amino acid sites (with accompanying low $d_N/d_S$), while non-pathogenic sites show more variation in both dimensions. This pattern is even stronger in the ‘confirmed’ subset of RHO pathogenic sites (Fig. 3A).

Since FUBAR allows $d_N$ and $d_S$ to be estimated independently, we were able to obtain individual values of $d_N$ and $d_S$ for each site in our vertebrate RHO and RPE65
alignments. Vertebrate RHO shows significant differences in $d_N$ between all reported pathogenic sites and non-pathogenic sites (Fig. 4A,B; Supplemental table 9), but no differences in $d_S$. The confirmed pathogenic set of RHO sites, however, shows significantly lower $d_N$ and higher $d_S$. In vertebrate RPE65, $d_N$ is significantly lower and $d_S$ significantly higher compared to pathogenic sites (Fig. 4C,D; Supplemental table 9).

Disease sites and $d_N/d_S$ on the crystal structure

Using FUBAR-estimated vertebrate $d_N/d_S$ values, we created a $d_N/d_S$ heatmap of conservation (where dark blue = most conserved and white = least conserved) to illustrate the relationship between structure and disease in both RHO and RPE65 (Fig. 5). The interior views of RHO (Fig. 5A) and RPE65 (Fig. 5B) show conserved residues surrounding the retinal chromophore in RHO and the catalytically active iron cofactor in RPE65. Conversely, the exterior of the proteins (the smaller structures shown above the interior views of the respective proteins) have higher variability in $d_N/d_S$. Pathogenic sites, outlined in red, are distributed throughout both proteins, tending to occur in areas of low $d_N/d_S$. 

15
Discussion

This study used a phylomedicine-based approach to investigate retinitis pigmentosa-associated mutations in two critical visual genes using interspecific sequence variation across vertebrates and mammals. In vertebrate and mammalian RHO sequences we found that sites implicated in RP had significantly lower rates of non-synonymous to synonymous substitution ($d_N/d_S$) relative to those sites that were not associated with disease. We also found similarly significant lower rates for pathogenic sites in the vertebrate dataset for the visual cycle gene RPE65, but we did not find significant differences in the mammalian dataset, likely due to the smaller sample size and reduced variation in these sequences. In both genes, our vertebrate dataset yielded more statistically significant differences in mean $d_N/d_S$ between pathogenic and non-pathogenic amino acid sites, as compared with a more restricted mammalian dataset. For RHO, these differences were even more striking when we only included disease-associated sites that had substantial biochemical or clinical support for pathogenicity. Our results were consistent regardless of the molecular evolutionary approach used, but here we discuss some important differences among these methods and their implications for future studies of this nature. Finally, we found that differences in the non-synonymous substitution rate ($d_N$), rather than the synonymous substitution rate ($d_S$), between pathogenic and non-pathogenic sites, drives the overall differences in $d_N/d_S$ rates in most instances. We discuss the importance of dataset properties, pathogenic site selection, and molecular evolutionary method for our results, and compare our approach to similar evolutionary studies of visual disease genes and human disease in general.
Implications of dataset properties on analyses of disease sites

In both visual disease genes, we consistently found significantly lower mean $d_N/d_S$ in all analyses of pathogenic sites compared to non-pathogenic sites, with the single exception of the mammalian RPE65 dataset. This is most likely due to the absence of sufficient interspecific variation to generate (detectable) differences in $d_N/d_S$ among sites. Though both genes are highly conserved among mammals, the additional mammalian sequences included in our RHO dataset may have provided sufficient additional variation to detect significant differences between pathogenic and non-pathogenic sites. In general, our analyses of the vertebrate dataset had greater differences between pathogenic and non-pathogenic sites and higher significance levels. Previous comparative studies of human disease genes have used various levels of evolutionary depth and taxonomic sampling (e.g., retinitis pigmentosa (Briscoe et al., 2004); deafness (Kirwan et al., 2013); and cystic fibrosis (Rishishwar et al., 2012)). Past work has suggested that analyses of mammalian sequences are most appropriate for studies of human disease (Cooper, 2003). While this is feasible for extensively sequenced genes (e.g., RHO), genes with less sequence information, such as RPE65, may require a wider evolutionary scope for statistically robust analyses. In general, a wider scope when fewer interspecific sequences are available is likely to enhance the ability to detect differences between pathogenic and non-pathogenic sites, due to an increased amount of variation at non-pathogenic sites as a result of increased evolutionary divergence, when function is conserved. It is important to note that increased sequence divergence can also lead to functional divergence, potentially confounding studies aiming to investigate human disease. In the case of RHO and RPE65, where function is highly conserved across vertebrates, this is not likely to be
in issue, but should be carefully considered for future studies on different genes.

Fortunately, with the increasing number of both mammalian and vertebrate genomes available, evolutionary studies of human disease using interspecific variation can be undertaken at multiple taxonomic levels.

Beyond varying the taxonomic scale in our study, we also performed our RHO analyses on two separate sets of pathogenic sites. We found that both our more stringent pathogenic site dataset (53 confirmed pathogenic sites, selected based on substantial clinical or biochemical evidence for deleterious effects; Supplemental table 2) and the ‘all reported’ pathogenic sites both showed significant differences between pathogenic and non-pathogenic sites. The confirmed pathogenic sites dataset was under greater constraint than all reported pathogenic sites and thus gave more significant results, suggesting the some of the ‘all reported’ pathogenic sites may not actually be pathogenic, or may be less severe and therefore not under as strong purifying selection. This result demonstrates that the method is robust to the inclusion of sites that do not have biochemical or clinical confirmation, since most disease genes lack substantial experimental investigation of point mutations and their impact on protein structure and function.

These results also highlight an application of the method for selecting putative disease-causing mutations that lack clinical or biochemical characterization, but have low $d_{S}/d_{S}$ similar to known pathogenic sites, for future in vitro mutagenesis experiments.

Mutations implicated in disease, but of unknown effect, that show comparatively lower evolutionary constraint (i.e., more interspecific variation) may also warrant further clinical or experimental investigation to confirm that they are actually deleterious rather
than benign polymorphisms (e.g., Vincent et al., 2013). For disease-associated genes with
less functional data available, deleterious mutations found in these databases can be
examined through comparative approaches outlined in this study to identify high-priority
sites (highly conserved sites) for further investigation.

Detection of higher constraint in disease sites achieved with multiple methods

The three molecular evolutionary methods used to investigate RHO and RPE65 (FUBAR,
FEL, and M8) yielded similar results, and each method’s mean $d_{\text{N}}/d_{\text{S}}$ estimates showed
significant differences between pathogenic and non-pathogenic sites. The distribution of
disease sites across varying levels of evolutionary conservation was comparable between
RHO and RPE65 regardless of the method implemented. Using ratios derived from FEL
may be problematic when $d_{\text{S}}$ is zero, as $d_{\text{N}}/d_{\text{S}}$ will be calculated as infinite (or undefined
when $d_{\text{N}}$ is also zero). In these instances, using normalized $d_{\text{N}}$-$d_{\text{S}}$ values for such
comparisons between site categories could be used instead of $d_{\text{N}}/d_{\text{S}}$; or, the issue can be
avoided via use of methods such as FUBAR that incorporate alignment-wide information
rather than calculating $d_{\text{N}}$ and $d_{\text{S}}$ for each site independently (Kosakovsky Pond, 2005;
Murrell et al., 2013). We found that in contrast to FUBAR and FEL, site-by-site $d_{\text{N}}/d_{\text{S}}$
estimates from M8 were assigned a much narrower range of values, since these estimates
are generally restricted to 10 rate categories. This discretization of the M8 $\beta$ distribution
into rate categories also results in $d_{\text{N}}/d_{\text{S}}$ estimates that do not fall below a certain
threshold (in the case of these analyses, 0.05). In contrast, FUBAR avoids this due to its
larger number of site classes that are able to capture sequence variation that cannot
otherwise be captured by a (discretized) $\beta$ distribution. Despite these discrepancies across
methods, we were still able to detect significant differences between pathogenic and non-
pathogenic sites in RHO, and the M8 model in particular has been successfully used in
previous studies to investigate auditory disease mutations in mammals (Kirwan et al.,
2013).

An advantage of the FUBAR and FEL analyses implemented in this investigation
is the ability to separately examine $d_N$ and $d_S$. Differing levels of evolutionary constraint
on non-synonymous substitutions at disease-associated amino acid sites is intuitive, since
these mutations alter the amino acid sequence, and therefore the biochemical properties,
of a particular protein (Li et al., 1985). Elevated mean $d_S$ at pathogenic sites in RHO and
RPE65 may be driven by outlier pathogenic sites with high $d_S$, or perhaps interspecific
variation in codon usage bias at these sites (Plotkin & Kudla, 2010). The role of
synonymous variation in human disease has gained prominence in recent years (e.g.,
Sauna et al., 2007; Sauna & Kimchi-Sarfaty, 2011), and work on mammalian RHO in
particular has identified selection on synonymous sites, likely contributing to mRNA
stability and translational efficiency (Du et al., 2014). A broader investigation of many
protein-coding gene families also found that genes with synonymous rate variation were
overrepresented in genetic diseases (Dimitrieva & Anisimova, 2014). Our finding that $d_S$
may be consistently higher in pathogenic sites suggests that future work investigating the
role of synonymous rate variation in genes involved in human disease is warranted.

Protein structure, $d_N/d_S$, and pathogenicity

Although an examination of specific structural defects in protein function is not
readily available with the codon-based models used in this study, incorporating metrics of
conservation from multiple methods can be an important first step for future analyses integrating protein structure and amino acid identity into disease site prediction, and downstream phenotypic characterization of mutant proteins. For instance, RP-causing mutations in RHO are found throughout the protein, with the majority located within the transmembrane domains. Known pathogenic mutations disrupt protein folding, and several restricted to the C-terminus interfere with protein translocation but proper protein structure is maintained (Mendes et al., 2005). Critical structural sites, such as those that participate in glycosylation and those responsible for correct protein folding, are highly conserved across vertebrates and mutations at these sites invariably result in protein defects (Pope et al., 2014). Conserved $d_{\text{s}}/d_{\text{s}}$ estimates mapped onto the RHO crystal structure coincide with these known highly conserved regions, as well as confirmed pathogenic sites. In some cases, however, a pathogenic site may be tolerant of a non-synonymous substitution to a particular amino acid residue, while the true disease-associated residue differs considerably from the wild type with respect to size or charge. For instance, the A292S substitution in RHO occurs frequently across vertebrates (e.g., Sugawara et al., 2010), while the A292E substitution is not found in vertebrate sequences and is known to cause constitutive activation of RHO (Weitz & Nathans, 1992). Moreover, different amino acid substitutions at the same amino acid site in rhodopsin may lead to different clinical disease phenotypes (Neidhardt et al., 2006). In our vertebrate analyses, we found that in both RHO and RPE65 certain pathogenic amino acid sites (including those with functional validation in RHO) exhibited some variability, supporting the use of amino acid characterization methods to determine which residue in particular is deleterious. In these instances, further in silico investigations of the mutation
in question could be undertaken with disease-prediction methods such as SIFT, PolyPhen, or Provean (Ng, 2003; Adzhubei et al., 2010; Flanagan et al., 2010; Choi et al., 2012), with additional *in vitro* characterization to corroborate the mutation.

Intuitively, sites in both RHO and RPE65 showed high evolutionary conservation in structurally important regions, particularly when contrasting the interior with the exterior of the protein. Much like RHO, pathogenic residues in RPE65 are located throughout the protein and their deleterious effects are mediated via a variety of mechanisms (Bereta et al., 2008). In RPE65, amino acid sites associated with isomerase activity, those pointing toward the catalytic interior of the protein (specifically, four highly conserved histidine residues), and those that may mediate membrane association are tightly constrained (Kiser & Palczewski, 2010; Li et al., 2015). Recent mutagenesis work has begun to elucidate the importance of these and other disease-associated residues for RPE65 function (Li et al. 2014). Since known pathogenic mutations are widespread in the RPE65 structure, rather than localized to a few key regions, further work investigating mutations *in vitro* will lend valuable insight into its function.

*Evolutionary conservation and disease site prediction*

The comparative approach used in this study is ideal for visual genes that are known to have many point mutations associated with RP. However, in some instances only one or two mutations in a given visual protein are known, but many more are likely to be deleterious and result in a disease phenotype that have not yet been found. Other techniques complementary to the ones employed in this paper can be used to elucidate the nature of these mutations. Comparing observed and expected numbers of disease-
associated non-synonymous substitutions across sites in a protein has been used to explore the spectrum of selective constraint across genes (reviewed in Miller & Kumar, 2001; Kumar et al., 2011). For these genes, even a small number of disease sites can be investigated, by testing whether disease mutations occur at highly conserved sites more frequently than expected by chance. An evolutionary prior, or “conservation index” ($1 - d_S/d_S$) has also been used to evaluate selective constraint across sites undergoing purifying selection (Burk-Herrick et al., 2006). In this approach, sites undergoing neutral or positive selection have an index set to 0, generating a spectrum of sites ranging from weak to strong purifying selection, wherein the most highly conserved sites have a conservation index approaching 1. Similar to our expectations for this study, pathogenic sites in the mammalian BRCA1 gene and ten mammalian auditory genes were found to predominantly occur in regions undergoing strong purifying selection (Burk-Herrick et al., 2006; Kirwan et al., 2013). An additional caveat to the methods used in this study is that missense/nonsense mutations were the sole target of our analyses. In addition to defects in protein structure/function, the role of pre-mRNA splicing mutations in human disease has been documented, including in hereditary visual diseases (e.g., Ueyama et al., 2012). Extending these existing phylomedicine approaches to incorporate splicing mutations would be a valuable avenue of future study.

Current tools aiming to predict the occurrence of novel disease sites often incorporate structural information for specific amino acids as well as entire proteins (Ng, 2003; Adzhubei et al., 2010), and can filter sequence variants that are of potential functional importance. In some cases, these methods may be sensitive to changes in alignment and taxonomic sampling, and may not be sufficiently sensitive to resolve
ambiguity surrounding particular disease sites (Flanagan et al., 2010; Hicks et al., 2011; Castellana et al., 2015); however, incorporating structural information and amino acid properties is an important aspect of disease prediction that is not explicitly addressed in the molecular evolutionary analyses undertaken here. Statistical estimates of selection obtained from several different molecular evolutionary methods could be integrated into existing techniques that use interspecific alignments to assess the severity of a given mutation (e.g., PolyPhen or SIFT) or complement these alignment-based analyses, improving the ability for interspecific variation to inform studies of human disease and predict deleterious mutations.

Conclusions

With the advancement of high-throughput sequencing technology, rapid and extensive recovery of putative genes and mutations implicated in visual disease is increasingly feasible (Neveling et al., 2012). This volume of data, particularly for newly recovered missense mutations, means that further validation of their clinical relevance is necessary. How likely are newly recovered mutations or genes to result in a disease phenotype? In instances where insufficient clinical or sequence information precludes the identification of important amino acid sites, computational approaches drawing on interspecific data found in vertebrate genomes, such as those employed in this study, may serve as an important first step in the characterization of novel genes and mutations. Genes implicated in hereditary visual disease are good candidates for these studies since there is abundant homologous sequence data for many genes, and a strong relationship
between genotype and visual degeneration phenotype can be established. In this study, we used several molecular evolutionary methods and interspecific datasets to investigate disease-causing mutations in RHO and RPE65, and found that pathogenic sites in these genes were under significantly stronger evolutionary constraint than non-pathogenic sites. Models of molecular evolution may aid in the identification, prediction, and investigation of point mutations associated with human disease. Moreover, at the codon level, they may reveal patterns of selection not easily recovered with amino acid-based approaches. Current disease prediction methods using homologous sequences and amino acid properties may be complemented by codon-based molecular evolutionary analyses, advancing our understanding of conservation, protein function, and pathogenicity.

Acknowledgements

This work was supported by an NSERC Discovery Grant and a Foundation for Fighting Blindness seed grant to BSWC, Vision Science Research Fellowships to FEH, RKS, GMC, and AVN, and Ontario Graduate Fellowships to FEH and RKS. Amir Sabouhanian provided valuable feedback on the statistical analyses.
Figure 1. Frequency histograms comparing the assignment of $d_N/d_S$ ratios of pathogenic and non-pathogenic sites for FUBAR, FEL, and M8 analyses of vertebrate RHO and RPE65. For RHO, all 84 disease sites were used to generate histograms. Statistical comparisons of $d_N/d_S$ values between pathogenic vs non-pathogenic sites for all three methods are summarized in Supplemental table 6.
A. RH1

- Confirmed pathogenic
- All reported pathogenic
- Non-pathogenic

B. RPE65

- All reported pathogenic
- Non-pathogenic
Figure 2. Differences in mean $d_N/d_S$ of pathogenic vs non-pathogenic sites in RHO and RPE65 datasets. Estimates of $d_N/d_S$ were obtained with FUBAR. In all instances pathogenic sites had significantly lower $d_N/d_S$ with the exception of the RPE65 mammal dataset (Mann-Whitney test, P<0.05). Statistical comparisons are summarized in Supplemental table 7. Error bars represent +/-SEM.
Figure 3. Distribution of all alignment sites in A) Vertebrate RHO and B) Vertebrate RPE65 with respect to number of different amino acid identities at a particular site, and FUBAR-estimated $d_S/d_S$ values. All reported RHO pathogenic sites are shown in blue, and confirmed pathogenic RHO sites in red. To illustrate the number of disease-associated alignment sites with respect to amino acid variation (where 1 = one amino acid, i.e., no variation; 2 = two different amino acids, etc.), raw counts of the pathogenic sites relative to number of different amino acid identities at a particular alignment site are shown to the left of each graph.
Figure 4. Differences in mean non-synonymous and mean synonymous substitution rates of pathogenic vs. non-pathogenic sites in the RHO vertebrate datasets (A,B) and RPE65 (C,D), determined through FUBAR analyses. In both ‘all reported’ and ‘confirmed pathogenic’ site categories, \(d_N\) values were significantly lower relative to non-disease sites (Mann-Whitney test, \(P<0.05\)). For confirmed sites, \(d_S\) was significantly higher in disease sites, but this was not the case for ‘all reported’ sites. Statistical comparisons are summarized in Supplemental table 9. Error bars represent +/-SEM.
A. RHO

B. RPE65

highest $d_v/d_s$

lowest $d_n/d_s$
Figure 5. $d_{\text{N}}/d_{\text{S}}$ values estimated from vertebrate FUBAR analyses converted into a conservation heatmap and mapped on the 3D crystal structures of A) RHO and B) RPE65. The structures have been bisected longitudinally and internal views of both structures are shown. The rhodopsin chromophore is shown in yellow, and RPE65 iron cofactor as an orange dot. Pathogenic sites are outlined in dotted red lines.
References


BREIKERS G, PORTIER-VANDELUYTGAARDEN MJM, BOVEE-GEURTS PHM & DEGRIJF WJ (2002). Retinitis pigmentosa-associated rhodopsin mutations in three membrane-
located cysteine residues present three different biochemical phenotypes.


Biomarkers 14, 533–537.


Nishiguchi KM, Tearle RG, Liu YP, et al. (2013). Whole genome sequencing in patients with retinitis pigmentosa reveals pathogenic DNA structural changes and


Piechnick R, Ritter E, Hildebrand PW, Ernst OP, Scheerer P, Hofmann KP &


STENSON PD, MORT M, BALL EV, SHAW K, PHILLIPS AD & COOPER DN (2013). The Human Gene Mutation Database: building a comprehensive mutation repository for


Figure legends

Figure 1. Frequency histograms comparing the assignment of $d_N/d_S$ ratios of pathogenic and non-pathogenic sites for FUBAR, FEL, and M8 analyses of vertebrate RHO and RPE65. For RHO, all 84 disease sites were used to generate histograms. Statistical comparisons of $d_N/d_S$ values between pathogenic vs non-pathogenic sites for all three methods are summarized in Supplemental table 6.

Figure 2. Differences in mean $d_N/d_S$ of pathogenic vs non-pathogenic sites in RHO and RPE65 datasets. Estimates of $d_N/d_S$ were obtained with FUBAR. In all instances pathogenic sites had significantly lower $d_N/d_S$ with the exception of the RPE65 mammal dataset (Mann-Whitney test, $P<0.05$). Statistical comparisons are summarized in Supplemental table 7. Error bars represent +/-SEM.

Figure 3. Distribution of all alignment sites in A) Vertebrate RHO and B) Vertebrate RPE65 with respect to number of different amino acid identities at a particular site, and
FUBAR-estimated $d_N/d_S$ values. All reported RHO pathogenic sites are shown in blue, and confirmed pathogenic RHO sites in red. To illustrate the number of disease-associated alignment sites with respect to amino acid variation (where 1 = one amino acid, i.e., no variation; 2 = two different amino acids, etc.), raw counts of the pathogenic sites relative to number of different amino acid identities at a particular alignment site are shown to the left of each graph.

**Figure 4.** Differences in mean non-synonymous and mean synonymous substitution rates of pathogenic vs. non-pathogenic sites in the RHO vertebrate datasets (A,B) and RPE65 (C,D), determined through FUBAR analyses. In both ‘all reported’ and ‘confirmed pathogenic’ site categories, $d_N$ values were significantly lower relative to non-disease sites (Mann-Whitney test, $P<0.05$). For confirmed sites, $d_S$ was significantly higher in disease sites, but this was not the case for ‘all reported’ sites. Statistical comparisons are summarized in Supplemental table 9. Error bars represent +/-SEM.

**Figure 5.** $d_N/d_S$ values estimated from vertebrate FUBAR analyses converted into a conservation heatmap and mapped on the 3D crystal structures of A) RHO and B) RPE65. The structures have been bisected longitudinally and internal views of both structures are shown. The rhodopsin chromophore is shown in yellow, and RPE65 iron cofactor as an orange dot. Pathogenic sites are outlined in dotted red lines.