doi: 10.1046/j.1529-8817.2003.00072.x

Quantifying the Intragenic Distribution of Human Disease Mutations

M. P. Miller[1,3], J. D. Parker[1], S. W. Rissing[1] and S. Kumar[1,2,*]

[1]*School of Life Sciences and* [2]*Center for Evolutionary Functional Genomics, Arizona State University, Tempe, Arizona 85287-4501 USA,* [3]*Department of Biology and Department of Forest, Range, and Wildlife, Utah State University, Logan, UT 84322, USA*

Summary

A wide variety of functional domains exist within human genes. Since different domains vary in their roles regarding overall gene function, the ability for a mutation in a gene region to produce disease varies among domains. We tested two hypotheses regarding distributions of mutations among functional domains by using (1) sets of single nucleotide disease mutations for six genes (*CFTR*, *TSC2*, *G6PD*, *PAX6*, *RS1*, and *PAH*) and (2) sets of polymorphic replacement and silent mutations found in two genes (*CFTR* and *TSC2*). First, we tested the null hypothesis that sets of mutations are uniformly distributed among functional domains within genes. Second, we tested the null hypothesis that disease mutations are distributed among gene regions according to expectations derived from the distribution of evolutionary conserved and variable amino acid sites throughout each gene. In contrast to the mainly uniform distribution of sets of silent and polymorphic mutations, sets of disease mutations generally rejected the null hypotheses of both uniform and evolutionary-influenced distributions. Although the disease mutation data showed a better agreement with the evolutionary-derived expectations, disease mutations were found to be statistically overabundant in conserved domains, and under-represented in variable regions, even after accounting for amino acid site variability of domains over long-term evolutionary history. This finding suggests that there is a non-additive influence of amino acid site conservation on the observed intragenic distribution of disease mutations, and underscores the importance of understanding the patterns of neutral amino acid substitutions permitted in a gene over long-term evolutionary history.

Introduction

Over 900 disease-associated human genes have been identified (Jimenez-Sanchez *et al*. 2001). Although these genes are generally associated with specific functions, different domains within a given gene perform different roles related to the overall function of the protein product. In fact, over 1,800 different putative types of domains have been identified in human genes (Li *et al*. 2001). Since different regions within a gene vary in their roles with respect to overall gene function, amino acid sites in different domains within a gene may vary in their ability to produce a given disease phenotype if mutated. For example, Maheshwar *et al.* (1997) noted multiple recurrent missense mutations in only a single putative GAP domain of *TSC2* in unrelated tuberous sclerosis patients, indicating that this gene region may be overall more important than others for the production of the disease phenotype. In contrast, mutations in the cystic fibrosis transmembrane conductance regulator gene (*CFTR*) resulting in cystic fibrosis have been reported among all of the functional domains contained within it, despite the fact that different regions play different roles in the overall function of the gene product (Devidas & Guggino, 1997; Welsh *et al*. 2000). Therefore, it is of use to understand the intragenic distribution of mutations within a disease-associated gene. This information may

*Address for Correspondence: Sudhir Kumar, Ph.D., School of Life Sciences, Arizona State University, Tempe, AZ 85287-4501, USA. Tel: 480-727-6949. Fax: 480-965-2519. E-mail: S.KUMAR@ASU.EDU

reveal regions that have the greatest influence on the development of a given disease phenotype, and identify regions that are likely to be over- or under-represented in a disease mutation data set.

The objective of this project was to quantify and test two hypotheses about the intragenic distribution of human disease mutations. First, we describe a test of the null hypothesis that disease mutations are distributed uniformly among regions within a gene and therefore reflect a random mutational process. Second, we describe a test of the null hypothesis that disease mutations are distributed among gene regions according to distributions derived from an understanding of the proportion of evolutionarily conserved and variable amino acid residues (among species) within and among domains. Recent literature clearly underscores the importance of understanding human disease mutations from an evolutionary perspective (Botstein & Risch, 2003). Therefore, this second test is necessary because there is a known overabundance of disease mutations at evolutionarily conserved amino acid sites (Greenblatt *et al.* 2003; Miller & Kumar, 2001; Mooney & Klein, 2002; Notaro *et al.* 2000), thus illustrating the importance of conserved residues in the proper functioning of protein products.

Our two tests were applied to sets of disease mutations found in 6 different disease genes: the cystic fibrosis transmembrane conductance regulator (*CFTR*), glucose-6-phosphate dehydrogenase (*G6PD*), phenylalenine hydroxylase (*PAH*), paired box 6 (*PAX6*), the X-linked retinoschisis gene (*RS1*), and a gene associated with the development of tuberous sclerosis (*TSC2*).

Data Acquisition

We obtained data for unique disease-associated single base pair replacement mutations observed in 6 different disease genes from on-line databases (Table 1), and further obtained human and orthologous metazoan cDNA sequences for each gene (Fig. 1). Databases for *CFTR*

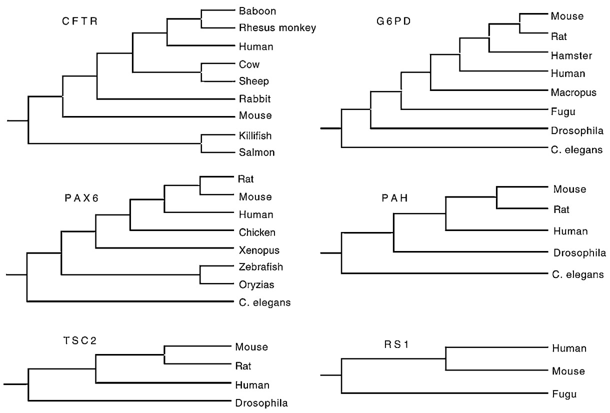

Figure 1 Model phylogenetic trees used to determine the number of amino acid substitutions that have occurred at each site in each gene throughout evolutionary history. GenBank accession numbers for sequences used in analyses are as follows: *CFTR*: human (NM_000492), baboon (AF162401), rhesus monkey (AF013753), cow (M76128), sheep (U20418), rabbit (U40227), mouse (M69298), killifish (AF000271), salmon (AF155237); *G6PD*: human (NM_000402), mouse (Z11911), rat (NM_017006), hamster (AF044676), *Macropus* (U13899), pufferfish (X83611), *Drosophila* (AH002543), *C. elegans* (Z73102); *PAH*: human (K03020), mouse (X51942), rat (NM_012619), *Drosophila* (M32802), *C. elegans* (AF119388); *PAX6*: human (12736585), mouse (X63963), rat (NM_013001), chicken (D87837), *Xenopus* (U77532), zebrafish (AF061252), *Oryzias* (AJ000938), *C. elegans* (U31537); *RS1*: human (AF014459), mouse (NM_011302), *Fugu* (AF146687); *TSC2*: human (X75621), mouse (NM_011647), rat (D50413), *Drosophila* (AF172995).

Table 1 Disease genes examined, numbers of mutations analyzed, and sources of the definitions of gene regions used for analyses

Disease gene	of mutations analyzed (disease/polymorphic/silent)	Web address of mutation database	Source for definitions of gene regions
CFTR	436/32/61	www.genet.sickkids.on.ca/cftr	Bianchet *et al.* (1997)
G6PD	110[a]/-/-	rialto.com/favism/mutat.htm[b]	Au *et al.* (2000)
PAH	270/-/-	www.mcgill.ca/pahdb/[c]	Erlandsen & Stevens (1999)
PAX6	29/-/-	www.hgu.mrc.ac.uk/Softdata/PAX6/[d]	Prosser & van Heyningen (1998)
RS1	71/-/-	www.dmd.nl/rs/rs.html	http://www.dmd.nl/rs/rshome.html
TSC2	47/18/33	expmed.bwh.harvard.edu/ts/	Maheshwar *et al.* (1997)

[a] 48 type I mutations, 62 types II, III, and IV mutations.

[b] Vulliamy *et al.* (1997).

[c] Scriver *et al.* (2000).

[d] Brown *et al.* (1998).

and *TSC2* also contained sufficient quantities of individual polymorphic replacement (presumably not disease associated) and silent (those not altering the encoded amino acid) mutations for use in separate analyses. In total, we analyzed 963 disease-associated replacement mutations, 50 polymorphic replacement mutations, and 94 silent mutations (Table 1). Furthermore, in the case of *G6PD*, information on the severity of the disease phenotype (Vulliamy *et al.* 1997) permitted us to separately analyze sets of severe (type I) mutations resulting in chronic non-spherocytic hemolytic anemia and milder (type II, III, and IV) mutations resulting only in enzyme deficiencies. Definitions of the specific amino acid residues contained in domains of each gene were obtained from the sources listed in Table 1 (see also Methods in Miller & Kumar, 2001). Not all the mutation databases examined contained information on the observation frequencies of a given mutation. Therefore, we included each mutation only once in our analyses to ensure that comparable data were present for each gene. In addition, the elimination of frequency information prevented bias in our results towards the properties of commonly observed mutations over those less frequently reported for a specific genetic disease.

Statistical Methods and Analysis Results

Testing for a Uniform Distribution of Mutations Among Gene Regions

In this analysis we tested the null hypothesis that sets of disease mutations are uniformly distributed among gene regions. This null hypothesis is based on the assumption that point mutations occur at random throughout disease genes. Since gene regions vary in coding sequence lengths and because the number of nucleotide sites that experience replacement mutations is much higher than those for silent (synonymous) mutations due to the properties of the genetic code, it is necessary to account for these factors when comparing numbers of observed mutations among domains. Thus, for each gene, we first calculated the number of potential replacement sites, R_j, for each of the j domains of the human gene sequence and $R = \Sigma R_j$, the total number of replacement sites in the gene. A convenient measure of R_j can be obtained using the method of Nei & Gojobori (1986), which accounts for both the length of the gene region and the mutability of codons within the region. If gene region j contains R_j replacement sites, then on average, we expect to observe the fraction R_j/R of the D total human mutations within region j. Therefore, the expected number of human mutations in region j under the assumption of an underlying uniform (on replacement sites) mutational process is

$$D_j^{\text{expected}} = (R_j/R) \times D, \tag{1}$$

where $D = \Sigma D_j^{\text{observed}}$ and D_j^{observed} is the observed number of mutations within the jth gene region. A global test of this null hypothesis can be performed by relating D_j^{observed} and D_j^{expected} as

$$X^2 = \sum_j \left(D_j^{\text{observed}} - D_j^{\text{expected}}\right)^2 / D_j^{\text{expected}} \tag{2}$$

and testing the X^2 statistic using a chi-square distribution with j-1 degrees of freedom. However, in the course of conducting this study, we frequently encountered situations where small expected counts for a domain appeared to overly influence the analysis outcome. Small expected counts can artificially inflate X^2 relative to the given underlying degrees of freedom for the analysis and result in the liberal rejection of the null hypothesis (Sokal & Rohlf, 1995). Therefore, we relied on a randomization procedure to evaluate the significance of the X^2 statistic. Here, we randomly (under a uniform distribution) allocated mutations among sites in the gene and quantified the global deviation of the randomized values from expected as

$$X^2_{\text{RND}} = \sum_j \left(D_j^{\text{RND}} - D_j^{\text{expected}}\right)^2 / D_j^{\text{expected}}, \tag{3}$$

where D_j^{RND} is the number of randomly allocated mutations to region j. This process generates a simulated empirical null (H_0) distribution of X^2, which approximates the true distribution. The P-value for this global randomization test is the proportion of randomization replicates where $X^2_{\text{RND}} \geq X^2_{\text{Observed}}$. Our use of this randomization-based procedure produced results that are more conservative in rejecting the null hypothesis than the asymptotic use of the chi-square distribution (results not shown). Therefore results from global randomization tests conducted with 10,000 replicates are presented throughout the remainder of this paper. Furthermore, in the course of performing this global

Table 2 Testing null hypotheses concerning the distribution of human mutations among functional domains

	Reject null hypothesis?	
Data set	Uniform distribution	Evolutionary-influenced distribution
CFTR		
Disease replacement	Yes ($X^2 = 69.10$, $P < 0.001$)	Yes ($X^2 = 50.76$, $P < 0.001$)
Polymorphic replacement	No ($X^2 = 12.02$, $P = 0.160$)	n/a
Silent	No ($X^2 = 5.15$, $P = 0.730$)	n/a
TSC2		
Disease replacement	Yes ($X^2 = 9.13$, $P = 0.020$)	Yes ($X^2 = 10.955$, $P = 0.007$)
Polymorphic replacement	No ($X^2 = 2.01$, $P = 0.372$)	n/a
Silent	No ($X^2 = 0.91$, $P = 0.664$)	n/a
G6PD replacement		
Type I	No ($X^2 = 5.198$, $P = 0.139$)	No ($X^2 = 2.66$, $P = 0.232$)
Type II, III, IV	No ($X^2 = 3.668$, $P = 0.218$)	n/a
PAH		
Disease replacement	Yes ($X^2 = 33.46$, $P < 0.001$)	Yes ($X^2 = 11.66$, $P = 0.002$)
PAX6		
Disease replacement	Yes ($X^2 = 24.244$, $P < 0.001$)	Yes ($X^2 = 12.58$, $P = 0.003$)
RS1		
Disease replacement	Yes ($X^2 = 15.20$, $P = 0.002$)	No ($X^2 = 3.53$, $P = 0.060$)

randomization test, we could also apply a domain-specific test to identify the specific gene regions that deviate from expectations. Details of this randomization procedure are given in Appendix I.

The test described above can also be performed for polymorphic replacement mutations and silent mutations observed in databases. In the case of the latter, it is then appropriate to substitute S and S_j, the number of silent sites in the entire gene, and each gene region, for R and R_j in the above calculations.

An important implicit assumption of the analysis described above is that the data used in analyses are obtained from the complete genotyping of each disease gene in question. Indeed, if mutation database contributors routinely only examine specific regions of the genes in question, then significant deviations from uniform expectations may result. However we note, in the case of *TSC2* and *CFTR*, that databases contain sufficient numbers of silent mutations for analyses (Table 1). Thus, assuming that all silent mutations detected by researchers are submitted to databases for these genes, we expect to observe uniform distributions of mutations among gene regions if there is no significant ascertainment bias due to the unequal screening of gene regions. Although this type of confirmatory analysis was not possible with data from the remaining four genes, our analyses of these data nonetheless produced clear results that are consistent with those observed in *CFTR* and *TSC2* (see analysis results below). In addition, our analyses excluded mutation frequency data, which minimizes the effects of any ascertainment bias that might exist for commonly observed variants (see Miller & Kumar, 2001).

Analysis Results: In global tests of the uniform distribution hypothesis, the sets of polymorphic and silent mutations from *CFTR* and *TSC2* did not reject the null hypothesis (Table 2). This suggests that these sets of mutations are randomly distributed among gene regions and, in the case of silent mutations discovered in *CFTR* and *TSC2*, provides evidence for the presence of low (or no) ascertainment bias in the data sets. However, while the global analysis did not reject the null hypothesis for *CFTR*, the domain-specific test indicated that there were more polymorphic replacement mutations observed in the first nucleotide-binding domain (NBD1) of the gene than expected by chance alone (Fig. 2, $P = 0.026$). An explanation for this observation is shown in Table 3. We found that when NBD1 is treated separately and observed and expected values from the other domains are pooled for analyses, that there is in fact a significant overabundance of polymorphic replacement mutations in this gene region. Often, when analyzing such data and an overall significant effect is found, it is then common procedure to pool sets of test categories to determine which are responsible for the significant deviation (Sokal & Rohlf, 1995). Thus, the randomization tests described here have the ability to identify gene regions that, if analyzed in comparison with pooled observations from the other regions of the gene, would result in a significant X^2 value when evaluated with fewer degrees of freedom.

The analysis of disease-associated replacement mutations revealed that most sets deviate from uniform expectations, indicating that disease mutations are overabundant in some domains and underrepresented in others. X^2 values from global tests were highly significant for *CFTR*, *TSC2*, *PAH*, *PAX6*, and *RS1* (Table 2), and use of the domain-specific randomization procedure showed that multiple gene regions deviated from uniform expectations (Fig. 3). Interestingly, global analysis of the set of less severe *G6PD* mutations (types II, III, and IV) did not reject the null hypothesis, nor did the domain-specific randomization procedure identify any gene regions with observed counts different from uniform expectations (Fig. 3c). Likewise, the global analysis of *G6PD* type I mutations suggested that these amino acid changes were uniformly distributed among gene regions (Table 2). However, domain-specific tests did identify the $\beta(\alpha$ gene region as having significantly more mutations than expected by chance alone ($P = 0.026$, Fig. 3b; see also Table 3).

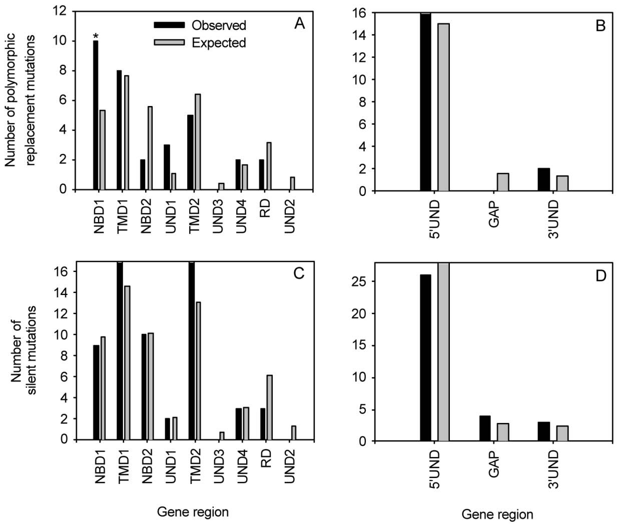

Figure 2 Observed (black bars) and expected (grey bars) numbers of human mutations for (A) polymorphic replacement mutations in *CFTR*, (B) polymorphic replacement mutations in *TSC2*, (C) silent mutations in *CFTR*, and (D) silent mutations in *TSC2*. Expected values were obtained under the null hypothesis that mutations are uniformly distributed among gene regions. See Fig. 3 caption for definitions of gene region name abbreviations. Gene regions are ordered on the abscissa from most conserved (left side) to least conserved (right side). Asterisks indicate observed counts for gene regions that were found to significantly differ from expected values derived under the null hypothesis of a uniform distribution of mutations among gene regions ($P < 0.05$).

Testing for the Distribution of Human Mutations using the Intragenic Distribution of Evolutionarily Conserved and Variable Amino Acid Residues

It has been shown previously that disease mutations are generally not distributed at random among amino acid sites within genes (for example, Botstein & Risch, 2003; Greenblatt *et al.* 2003; Miller & Kumar, 2001; Mooney & Klein, 2002; Notaro *et al.* 2000). Instead, disease-causing mutations are generally overabundant at evolutionarily conserved sites. Such patterns illustrate the importance of those conserved residues for the proper function of the protein product, as amino acid-altering mutations at these sites result in

Table 3 Examples of domain-specific randomization test results that gave significant *P*-values when the null hypothesis was not rejected in the global analysis

Gene region[a]	Observed	Expected	(Observed-Expected)^2/Expected
CFTR			
NBD1 only	10	5.319	4.118
Rest of the domains	22	26.681	0.821
Total	32	32	$X^2 = 4.94$, 1 df, $P = 0.026$
G6PD			
$\beta{+}\alpha$ domain	36	29.400	1.482
Rest of the domains	12	18.600	2.341
Total	48	48	$X^2 = 3.82$, 1 df, $P = 0.0505$

[a] ~~The gene region with the significant *P*-value from domain-specific analysis was treated~~ separately, while observed and expected counts for the other regions were pooled. Thus, in these cases, the analyses give significant or nearly significant results when X^2 values were considered with fewer degrees of freedom.

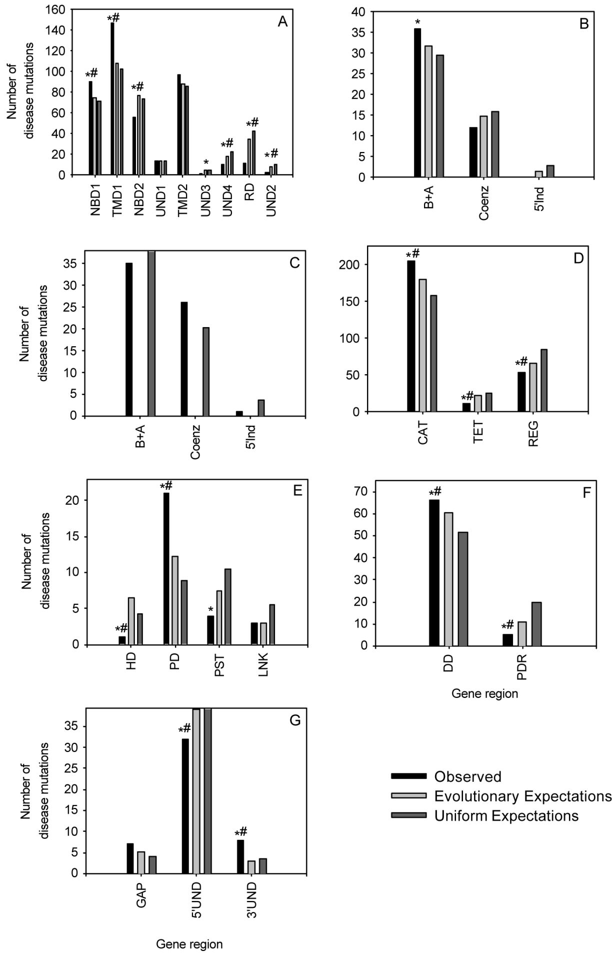

Figure 3 Observed and expected numbers of disease associated mutations in each region of the six genes examined: (A) *CFTR*, (B) *G6PD* type I mutations, (C) *G6PD* type II, III, and IV mutations, (D) *PAH*, (E) *PAX6*, (F) *RS1*, (G) *TSC2*. Black bars indicate the observed number of human mutations in each gene region, light grey bars indicate the expected number of human mutations based on an evolutionarily influence model, and dark grey bars indicate expected numbers assuming a uniform distribution among gene regions. Expected values under the evolutionarily influenced model were not obtained for *G6PD* type II, III, and IV mutations (see text for details). Abbreviations of gene region names are as follows: *CFTR*: TMD1 (transmembrane domain 1), TMD2 (transmembrane domain 2), NBD1 (nucleotide binding domain 1), NBD2 (nucleotide binding domain 2), RD (R-domain), UND1-UND4 (four independent undesignated regions); *G6PD*: 5′ Ind (5′ undesignated region), Coenz (coenzyme domain), B + A (β+α domain); *PAH*: TET (tetramerization domain), REG (regulatory domain), CAT (catalytic domain); *PAX6*: HD (homeodomain), PD (paired domain), PST (proline-serine-threonine domain), LNK (link domain); *RS1*: PDR (pre-discoidin region), DD (discoidin domain); *TSC2*: 5′ UND (5′ undesignated region), GAP (GAP-like domain), 3′ UND (3′ undesignated region). Gene regions are ordered on the abscissa from most conserved (left side) to least conserved

Figure 3 (continued) (right side). Asterisks indicate observed counts for gene regions that were found to differ significantly from expected values derived under the null hypothesis of a uniform distribution of mutations among gene regions ($P < 0.05$). # and ∗ indicate observed counts that significantly differ from expected values derived under the null hypothesis that mutations are distributed based on evolutionary expectations, or are uniformly distributed, respectively ($P < 0.05$).

phenotypic changes of sufficient severity to come to the attention of clinicians.

In the analysis that follows, we consider the possibility that if disease mutations as a whole are overabundant at conserved sites, then deviations of disease mutations from uniform/random expectations may be due to the fact that some regions of the gene have been differentially conserved throughout long-term evolutionary history. Thus, in this analysis, we are testing the null hypothesis that the observed disease mutations are distributed among gene regions relative to the abundance of evolutionarily conserved and variable amino acid residues in each gene region.

Consider an amino acid sequence alignment of length *A*. Through the use of the algorithm of Fitch (1974), we can obtain for each site in the gene an estimate of the minimum number of amino acid substitutions, *i*, that have occurred throughout evolutionary history given a known phylogenetic tree for all of the sequences in the alignment (Fig. 1). This procedure permits us to quantify the amino acid variability at a site while simultaneously accounting for the shared ancestral amino acid substitutions that appear within descendent phylogenetic lineages, i.e., the statistical non-independence of the interspecific sequence data used for analyses; (Felsenstein, 1985). It is important to estimate *i* for each site using an alignment containing as many species as possible, and including divergent species (e.g., distantly related vertebrates or animals). This is to ensure that a sufficient amount of evolutionary variability exists at the amino acid level to conduct a powerful test (see below). For the six genes examined in this study, these parsimony-based estimates of amino acid site rate variability were highly correlated with comparable maximum-likelihood based estimates (r = 0.987; Miller & Kumar, 2001). Using this information, we classified amino acid sites based on their amino acid variability, and obtained counts of sites of type i in the gene, thus obtaining an estimated frequency of invariant sites (a_0: type 0 sites), once-changed sites (a_1: type 1 sites), and sites that have changed i times (a_i: type i sites) over the course of evolutionary history. We then mapped the positions of the D human disease mutations to the interspecific alignment and obtained counts of the number of human disease mutations that occur at each type of amino acid site. Therefore, there will be D_0^{observed} human mutations among the a_0 type 0 sites, D_1^{observed} human mutations among the a_1 type 1 sites, and D_i^{observed} human mutations among the a_i type i sites, where $D = \sum_0^i D_i^{\text{observed}}$. Of the six disease mutation data sets examined in this study, all (with the exception of type II, III, and IV *G6PD* variants) show a significant overabundance of disease mutations at amino acid sites that were perfectly conserved (i.e., type 0 sites) among the taxa used to generate interspecific alignments (Fig. 1., Miller & Kumar, 2001).

Next, within each gene region j, we tabulated the estimated number of sites, a_{ij}, that have undergone i amino acid substitutions based on the phylogenetic analysis of the interspecific sequence alignments described above. Since gene region j contains a_{ij} of the total a_i sites that have experienced i substitutions, we expected gene region j to have $(a_{ij}/a_i) \times D_i^{\text{observed}}$ of the D_i^{observed} mutations found at sites that have evolved i times based on the analysis of interspecific data. Consequently, under the null hypothesis that the relative importance of conserved versus variable amino acid sites is equivalent among gene regions, we have the expected value

$$D_j^{\text{expected}} = \sum_0^i \left((a_{ij}/a_i) \times D_i^{\text{observed}}\right) \tag{4}$$

of total disease mutations for gene region j. Thus, we can relate D_j^{expected} and D_j^{observed} in a global test for deviation from the evolutionary expectations using equation 2 above. However, there are two types of problems that arise. First, because the proportion of disease mutations found at each rate variability category is estimated from the data, the distribution of the test statistic (X^2) becomes more spread out than the actual chi-square distribution (Greenwood & Nikulin, 1996; Sokal & Rohlf, 1995). Therefore it is inappropriate simply to perform an asymptotic chi-squared test with j-1 degrees of freedom. Furthermore, as with the test for a uniform distribution of mutations among domains, we also frequently encountered situations where domains have small-expected counts. Therefore, we can perform global and domain-specific randomization tests in lieu of the conventional global chi-square analysis described above. Details of these analyses are presented in Appendix II.

Since we have modified expectations for each gene region to reflect a potential evolutionary influence on distributions, there is no need to perform this analysis if, when using the site-by-site analysis procedure (i.e., Miller & Kumar, 2001), no overall excess of mutations is detected at conserved amino acid residues. In these cases, use of the evolutionary-influence analysis described above will produce expected values similar to those obtained from the test for a uniform distribution of mutations among gene regions (equation 2). Therefore, this analysis was not performed on silent and non-disease associated mutations in *CFTR*, *TSC2*, or on types II, III, and IV mutations in *G6PD*, which have already been shown to be randomly distributed among individual amino acid sites with respect to the level of conservation of each residue over long-term evolutionary history (Miller & Kumar, 2001).

Relationship Between Expected Values Obtained from Uniform and Evolutionary Models

While the statistical analyses described above were explicitly performed on observed and expected counts of mutations from each gene region, it is conceptually easier to visualize the relationship between the expected values used in each test by expressing expected counts for a domain as a proportion of the number of potential replacement sites within the gene region. Here, we calculated the number of replacement sites (Nei & Gojobori, 1986) in the gene region from the reference human sequence to account for variation in the length and mutability of coding regions for each functional domain. Fig. 4 shows the relationship between expected values from each test, normalized by the number of replacement sites in each gene region, and the average number of amino acid substitutions per amino acid site within different *CFTR* gene regions over long-term evolutionary history (a convenient measure of the average degree of conservation of amino acid residues within

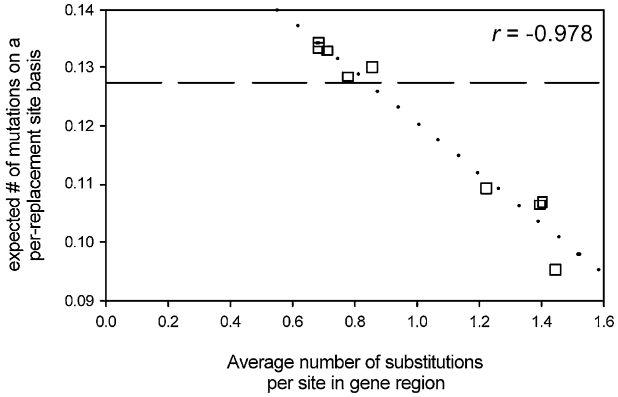

Figure 4 Scatter-plot illustrating the relationship between expected values obtained from the test for a uniform distribution (dashed line) and evolutionarily-influenced distribution (squares indicate actual expected values and the dotted line is the best fit regression line through those points). Expected values are plotted as a proportion of the number of replacement sites within the gene region to account for variation in the mutability and length of each region. In the test for a uniform distribution among gene regions, expected values are equivalent on a per-replacement site basis in all of the gene regions. In contrast, expected values for the evolutionary-influenced distribution test decrease in less conserved gene regions to account for the overabundance of disease mutations at conserved sites. The expected values shown here were obtained for *CFTR* disease-associated mutations.

a domain). On a per replacement site basis, we expected to observe equal numbers of mutations in each gene region (dashed line) under the null hypothesis of a uniform distribution of disease mutations. In contrast, if expected counts are modified to account for an overabundance of disease mutations at invariant sites, we expected to see more disease mutations within conserved gene regions as opposed to those that have a greater proportion of sites that vary among species (dotted line). Based on these factors, the correlation between the average number of amino acid substitutions experienced per site in a gene region throughout evolutionary history and the expected numbers of disease mutations for a given gene region based on evolutionary modifications (computed using equation 4) was high for *CFTR* ($r = -0.979$) (Fig. 4; open square). A similar result was obtained for the remaining genes examined (*G6PD* type I mutations: $r = -0.995$; *PAH*: $r = -0.999$; *PAX6*: $r = -0.998$; *RS1*: $r = -1.000$; *TSC2*: $r = -0.999$).

Analysis Results: In global tests for an evolutionarily-influenced distribution, the null hypothesis was rejected for the sets of disease mutations in *CFTR*, *TSC2*, *PAH*, and *PAX6* (Table 2, Fig. 3). In contrast, the null hypothesis was not rejected for *G6PD* type I mutations or *RS1* (Table 2, Fig. 3), indicating that mutational distributions among domains were highly influenced by the level of conservation of sites within those genes. However, the *P*-value in the case of *RS1* was approximately significant ($X^2 = 3.53$, P = 0.060) at the 5% level. When the randomization-based procedure was used to identify gene regions that significantly deviate from evolutionary expectations, the test indicated that in all cases, except for *G6PD* type I mutations, there were one or more regions that significantly deviated from expected values (Fig. 3). Thus, while the global test for *RS1* did not produce a significant result at the 0.05 level (Table 2), the randomization procedure in fact suggested that each gene region deviated from expectations ($P = 0.026$), with the pre-discoidin region (PDR) showing a deficit of mutations, and the discoidin domain (DD) containing more mutations than expected (Fig. 3).

Discussion

With the exception of the two sets of *G6PD* mutations, global analyses of all remaining sets of disease mutations rejected the null hypothesis of a uniform distribution among gene regions (Table 2). However, the domain-specific randomization procedure suggested that the conserved $\beta+\alpha$ region of *G6PD* has more mutations than expected under the uniform distribution model (Fig. 3). Thus, our combined analyses indicated that the observed distributions of disease mutations (except for types II, III, and IV *G6PD* mutations) are not simply a set of randomly occurring mutations. Furthermore, global analyses rejected the null hypothesis of an evolutionarily influenced distribution for the sets of disease mutations found in *CFTR*, *TSC2*, *PAH* and *PAX6* (Table 2), and use of the domain specific randomization procedure indicated that both gene regions of *RS1* in fact deviated from evolutionary expectations (Fig. 3). Although statistical analyses generally rejected the evolutionarily influenced distribution hypothesis, with the exception of *TSC2*, X^2 values for the global tests of an evolutionarily influenced distribution were much lower than for the uniform distribution model (Table 2), and indicated a better fit of the data to the evolutionary model. To illustrate this, Fig. 5

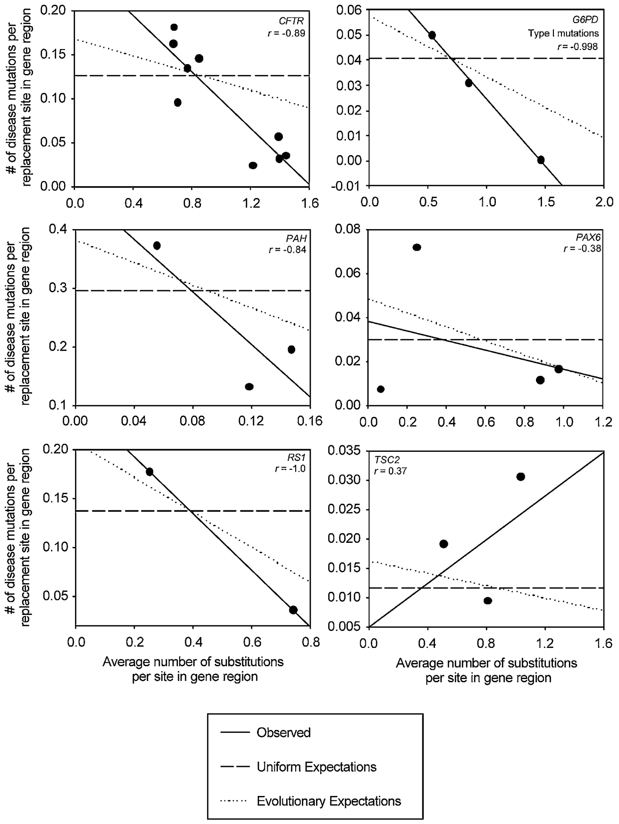

Figure 5 Scatter plots illustrating the relationship between the observed numbers of disease mutations found in a gene region (expressed on a per-replacement site basis) and the average number of amino acid substitutions per site. Circles indicate actual data points while the solid line is the best-fit regression line through those points. The dashed lines and dotted lines indicate expected values (also expressed on a per-replacement site basis) obtained under the assumptions of uniform and evolutionarily-influenced distributions, respectively.

shows the relationship between the observed number of disease mutations (expressed as a proportion of the number of replacement sites) and the average number of amino acid substitutions per site within a gene region for the six data sets that were appropriate for analysis with the evolutionary-influence model. Except for *TSC2*, the emergent pattern was consistent with expectations from the evolutionary-influence model, as there are more mutations on a per-replacement site basis within conserved regions than highly variable ones. This suggests that the use of an evolutionary perspective can be an important first step towards understanding the intragenic distribution of disease mutations.

Examination of Figs 3 and 5 reveals a common trend among the majority of the genes analyzed. Namely, in the cases of *CFTR*, *G6PD* type I mutations, *PAH*, and *RS1*, there is a strong tendency for the observed number of mutations to be greater than evolutionary expectations in conserved domains (left side of X-axes in Figs 3 and 5), while the observed number of mutations in less conserved regions (right side of X-axes in Figs 3 and 5) tends to be lower than values obtained based on the evolutionary model. These analyses provide evidence that the relative importance of conserved and variable sites for the development of disease is not comparable among domains. More specifically, our results indicate that the numbers of evolutionary conserved and variable amino acid residues within the gene region non-additively influence the number of disease mutations found within a domian. Perhaps, of the full complement of random mutations that may possibly be identified in a gene, mostly those found in conserved gene regions produce disease phenotypes. Replacement mutations in less-conserved regions, despite containing some fully conserved residues, may have a less severe effect on an individual's phenotype, and therefore not be detected in routine studies of disease patients.

An exception to this pattern is apparent from the analysis of *PAX6*. This gene is an important regulator of the development of the central nervous system (Callaerts *et al.* 1997; Prosser & van Heyningen, 1998), and missense mutations in the homeodomain (HD), a DNA binding region, are significantly underrepresented relative to both uniform and evolutionary-influenced expectations (Figs 3E and 5). This gene region, however, is the most conserved of all of the domains analyzed in this study, displaying only four variable amino acid sites among eight species ranging from humans to *C. elegans*; interspecific variation was observed only in the most distantly related species examined. The apparent deficit of mutations in this region may reflect the critical function of the domain, as replacement mutations to this region may in fact be lethal and rarely come to the attention of clinicians. Alternately, this finding may suggest that replacement mutations in the HD do not produce the degenerative eye disorders typically attributed to *PAX6* mutations (Prosser & van Heyningen, 1998), and instead produce other deleterious disease phenotypes that have not yet been extensively studied.

Our analysis of disease-associated *TSC2* mutations yielded results unlike those for any of the other genes examined (Table 2, Fig. 3). Here, use of the domain-specific randomization procedure suggested that the observed number of mutations in the highly conserved GAP domain was not significantly different from either uniform or evolutionary expectations. However, counts of mutations in the less conserved 5′ and 3′ undesignated regions were significantly different from expected values under both models (Fig. 3G). Interestingly, based on the overall X^2-values from global tests, the evolutionary-based model was not a better descriptor of the distribution of mutations among regions of this gene. The exceptional pattern observed in this gene relative to the others examined may be due to the fact that, currently only a single functional domain has been identified in *TSC2* and, overall very little is known about the true function of the full gene product (Cheadle *et al.* 2000). However, comparative sequence analysis of human and pufferfish *TSC2* sequences has revealed additional regions of high conservation (Maheshwar *et al.* 1997), suggesting the presence of additional functional domains. At this time, however, no known homologies to other functional domains have been found (Cheadle *et al.* 2000). Therefore, we grouped all undesignated amino acids at the 5′ and 3′ ends of the GAP-like domain to create the three gene regions used for analysis purposes. Future re-analysis of the *TSC2* mutation data from this region-by-region perspective may reveal patterns more similar to those observed in the other genes examined, as more detailed information on functional regions of the gene becomes available.

Results of global analyses indicated that the sets of polymorphic replacement and silent mutations in *CFTR* and *TSC2* generally fit the null hypothesis of a uniform distribution among domains (Table 2, Fig. 2). Likewise, less severe type II, III, and IV mutations in *G6PD*, which are not overabundant at conserved sites (Miller & Kumar, 2001), were uniformly distributed among gene regions (Table 2, Fig. 3c). It should be noted that the biological significance of the uniform distribution is not equivalent for both types of mutations. In the case of silent mutations, the uniform distribution reflects the random mutational process within the gene, as these types of mutations may occur in any location without altering an individual's phenotype. However, the uniform distribution of polymorphic replacement mutations may point to the presence of slightly deleterious mutations within human populations. Since gene regions that have varied little throughout evolutionary history provide evidence for domains that are most critical for proper protein function, presumably only replacement mutations found in variable gene regions are likely to be tolerated due to relaxed selective constraints. This pattern may possibly be due to the fact that observation frequencies of these polymorphic mutations were not available from databases. Perhaps, of the human polymorphic variants, those mutations found in the less conserved gene regions are far more common than those seen in conserved domains.

While the global analysis indicated that polymorphic replacement mutations in *CFTR* were uniformly distributed, our use of domain-specific randomization tests suggested that the *CFTR* database contained a slight overabundance of polymorphic mutations in the most conserved domain of that gene (NBD1; Fig. 2a). This gene region also contains the most commonly observed disease-associated *CFTR* mutation, ΔF508 (Welsh *et al.* 2000). If large numbers of studies of non-affected individuals have focused solely on exons containing this mutation, then replacement mutations in this functional domain may in fact be over-represented in databases as a result of unequal screening of gene regions. However, assuming diligent reporting of all silent mutations discovered in this gene during such screenings, we would also expect to observe more silent mutations within NBD1 as well, which was clearly not the case (Fig. 2c). Alternately, the slight overabundance of polymorphic replacement mutations in NBD1 may be due to Darwinian natural selection. For example, Pier *et al.* (1998) suggested that there may be a selective advantage for ΔF508 in terms of an increased resistance to infection by *Salmonella typhi*. Further, it has been suggested that individuals heterozygous for ΔF508 may have a generalized resistance to diarrheal diseases (Baxter *et al.* 1988; Guggino, 1999) or reduced incidence of bronchial asthma (Schroeder *et al.* 1995). Possibly some of the amino acid-altering polymorphic mutations found in NBD1 have selective advantages. Such benefits would cause the frequencies of these allelic variants to increase in populations faster than neutral amino acid changes, and hence they would be observed more often in routine *CFTR* genotyping studies.

Acknowledgements

We thank Dr. Michael Rosenberg and Jacob Reidhead for comments on this manuscript and Drs. Priscilla E. Greenwood and Susan Durham for statistical advice. This research was supported by grants from the National Science Foundation, National Institute of Health, and Burroughs Wellcome Fund to S.K.

Appendix I

A Randomization Test for Identifying Gene Regions that Contain Different Numbers of Mutations than Expected Under the Null Hypothesis of a Uniform Distribution Among Domains

A simple randomization procedure can be used to identify specific gene regions that deviate from uniform expectations. An algorithm for performing this analysis is as follows:

1. Obtain a test statistic for gene region j, which can be simply calculated as the difference between the observed and expected values for region j. Thus, the test statistic for region j is $T_j = D_j^{\text{observed}} - D_j^{\text{expected}}$, with D_j^{observed} being the observed number of mutations in region j and D_j^{expected} calculated as in equation 1.
2. Initialize a counter for each gene region, $K_j = 0$.
3. Create an index vector, $\boldsymbol{X}$, of length A, where A is the number of amino acid sites in the entire gene. Within this array, associate a_1 positions to gene region 1, a_2 positions to gene region $2, \ldots,$ where a_j is the number of amino acid sites within region j. A diagrammatic representation of $\boldsymbol{X}$ is presented in Fig. 6.
4. For each of the specified number of randomization replicates m ($m = 1{,}000$ or more) carry out the following:
   (i) Using a uniform random number generator, randomly assign D mutations to the elements of $\boldsymbol{X}$, where D is the total number of mutations in the data set.
   (ii) Calculate D_j^{RND}, the number of randomly assigned mutations to each region j, by recording the number of mutations found in the appropriate a_j elements of $\boldsymbol{X}$.

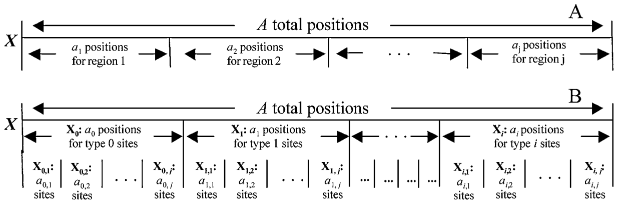
A
A total positions
X
a_1 positions for region 1
a_2 positions for region 2
. . .
a_j positions for region j
B
A total positions
X
$\mathbf{X}_0$: a_0 positions for type 0 sites
$\mathbf{X}_1$: a_1 positions for type 1 sites
. . .
$\mathbf{X}_i$: a_i positions for type i sites
$\mathbf{X}_{0,1}$: $a_{0,1}$ sites
$\mathbf{X}_{0,2}$: $a_{0,2}$ sites
. . .
$\mathbf{X}_{0,j}$: $a_{0,j}$ sites
$\mathbf{X}_{1,1}$: $a_{1,1}$ sites
$\mathbf{X}_{1,2}$: $a_{1,2}$ sites
. . .
$\mathbf{X}_{1,j}$: $a_{1,j}$ sites
...
...
...
...
$\mathbf{X}_{i,1}$: $a_{i,1}$ sites
$\mathbf{X}_{i,2}$: $a_{i,2}$ sites
. . .
$\mathbf{X}_{i,j}$: $a_{i,j}$ sites

Figure 6 Diagrammatic representations of the arrangement of positions within sampling vector $\boldsymbol{X}$, which was used to (A) determine which gene regions deviated from uniform expectations, and (B) determine which gene regions deviated from evolutionarily-influenced expectations. In panel A, $\boldsymbol{X}$ is subdivided into j regions (representing domains), each containing a_j of the A total amino acid residues in the gene being examined. In panel B, $\boldsymbol{X}$ is subdivided into each of i regions to reflect the sets of amino acid sites in the gene that have undergone i amino acid substitutions over the course of evolutionary history. Each $\boldsymbol{X}_i$ is further subdivided into j sub-regions denoting the abundance of type i sites in each of the j gene regions. See appendices for further details.

   (iii) Calculate new test statistics for each of the j gene regions as $T_j^{\text{RND}} = D_j^{\text{RND}} - D_j^{\text{expected}}$.
   (iv) For each gene region j, increment the counter K_j under the following conditions:

   $$\text{if } T_j > 0 \quad \text{and} \quad T_j^{\text{RND}} \geq T_j$$

   or

   $$\text{if } T_j < 0 \quad \text{and} \quad T_j^{\text{RND}} \leq T_j.$$

   Thus, this analysis is 1-tailed for each gene region, as we are estimating the probability of observing both differences and signs of observed and expected values as large or larger than random expectations.
5. The Monte-Carlo P-value for region j is $P_j = K_j/m$.

Appendix II

Randomization Tests for Evaluating Evolutionary-Influenced Distribution Hypotheses

As with the test for a uniform distribution of mutations among gene regions (Appendix I), we can perform a randomization test to determine which gene regions significantly deviate from evolutionary expectations. The analysis proceeds as follows:

1. Obtain a test statistic for each gene region, which is calculated as the difference between the observed and

expected values for region j. Thus, the test statistic for region j is $T_j = D_j^{\text{observed}} - D_j^{\text{expected}}$, with D_j^{observed} being the observed number of mutations in region j and D_j^{expected} calculated as in equation 4.

2. Initialize a counter for each gene region, $K_j = 0$.
3. Create an index vector $\boldsymbol{X}$ of length A, where A is the number of amino acid sites in the gene. Within this array, associate the first a_0 positions with sites of type 0, the next a_1 positions with sites of type $1, \ldots,$ and the final a_i positions with sites of type i, where a_i is the number of amino acid sites in the gene that have undergone i substitutions throughout evolutionary history. We will refer to the set of positions in $\boldsymbol{X}$ associated with sites of type i as $\mathbf{X_i}$. Next, further subdivide each $\mathbf{X_i}$ to reflect the numbers of type i sites found within region j. Thus, within $\mathbf{X_i}$, we indicate j separate sub-regions, each encompassing a_{ij} of the a_i total sites of type i found among the j gene regions. We refer to the set of positions in $\mathbf{X_i}$ associated with gene region j as $\mathbf{X_{ij}}$. A diagrammatic representation of $\boldsymbol{X}$ is given in Fig. 6b.
4. For each of the specified number of randomization replicates m ($m = 1{,}000$ or more) carry out the following:
   (i) Using a uniform random number generator, randomly assign D_0 mutations to region $\mathbf{X_0}$ of $\boldsymbol{X}$, D_1 mutations to region $\mathbf{X_1}$ of $\boldsymbol{X}, \ldots,$ and D_i mutations to region $\mathbf{X_i}$ of $\boldsymbol{X}$.
   (ii) Record the number of randomly assigned mutations to gene region j, D_j^{RND}, by noting the number of random mutations in $\boldsymbol{X}$ associated with region j.
   (iii) Calculate new test statistics for each of the j gene regions as $T_j^{\text{RND}} = D_j^{\text{RND}} - D_j^{\text{expected}}$, with D_j^{expected} calculated from equation 4.
   (iv) For each gene region j, increment K_j under the following conditions:

   $$\text{if } T_j > 0 \quad \text{and} \quad T_j^{\text{RND}} \geq T_j$$

   or

   $$\text{if } T_j < 0 \quad \text{and} \quad T_j^{\text{RND}} \leq T_j.$$
5. The Monte-Carlo P-value for region j is $P_j = (K_j)/m$.

As with the test for a uniform distribution of mutations among gene regions (Appendix I), this analysis is 1-tailed for each gene region, since we are estimating the probability of observing both differences and signs of observed and expected values as large or larger than evolutionary expectations. Furthermore, in the course of performing this region-by-region analysis we can also perform a global randomization test by calculating X^2_{RND} for each of the randomization replicates and evaluating the significance of X^2_{observed} in a manner similar to that described in the text.

References

Au, S. W., Gover, S., Lam, V. M. & Adams, M. J. (2000) Human glucose-6-phosphate dehydrogenase: the crystal structure reveals a structural NADP(+) molecule and provides insights into enzyme deficiency. *Structure Fold Des* **8**, 293–303.

Baxter, P. S., Goldhill, J., Hardcastle, J., Hardcastle, P. T. & Taylor, C. J. (1988) Accounting for cystic fibrosis. *Nature* **335**, 211.

Bianchet, M. A., Ko, Y. H., Amzel, L. M. & Pedersen, P. L. (1997) Modeling of nucleotide binding domains of ABC transporter proteins based on a F1-ATPase/recA topology: structural model of the nucleotide binding domains of the cystic fibrosis transmembrane conductance regulator (CFTR). *J Bioenerg Biomembr* **29**, 503–524.

Botstein, D. & Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33 Suppl**, 228–237.

Brown, A., McKie, M., van Heyningen, V. & Prosser, J. (1998) The Human PAX6 Mutation Database. *Nucleic Acids Res* **26**, 259–264.

Callaerts, P., Halder, G. & Gehring, W. J. (1997) PAX-6 in development and evolution. *Ann Rev Neurosci* **20**, 483–532.

Cheadle, J. P., Reeve, M. P., Sampson, J. R. & Kwiatkowski, D. J. (2000) Molecular genetic advances in tuberous sclerosis. *Hum Genet* **107**, 97–114.

Devidas, S. & Guggino, W. B. (1997) CFTR: domains, structure, and function. *J Bioenerg Biomembr* **29**, 443–451.

Erlandsen, H. & Stevens, R. C. (1999) The structural basis of phenylketonuria. *Mol Genet Metab* **68**, 103–125.

Felsenstein, J. (1985) Phylogenetics and the comparative Method. *Am Naturalist* **125**, 1–15.

Greenblatt, M. S., Beaudet, J. G., Gump, J. R., Godin, K. S., Trombley, L., Koh, J. & Bond, J. P. (2003) Detailed computational study of p53 and p16: using evolutionary sequence analysis and disease-associated mutations to predict the functional consequences of allelic variants. *Oncogene* **22**, 1150–1163.

Greenwood, P. E. & Nikulin, M. S. (1996) A guide to chi-squared testing, pp. 280, John Wiley & Sons, New York.

Guggino, S. E. (1999) Evolution of the delta F508 CFTR mutation. *Trends Microbiol* **7**, 55–56; discussion 56–58.

Jimenez-Sanchez, G., Childs, B. & Valle, D. (2001) Human disease genes. *Nature* **409**, 853–855.

Li, W. H., Gu, Z., Wang, H. & Nekrutenko, A. (2001) Evolutionary analyses of the human genome. *Nature* **409**, 847–849.

Maheshwar, M. M., Cheadle, J. P., Jones, A. C., Myring, J., Fryer, A. E., Harris, P. C. & Sampson, J. R. (1997) The GAP-related domain of tuberin, the product of the TSC2 gene, is a target for missense mutations in tuberous sclerosis. *Hum Mol Genet* **6**, 1991–1996.

Miller, M. P. & Kumar, S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet* **10**, 2319–2328.

Mooney, S. D. & Klein, T. E. (2002) The functional importance of disease-associated mutation. *BMC Bioinformatics* **3**, 24.

Nei, M. & Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418–426.

Notaro, R., Afolayan, A. & Luzzatto, L. (2000) Human mutations in glucose 6-phosphate dehydrogenase reflect evolutionary history. *FASEB J* **14**, 485–494.

Pier, G. B., Grout, M., Zaidi, T., Meluleni, G., Mueschenborn, S. S., Banting, G., Ratcliff, R., Evans, M. J. & Colledge, W. H. (1998) Salmonella typhi uses CFTR to enter intestinal epithelial cells. *Nature* **393**, 79–82.

Prosser, J. & van Heyningen, V. (1998) PAX6 mutations reviewed. *Hum Mutat* **11**, 93–108.

Schroeder, S. A., Gaughan, D. M. & Swift, M. (1995) Protection against bronchial asthma by CFTR delta F508 mutation: a heterozygote advantage in cystic fibrosis. *Nat Med* **1**, 703–705.

Scriver, C. R., Waters, P. J., Sarkissian, C., Ryan, S., Prevost, L., Cote, D., Novak, J., Teebi, S. & Nowacki, P. M. (2000) PAHdb: a locus-specific knowledgebase. *Hum Mutat* **15**, 99–104.

Sokal, R. R. & Rohlf, F. J. (1995) Biometry, W. H. Freeman and Company, New York.

Vulliamy, T., Luzzatto, L., Hirono, A. & Beutler, E. (1997) Hematologically important mutations: Glucose-6-Phosphate Dehydrogenase. *Blood Cells Mol Dis* **23**, 302–313.

Welsh, M. J., Ramsey, B. W., Accurso, F. & Cutting, G. R. (2000) Cystic Fibrosis. In: *The metabolic and molecular bases of inherited disease*, Vol. 1, (eds. C. R. Scriver, A. L. Beaudet, W. S. Sly & D. Valle), pp. 5121–5188, McGraw-Hill, New York.

Received: 7 February 2003
Accepted: 31 July 2003