Deep mutational scanning quantifies DNA binding and predicts clinical outcomes of PAX6 variants

Alexander F McDonnell[1], Marcin Plech[1], Benjamin J Livesey[1], Lukas Gerasimavicius[1], Liusaidh J Owen[1], Hildegard Nikki Hall[1], David R FitzPatrick[1], Joseph A Marsh[1], Grzegorz Kudla[1]

[1]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XU, UK.

ABSTRACT

Nonsense and missense mutations in the transcription factor PAX6 cause a wide range of eye development defects, including aniridia, microphthalmia and coloboma. To understand how changes of PAX6:DNA binding cause these phenotypes, we combined saturation mutagenesis of the paired domain of PAX6 with a yeast one-hybrid (Y1H) assay in which expression of a PAX6-GAL4 fusion gene drives antibiotic resistance. We quantified binding of more than 2,700 single amino-acid variants to two DNA sequence elements. Mutations in DNA-facing residues of the N-terminal subdomain and linker region were particularly detrimental, as were mutations to prolines and to negatively charged residues. Many variants caused molecular gain-of-function effects, including variants in position Ile71 that increased binding to the LE9 enhancer but decreased binding to a SELEX-derived binding site. Benchmarking against known patient variants and applying ACMG/AMP guidelines to variant classification, we obtained supporting to moderate evidence to suggest that 1,306 variants are likely benign, and 977, likely pathogenic. Our analysis shows that most pathogenic mutations in the paired domain of PAX6 can be explained simply by the effects of these mutations on PAX6:DNA association, and establishes Y1H as a generalisable assay for the interpretation of variant effects in transcription factors.

INTRODUCTION

Transcription factors (TFs) are one of the most abundant human gene families with approximately 1,600 genes[1], and are key for development, signal transduction and cellular housekeeping[2-4]. They act by selectively binding DNA elements and facilitating or inhibiting transcription machinery recruitment onto the promoter regions of target genes, modulating their expression[5]. Many TFs sit atop gene regulatory networks and function as master regulators of development by coordinating cellular processes that include stem cell renewal, lineage specification, and patterning[1,5-7]. Mutations that disrupt the ability of these TFs to bind DNA are commonly associated with disease[6,8].

PAX6 is a highly conserved transcription factor essential for the correct development of the central nervous system, the pancreas, and the eye[9-12]. In humans, PAX6 is first expressed in the surface and neural ectoderm through the activity of a tightly orchestrated array of enhancer elements, and it induces the expression of multiple downstream genes that collectively regulate the balance between neural progenitor cell self-renewal and differentiation[13-24]. Studies using Hidden Markov models and chromatin immunoprecipitation (ChIP) experiments confirm that PAX6 binds to hundreds of genomic loci associated with embryonic development, patterning, and transcriptional regulation[14,24]. The pivotal role of PAX6 in oculogenesis was famously highlighted in early experiments where its misexpression resulted in the formation of ectopic eye structures in both vertebrates and invertebrates[25,26].

Mutations in PAX6 cause a range of developmental disorders (reviewed here[27]), with homozygous deletions being lethal in mice and humans, while heterozygous deletions, nonsense, and frameshift mutations typically cause aniridia in humans. By contrast, most missense variants produce a broad range of other ocular pathologies, such as microphthalmia, anophthalmia, coloboma, and isolated foveal hypoplasia. Knowledge of the impact of PAX6 missense variants on binding to DNA is key for genetic diagnosis of eye development disorders[28,29]. However, exploring TF-DNA interactions has so far been limited by relatively low-throughput techniques that are resource-intensive and can only feasibly be performed on a handful of variants.

Deep mutational scanning (DMS)[30-33] is a high-throughput experimental strategy to systematically map the effects of mutations in a gene on a selected molecular or cellular

phenotype, such as cell growth rate, expression level, or interaction with a ligand. In DMS, a library of mutants is constructed and the phenotypes of all mutants are measured in parallel, typically in a pooled cell-based assay using a next-generation sequencing readout. DMS has been applied to classify disease mutations, predict the evolution of viruses, inform precision medicine approaches and study the mechanisms through which mutations influence function[34–36]. DMS has also been applied to human TFs[37–39], including TP53. Most of these DMS assays used cell growth rate as phenotypic readout. However, only a minority of TFs have a measurable effect on growth in human cell models, making it difficult to generalise the TP53 experimental approach to other TFs.

We hypothesised that the functional effects of transcription factor mutations could be systematically studied using an adaptation of a yeast one-hybrid (Y1H) assay. In a Y1H assay, a TF binding site (the "bait") is cloned upstream of a reporter construct such as an antibiotic resistance gene that is expressed upon binding of a cognate TF (the "prey"), resulting in antibiotic resistance[40]. Y1H assays have typically been used to screen libraries of putative binding sites for their ability to interact with a TF of choice in TF-centred approaches[41], but they can also be used to screen libraries of TFs for the ability to bind a specific DNA element (gene-centred)[42]. By applying the assay to a library of single amino acid variants of the PAX6 paired domain, we validate Y1H as a generally applicable modality for DMS assays, we identify positions in PAX6 with differential effects upon binding to different DNA sequence elements, and we build an accurate classifier of the clinical effects of mutations in PAX6.

RESULTS

A high-throughput yeast one-hybrid assay to characterise *PAX6* variants

To systematically measure the effects of mutations in the paired domain of human *PAX6*, we developed a Y1H assay in which expression of the PAX6 paired domain fused to the GAL4 activation domain (henceforth collectively referred to as simply PAX6) conferred Geneticin resistance to yeast (S. cerevisiae) cells (Fig 1a). Geneticin resistance was mediated by binding of PAX6 to a DNA response element upstream of the Geneticin resistance gene and integrated into the yeast genome. We tested 16 candidate PAX6 DNA response elements (Fig S1a), from which 2 that conferred specific antibiotic resistance upon PAX6 binding were chosen: LE9, a minimum PAX6-responsive element derived from the head surface ectoderm enhancer that autoregulates expression of *Pax6* in ectoderm derivatives in the eye[17], and BLX, a sequence derived from a SELEX screen for shared PAX6 paired domain and SOX2-binding sequences (unpublished data). The remaining 14 candidate DNA response elements were rejected because they did not offer antibiotic resistance (possibly due to insufficient copy numbers, low PAX6 affinity, or PAX6 binding requiring the presence of cofactors such as SOX2[19,21,43]), or they resulted in antibiotic resistance in the absence of PAX6 (likely caused by interactions with endogenous yeast transcription factors). Fusion of GAL4 activation domain at the C-terminus of the PAX6 paired domain conferred antibiotic resistance whereas an N-terminal fusion did not. We replicated these results in liquid culture assays, and confirmed that previously studied PAX6 amino acid disease variants found in humans with varying degrees of disruption to DNA binding function did not generate antibiotic resistance[29] (Fig S1b).

Encouraged by these results, we used one-pot saturation mutagenesis[44] to generate a library of single amino acid variants of the PAX6 paired domain and its flanking sequence (amino acid positions 1-150; Fig 1c). We then used the Gateway system to clone the variant library into a centromeric yeast expression plasmid library that included a strong synthetic promoter[45] and a collection of 30-nt barcodes in the 3' UTR. The linking PAX6 variants with their corresponding barcodes was done using PacBio sequencing of the plasmid library followed by phasing using alignparse[46]. The final library contained 96.5% (2731/2831) of all possible single nonsynonymous (missense) variants, 99.3% (148/149) of all possible single stop codon (nonsense) variants, 227

synonymous variants, and 5,266 unique variants with two or more mutations (Fig S2a). Missense and nonsense single variants were associated with a median of 19 barcodes per variant, and 8,655 barcodes were associated with wild-type PAX6.

PAX6 variants cause sequence-specific perturbations in DNA-binding

To quantify DNA binding of PAX6 variants, we transformed the mutant library into LE9 and BLX yeast strains, treated cells with Geneticin, collected samples at 12 hr intervals, and quantified barcodes by Illumina sequencing (see Methods). Barcode counts were analysed with the Enrich2 pipeline[47], yielding fitness scores for 95.8% (2856/2980) of variants in the LE9 assay and 95.2% (2838/2980) of variants in the BLX assay. Fitness scores were normalised so the average of wild-type variants was 0 in both assays. To quantify the reproducibility of these scores, we analysed sets of barcodes associated with the same PAX6 variant. After separating all barcodes into two equal sets and calculating variant scores for each set, we obtained Pearson correlations of 0.77 for the LE9 assay and 0.63 for the BLX assay for all barcodes. As expected, reproducibility increased with the number of barcodes per variant and with read coverage per barcode (Fig S2).

Variant fitness scores showed broadly similar patterns in the LE9 and BLX assays, with a Spearman correlation of 0.64 between assays (Fig 2b,d,e). The BLX assay appeared more tolerant of PAX6 mutations (Fig 2e), perhaps reflecting stronger binding of PAX6 to the SELEX-derived consensus binding site. Premature termination codons (PTC; nonsense) were detrimental throughout the gene in each assay. This was expected because a PTC in PAX6 paired domain results in a truncated protein that lacks GAL4 activation domain and should not induce antibiotic resistance. The mean scores of synonymous variants (LE9 mean = -0.09; BLX mean = -0.24) were slightly lower than wild-type PAX6, but this was most likely caused by a small number of outliers with a single barcode per variant, as the effect all but disappeared when variants with three or more barcodes were considered (Fig S2c). Most missense variants were detrimental to function (Fig 2b, with some positions appearing completely intolerant of amino acid substitutions (blue columns) while other positions tolerated substitutions well (white columns). Substitutions to proline and to negatively charged amino acids were most detrimental,

consistent with their roles in disrupting alpha helices and electrostatic interactions with DNA, respectively. Mutational sensitivity was correlated with PAX6 structural elements (Fig 2a) and with sequence conservation (Fig 2c). The apparent periodicity in α-helices α1, α4 and α5 was reminiscent of α-helix periodicity of 3.6 residues per helical turn, suggesting that one face of the helix may be more sensitive to mutations than the other. Periodicity was also observed in the β1 and β2 beta sheets (Fig 2b).

Some positions showed distinct patterns of mutational sensitivity: positions 50, 70, and 71 in the LE9 assay, and 50 and 54 in the BLX assay, were enriched for mutations that apparently increased DNA binding compared to wild-type PAX6 (Fig 2b, S3). Position 17 was unique as many mutations in this position reduced yeast fitness to a larger extent than PTCs or other presumed loss of function mutations; we hypothesise that mutations in position 17 generate PAX6 variants that are toxic to yeast through a molecular gain-of-function mechanism. Certain variants appear to have opposing effects on fitness across the two assays – for example, Asn50Arg, Gly51Arg, Ser54Arg, Ser74Gly, Ser121Asn, and multiple substitutions at position Ile71 all increased the fitness score in one DMS assay (LE9 or BLX) while decreasing fitness in the other (Fig S3). Similarly, there were also numerous variants such as Cys52Gly, Ala70Ser, Ile71Leu, Gly72Ser, and Arg128Gln that showed either a significant change in fitness score or no effect on fitness, depending on the DNA response element. Analysis of replicate barcodes in these variants and other cases suggested that they were genuine effects of PAX6 binding to different DNA response elements, rather than experimental noise (Fig S3).

Structural Sensitivity of PAX6 to Amino Acid Substitutions

Mutations in the interior of proteins or at protein:DNA interfaces are known to be more detrimental than mutations in surface residues[48-50]. To verify that this principle applies to PAX6, we analysed the effects of PAX6 mutations in the context of the PAX6:DNA complex structure. We used the 6PAX crystal structure available in PDB to model the interactions of PAX6 paired domain with LE9 and BLX binding sites[23]. As expected, we found greater effects of mutations in inward-facing residues of the subdomain protein cores and the protein:DNA interface (Fig 3a,b, left panels). The proximity of each residue

to DNA correlated with fitness score, with residues closest to DNA having the lowest scores (Fig 3a,b, central panels).

Previous studies also suggested that the N-terminal subdomain (NTS) had a larger impact on PAX6:DNA interactions, when compared to the C-terminal subdomain (CTS)[23,51,52]. Our analysis confirms this, with NTS mutations being more consistently detrimental than CTS mutations or mutations outside the paired domain (Fig 3A,B, right panels). Mutations in the linker sequence between the two subdomains were slightly less deleterious than NTS mutations in the LE9 assay but the opposite was observed in the BLX assay. We speculate that the effects of mutations in the linker sequence depend on the relative arrangement of the binding sites of the NTS and CTS subdomains in the DNA structure. For example, a larger distance between these sites will favour linker mutations that promote a more extended conformation.

DMS accurately classifies pathogenic variants and gives insight into disease phenotypes

Recent studies show the utility of deep mutational scanning in the annotation of variants of uncertain significance (VUS)[34]. We therefore examined if PAX6 fitness scores could explain the pathogenicity of mutations. Putatively benign variants found in the gnomAD database had fitness scores comparable to synonymous variants (Fig. 4a). By contrast, pathogenic variants collected from the from the HGUeye, ClinVar and HGMD databases were overwhelmingly detrimental, with a mean fitness score close to nonsense variants (Fig 4a). While most variants followed these patterns, this was not absolute, with some pathogenic variants showing neutral or increased function relative to the wild-type, and some gnomAD variants demonstrating varying degrees of reduced function. Moreover these effects were often DNA response element-specific, suggesting the observed effects on fitness were a result of PAX6:DNA binding, rather than stability-related disruption.

Deep mutational scanning has also been used to benchmark in-silico variant effect predictors (VEPs). We therefore compared the accuracy of classification of PAX6 disease mutations based on DMS and VEPs (Fig 4b, Table S2). Raw PAX6 fitness scores classified

benign and pathogenic variants better than most VEPs (LE9 ROC AUC: 0.885, BLX ROC AUC: 0.891). Previous studies indicated that the magnitude of change in function rather than the sign caused by amino acid substitution offers greater accuracy in classifying pathogenic variants[48,53]. Accordingly, we calculated absolute values (ABS) for fitness scores and assessed the predictive performance (Fig 4b). Interestingly, we saw an improvement in both DMS assays (LE9 ABS ROC AUC: 0.913, BLX ABS ROC AUC: 0.954), with BLX ABS fitness scores outperforming all VEPs. LE9 ABS fitness score outperformed all unsupervised VEPs, with only MetaRNN and MutPred surpassing it in ROC AUC[54,55].

We next used a proposed modification to American College of Medical Genetics (ACMG)/Association for Medical Pathology (AMP) guidelines for variant interpretation to quantify the evidence of pathogenicity/benignity generated in this study (Fig 4b)[56,57]. Our findings indicate that the BLX ABS assay provides moderate evidence of pathogenicity for 941 PAX6 variants, supporting evidence of pathogenicity for 36 variants, moderate evidence of benignity for 1184 variants and supporting evidence of benignity for 122 variants. The remaining 411 variants provide evidence for neither pathogenicity nor benignity.

Following the observation that absolute fitness scores offered better classification of pathogenic variants, we explored whether atypical (non-aniridia) phenotypes could be explained using our approach (Fig 4c). Taking advantage of a recent publication in which ocular disease phenotypes were systematically annotated in individuals possessing PAX6 missense variants[29], we subdivided variants into three groups; (1) individuals with mild or classical aniridia; (2) individuals with atypical phenotypes that included anterior segmentation, MAC (microphthalmia, anophthalmia, coloboma) spectrum, or classical aniridia with any of the following: optic disc coloboma, morning glory disc anomaly, coloboma (iris/choroid/retina), microphthalmia, anophthalmia, or microcornea; (3) individuals where assignation to (1) or (2) could not be determined. No significant difference was detectable in the BLX_ABS assay across variants with mild/ classical aniridia or atypical phenotype-linked variants (Fig 4c, right). Interestingly, atypical phenotype-linked variants had significantly higher (p=1.8 x 10^{-3}) absolute effect sizes than mild/ classical aniridia phenotype-linked variants according to the LE9_ABS assay (Fig 4c, left) implying these variants experienced a greater degree of

DNA-binding disruption specifically. Additionally, mild/ classical aniridia phenotype-linked variants were also indistinguishable from nonsense variants in the LE9_ABS assay while atypical phenotype-linked variants had the greatest absolute change in DNA binding compared to nonsense variants (p=1.3 x 10^{-9}).

PAX6 reduces yeast growth through promiscuous binding to the genome

While PAX6 allows yeast strains containing LE9 and BLX bait sequences to grow in Geneticin-containing media, we noticed that expression of wild-type PAX6 caused a small but reproducible reduction of growth in the absence of antibiotic. To test whether this effect was specific to wild-type PAX6, we repeated the DMS assays with no antibiotic in the culture media (Fig. 5b). Intriguingly, we observed that wild-type PAX6 and variants that can bind DNA reduced growth in the absence of antibiotic, whereas loss-of-function variants showed relatively faster growth. In particular, amino-acid mutations in the PAX6:DNA binding interface prevented growth in media containing Geneticin, but increased growth without Geneticin (Fig. 5b) We observed negative correlations of variant effects when comparing DMS assays with and without Geneticin (Fig 5d,e. LE9 +Geneticin vs -Geneticin: R=-0.73; BLX +Geneticin vs -Geneticin: R=-0.54).

We considered several possible explanations for these observations: (1) expression of PAX6 may impose a metabolic load and divert resources from essential cellular pathways; (2) misfolding of mutant PAX6 may cause a cellular stress response; (3) gratuitous induction of the Geneticin^R gene may impose a metabolic cost; (4) non-specific binding of PAX6 to yeast genomic DNA might reduce growth by perturbing the yeast transcriptome. Scenarios (1) and (2) do not explain the results as they predict mutant PAX6 to have similar (1) or lower (2) growth rates than wild-type PAX6 in the absence of Geneticin, whereas we observed an increased growth rate in PAX6 mutants. To distinguish between (3) and (4), we reasoned that cost related to Geneticin^R gene expression should result in strong negative correlations within the same bait strain grown with and without antibiotic, whereas the cost related to non-specific genome binding of PAX6 should result in a strong positive correlation between LE9 and BLX strains grown without antibiotic. We found the strongest correlation (R=0.95) between LE9 and BLX strains without antibiotic (Fig. 5f), suggesting that the identity of the bait

sequence is not relevant in this condition. This analysis supports the idea that functional (DNA-interacting) variants of PAX6 reduce growth in the absence of Geneticin through promiscuous binding to the yeast genome.

The observation that sequence-function maps with (Fig. 2b) and without (Fig. 5b) Geneticin are nearly mirror images of each other prompted us to explore whether assays without Geneticin reflected PAX6 physiological function. To this end, we tested if the assays without Geneticin accurately predict the pathogenicity of PAX6 mutations (Fig. 4b). Remarkably, the LE9 and BLX assays without Geneticin offered greater predictive accuracy compared to their respective assays with Geneticin (LE9 -Geneticin ROC AUC: 0.892; BLX -Geneticin ROC AUC: 0.898). Using absolute fitness scores improved accuracy further (LE9 ABS -Geneticin ROC AUC: 0.922; BLX ABS -Geneticin ROC AUC: 0.926), with both assays surpassing all other DMS assays and unsupervised VEPs in ROC AUC with the exception of BLX ABS with Geneticin, and were only surpassed by the supervised VEPs MetaRNN and MutPred[54,55].

DISCUSSION

PAX6 genotype-phenotype relationship

Here we systematically measured the effects of missense mutations in the paired domain of PAX6 on interactions with DNA and we predicted the pathogenicity of known PAX6 variants with 92% accuracy. Of the variants previously observed in the human population, 31 were known as benign and 91 as pathogenic. We quantified the effects of 2,574 additional variants and we provide supporting to moderate evidence to suggest 1,306 are likely benign and 977 are likely pathogenic.

While heterozygous nonsense mutations in PAX6 typically lead to aniridia, missense mutations cause diverse phenotypes that include microphthalmia, anophthalmia, and coloboma (MAC), isolated foveal hypoplasia, and congenital nystagmus. The reasons for these phenotypic differences are not well understood: in theory, missense mutations can cause partial loss of function, gain-of-function, or dominant negative effects, and indeed functional studies show that some missense mutations lead to different DNA binding outcomes[28,29]. In the present study, approximately 20% of missense mutations caused partial loss of function and an additional 9-16% showed either increased binding to DNA or toxicity, which could both be considered as molecular gain-of-function. There is also indication that specific phenotypes in yeast can be related to patients' phenotypes. For example, variants Asn17Ser and Asn17Lys were found respectively in aniridia and Peter's anomaly; in our data, Asn17Ser caused a LOF or near-LOF phenotype, consistent with aniridia, whereas Asn17Lys was toxic to both LE9 and BLX strains, consistent with its more severe patient phenotype. Our results are also consistent with the survival of a patient with compound heterozygous mutations in PAX6 (p.Arg38Trp/p.Arg240*), since the p.Arg38Trp mutation showed some activity in the BLX strain[58]. Overall, statistical analysis of the yeast data showed a moderately strong association between atypical patient phenotypes and fitness scores in the LE9, but not BLX strains.

PAX6 exists in two major isoforms, canonical PAX6 and PAX6(5a), which differ in their expression patterns: the canonical isoform is expressed during embryonic development and controls differentiation and cell fate determination, whereas PAX6(5a) is expressed at later stages and plays a role in cell proliferation. It has been reported that the two isoforms also differ in their mode of binding to DNA, with the canonical isoform

preferentially binding through the NTS and PAX6(5a) through the CTS domain. In our experiments with canonical PAX6, we indeed observed a significantly higher effect of mutations in the NTS than in CTS, in both the LE9 and BLX strains. However, many CTS mutations, especially in the α6 helix, also disrupted binding, showing that both NTS and CTS domains contribute to canonical PAX6:DNA interactions.

PAX6 mutations alter DNA binding specificity

A possible explanation for the phenotypic diversity of missense mutations is that they may change DNA binding specificity of PAX6, leading to stronger binding of certain enhancer elements and weaker binding of others, which may in turn lead to variant-specific changes in development. This phenomenon has been reported in a missense variant in the transcription factor Krüppel-like factor 1 that is important in erythroid cell development: the variant caused an autosomal dominant disorder by corruption of binding preference from a canonical DNA motif, leading to dysregulation of gene expression[59]. In PAX6, Tang et al found that the I87R variant found in a patient with aniridia ablated DNA-binding, while R26G, a variant located proximally to the protein-DNA interface that caused relatively less disruption DNA-binding, lead to a non-aniridia phenotype[28]. Similarly, a recent study demonstrated that the MAC-associated variant, S54R, showed sequence-dependent DNA-binding disruption to two characterised PAX6-binding sites, LE9 and SIMO, while aniridia-associated variants such as C52R and S121L consistently ablated DNA-binding[29].

Here we found numerous examples of mutations that either alter PAX6 interactions with one of the DNA response elements without changing interactions with the other element, or affect DNA-binding in opposite ways. In agreement with the findings by Williamson et al[29], we observed that the mutation S54R results in a substantial loss of affinity with LE9, however rather than an inconsequential effect on binding on another element observed by the researchers, we saw increased binding with BLX. The most striking position in which we saw PAX6 target-specific effects was at I71, where numerous amino acid substitutions resulted in substantially increased affinity towards LE9, but reduced affinity to BLX. No human variants to date have been identified at this position, however the nearby pathogenic variant S74G showed sequence-specific

inverse effects on DNA-binding in our assay. In patients, S74G is associated with non-aniridia phenotypes, including ocular, neurological and cognitive impairments[60]. Interestingly, a recent study exploring the link between DNA-binding affinity and phenotypic differences found there was no dramatic difference in the affinity of wild-type PAX6 and S74G to three DNA targets, however these targets did not include LE9 or BLX[61].

The variant Gly51Arg has previously been reported recently in individuals with MAC[29], and according to our assay, results in a LOF with LE9 and a neutral effect with BLX which might explain the atypical phenotype under the hypothesis that DNA-binding exclusively causes pleiotropy. However, Gly51Arg has also been reported in individuals with aniridia[62] suggesting other factors affecting phenotypic heterogeneity. Intrafamilial phenotypic variability associated with the same mutation has been reported both in individuals with aniridia harbouring indel mutations as well as in individuals with missense variants that cause atypical phenotypes such as isolated foveal hypoplasia and nystagmus[58,63,64]. Mechanisms such as mosaicism, gene dosage and genetic modifiers are possible explanations of variability[58,65,66] so it is conceivable that pathogenic variants such as Gly51Arg are subject to similar confounding layers of control.

Effect of PAX6 in the absence of antibiotic

Although our assay was designed to probe PAX6:DNA interactions through an antibiotic resistance reporter gene, it also showed effects of PAX6 mutations in the absence of antibiotic. These effects were highly reproducible and they correlated with DNA binding activity of PAX6 and with clinical effects of PAX6 mutations; we speculate that they result from promiscuous binding of the PAX6-GAL4 fusion protein to yeast genomic DNA, which reduces growth by perturbing the transcriptome. Overexpression of GAL4 in yeast was previously observed to cause a growth defect[67], but the effect, known as squelching, was independent of GAL4:DNA binding and instead resulted from titration of transcription factors by the activating region of GAL4. In our results, expression of DNA binding-deficient PAX6-GAL4 fusions did not reduce growth, suggesting that the effect we observe is distinct from squelching.

Regardless of the mechanism, these results suggest that yeast growth can be a useful phenotypic readout for deep mutational scanning of other human transcription factors. Unlike the normal Y1H setup, which necessitates laborious design and testing of strains with appropriate bait sequences, the assay without antibiotic is predicted to work without a bait sequence and could potentially be applied directly to many different TFs. Of the 1,600 human TFs, 313 have been associated with at least one disease phenotype according to a recent study[1], and therefore may be good targets for this style of assay.

While PAX6 mutant phenotypes with and without antibiotic were negatively correlated with each other, the correlation was strongly nonlinear: a subset of variants had a range of negative or positive effects in the absence of antibiotics, but showed no effect in the presence of antibiotic. In addition, some variants, including most of the Asn17 mutants, showed negative effects on growth in the presence and absence of antibiotic. While the exact interpretation of these results is not clear, the pattern is reproducible between the LE9 and BLX strains and it indicates the complex interplay between the growth promoting effects of PAX6 through expression of the antibiotic resistance gene, and growth reducing effects caused by promiscuous binding to yeast genomic DNA.

Uses of deep mutational scanning in an age of variant effect predictors

Large-scale genome sequencing has greatly increased the number of genetic variants detected in the human population but it also caused a proliferation of variants of uncertain significance (VUS). DMS and VEPs are two approaches that have the potential to resolve VUSs at scale by classifying them as either benign or pathogenic, but classification accuracy depends on the target gene and on the type of VEP or DMS approach being used. For some genes such as BRCA1, DMS performs better than all computational predictors[34], whereas computational methods outperform DMS for other genes[68]. This raises questions about what experimental or computational variables influence the performance of DMS assays, and about the utility of DMS in general, given that as an experimental approach DMS is much more resource-intensive than computational predictors.

The four DMS datasets we collected showed consistently high performance in their ability to predict pathogenic variants in PAX6, with ROC AUC values between 0.885 and

0.954. This performance was slightly better than that of the top unsupervised predictors identified in a recent benchmarking study (EVE, ESM-1v and DeepSequence, ROC AUC 0.811-0.886)[69-72], but were on average slightly worse than top supervised predictors (MetaRNN, MutPred, and MutationAssessor, ROC AUC 0.901-0.932)[54,55,73]. These top supervised predictors may however have been trained on data used for evaluation and suffer from data circularity[74]. We identified several factors that influenced the predictive performance of DMS: (1) higher data coverage (sequencing depth and more importantly, number of barcodes) for a particular variant increased predictive power, though applying a coverage threshold also decreased the proportion of variants we could classify; (2) the number of experimental time points used for fitness calculations had a complex relationship with performance: too few data points resulted in lower accuracy, but too many data points reduced the numbers of fitness scores obtained, seemingly due to difficulty the model had in fitting all the data; (3) as found in some previous studies[48,53], the use of absolute fitness scores instead of raw scores substantially increased the predictive performance (ROC AUC for raw scores, 0.885-0.891; absolute scores, 0.913-0.954). We propose that the utility of absolute scoring depends on whether the gene in question harbours pathogenic gain-of-function (GOF) variants: transcription factors are known to contain such mutations, and indeed examples of pathogenic gain-of-function variants have also been described in PAX6[29]. Owing to the dosage sensitivity of PAX6, it has been shown that both overexpression[75-77] and heterozygous knockouts[78,79] can be pathogenic, so it is conceivable that any deviations in missense variant activity relative to homozygous wild-type functional levels, positive or negative, has the potential to be pathogenic.

Irrespective of its ability to predict disease variants, one advantage of DMS over the current generation of VEPs is that DMS can probe multiple molecular and cellular phenotypes in parallel. This can lead to a better mechanistic understanding of genetic variants, e.g. their effects on cell growth and survival, gene expression, protein-protein interactions and protein-DNA interactions (this study)[34,80,81]. DMS can also probe the effects of mutations depending on the cellular context, drug treatment, or the presence of additional mutations[82,83]. While we have not investigated double mutants systematically in this study, our dataset contains thousands of such mutants which could be analysed for intragenic epistasis in the DNA-binding domain of PAX6.

DMS also has limitations: for example, any mutations that influence splicing of human PAX6 would not have been detected in the present assay that used an intronless PAX6 cDNA in a yeast model system. Effects on splicing are detectable using other DMS techniques, such as saturation genome editing[34], and might be detected by a subset of generic VEPs or dedicated splicing prediction software[84,85]. In addition, we would not have detected effects of mutations on interactions of PAX6 with cofactors (such as SOX2) that are not found in yeast. An effect on splicing or cofactor interaction could explain the only known pathogenic variant of PAX6 (Glu93Lys) that was scored as benign in our analysis. Nevertheless, the high predictive power of our assays demonstrates that most pathogenic mutations in the paired domain of PAX6 can be explained simply by the effects of these mutations on PAX6:DNA association.

METHODS

Enhancer-specific *S. Cerevisiae* strain generation

All strains were derived from the S288C strain, Y1HGold (MATα, ura3-52, his3-200, ade2-101, trp1-901, leu2-3, 112, gal4Δ, gal80Δ, met–, MEL1), obtained from Takara Bio. Initially the Matchmaker® Gold Yeast One-Hybrid Library Screening System (Takara Bio) was employed, however owing to the incompatibility of Aureobasidin A with liquid cultures, an equivalent Geneticin resistance gene was subcloned in place of the Aureobasidin A resistance gene. PAX6 DNA response elements (baits) were synthesised either as single copies or tandem triple repeats (IDT) and individually cloned into a plasmid containing a downstream Geneticin resistance gene and URA3 marker (pKani). Each pKani+bait vector was then integrated into Y1HGold genomic DNA via homologous recombination. Each pKani+bait vector was linearised by restriction digestion with BstBI and incubation at 65 °C for 1 hour, followed by purification using DNA Clean & Concentrator-5 (Zymo). DNA was then transformed into Y1HGold using the protocol described by Gietz et al.[86], followed by selection on plates lacking uracil. Correct genomic integration was confirmed using colony PCR.

PAX6 variant library construction

The PAX6 (1-150) variant library was generated using a modified version of PCR-based saturation mutagenesis described by Wrenbeck et al[44]. Briefly, the top strand of a dsDNA plasmid (pAINt1) containing wild-type PAX6 was nicked using Nt.BbvCI, and degraded using Exonuclease I and III (NEB). The resulting ssDNA template was then split between 6 parallel reactions and custom mutagenic oligo primers (IDT) added such that they spanned 6 adjacent 75bp stretches of the entire PAX6 sequence construct (Table S1). Phusion polymerase (NEB) was then used to synthesise a mutated top strand, followed by Taq DNA ligase (NEB) to seal nicks. The 6 reactions were then pooled and the wild-type bottom strand nicked using Nb.BbvCI and digested using Exonuclease I and III. A second primer was then added and a complementary mutated bottom strand synthesised using Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) and Taq DNA ligase, creating a pooled dsDNA PAX6 variant library. This library was then transformed into XL10-Gold Ultracompetent Cells (Agilent) and transformants grown on

appropriate antibiotic plates. Colonies were harvested by resuspension in H_2O, followed by scraping, pooling, and centrifugation. Finally, plasmids were isolated using Qiaprep spin miniprep kit (Qiagen).

Assigning barcodes to variants

The PAX6 variant library within pAINt1 was transferred to a centromeric yeast expression plasmid (pASH) containing approximately 4×10^6 unique 30-nt barcodes, using the LR Gateway system (Thermo Scientific). The reaction product was then transformed into Library Efficiency™ DH5α Competent Cells (Thermo Scientific) and plated on appropriate antibiotic plates. Approximately 7.5×10^4 colonies were harvested by resuspension in H2O, followed by scraping, pooling, and centrifugation.

An aliquot of the barcoded PAX6 variant library was taken and a region of the plasmid containing the barcodes, PAX6, and GAL4 was excised from the vector using restriction digestions followed by BluPippin purification. This fragment was then sequenced using PacBio Sequel II and 20 hours of movie collection time. Finally, barcode:variant phasing was achieved using the alignparse Python package[46].

Competition assays

Competition assays were performed using a complete supplement mixture without histidine (Foremedium) and yeast nitrogen base without amino acids (Foremedium) media, supplemented with 2% glucose. To begin, the barcoded PAX6 variant library was transformed into two separate strains, LE9 and BLX, in two independent assays. The transformation process followed the protocol described in the Yeastmaker™ Yeast Transformation System 2 User Manual (Takara Bio), yielding approximately 2×10^6 and 7×10^6 transformants for LE9 and BLX strains, respectively. Transformed cells were then inoculated into 500 ml of liquid media and incubated at 30°C and 230 RPM for 12-24 hours until reaching an OD600 of approximately 1.0. Subsequently, approximately 7.5×10^8 cells were transferred to fresh media and grown for an additional 12 hours (T0). The T0 cultures were then split into two parallel lineages. All cultures were then maintained between an OD600 of 0.05 and 1.0 by passaging every 12 hours for a total of

36 hours (T1-3). Geneticin was then added to one lineage of each strain (LE9; 400 μg/ml, BLX; 500 μg/ml) starting from T0. Just before each passage, 10 to 20 OD600 units of cells were collected and pelleted by centrifugation at 700 x g for 5 minutes and the supernatant removed, followed by two rounds washing using resuspension in H2O and centrifugation to remove extracellular plasmid DNA.

Plasmid isolation and sequencing sample preparation

Plasmid isolation was achieved using the protocol described by Fowler et al.[87], which in summary used a freeze-thaw cycle and zymolyase cell wall digestion steps were coupled with Qiaprep spin miniprep kit (Qiagen) column purification. Samples from each timepoint were prepared for next-generation sequencing using a single limited-cycle PCR. Barcodes were amplified using primers possessing overhangs that contained sequencing primer annealing sequences, unique dual indexes, and P5/P7 Illumina adapter sequences (Table S1). 50 μl PCR reactions were performed for each timepoint using Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB) with 20 to 50ng of template isolated plasmid DNA. Each reaction was then thermocycled using the following conditions: (1) initial denaturation at 98°C for 30 seconds, (2) incubation at 98°C for 8 seconds, (3) incubation at 69°C for 20 seconds, (4) incubation at 72°C for 8 seconds, (5) incubation at 72°C for 7 minutes. Steps 2 to 4 were repeated for 12 to 19 cycles (< 20). Multiple replicate 50 μl reactions were performed for each timepoint to ensure enough PCR product was generated. 2.5ul of Exo I nuclease (NEB) was then added directly to each 'dirty' PCR product and incubated at 37°C for 1 hour, followed by purification using QIAquick PCR Purification Kit (Qiagen). Samples were further purified by 2% E-Gel SizeSelect agarose gel (Invitrogen). The resulting products were purified using QIAquick PCR Purification Kit (Qiagen), after which all samples were pooled at equimolar amounts (quantified by QuBit [Invitrogen]) and then quantitated by Bioanalyzer (Agilent).

Barcode sequencing and fitness score calculations

Samples were sequenced using either Nextseq 550 High Output flowcell with 75 cycles or Nextseq 2000 P2 flowcell with 100 cycles for the LE9 and BLX assays, respectively. A custom Read 1 sequencing primer (Table S1) was used in both libraries, and annealed to the template such that the first position in the read corresponded to the first nucleotide of the barcode. Samples were spiked with 10% PhiX. Following demultiplexing, the regions flanking the barcodes were trimmed using trimgalore[88] then barcodes counted and variant fitness scores calculated using the Enrich2 tool[47]. A minimum threshold of 25 reads per barcode was applied to the first timepoint (T0).

In addition to the p-values generated by Enrich2 for each variant using z-distribution under the null hypothesis that it behaves like wild-type (p_{WT}), an equivalent p-value was calculated under the null hypothesis that each variant behaves like a nonsense variant ($p_{nonsense}$). All p-values were adjusted to q-values using the False Discovery Rate (FDR) from the p.adjust() r package. Relative changes in function for each variant were assigned for each assay using the following categories: increased (fitness score > 0, & q_{WT} < 0.05), neutral (q_{WT} > 0.05), hypomorphic (0 > fitness score > mean nonsense score, & q_{WT} < 0.05, & $q_{nonsense}$ < 0.05), LOF ($q_{nonsense}$ > 0.05), and toxic (fitness score < mean nonsense score, & $q_{nonsense}$ < 0.05).

Structural analysis of fitness scores

PAX6 region amino acid residue ranges were defined as: NTS (4-64), Linker (64-80), CTS (81-133) and 'other' (1-3 and 134-150). LE9 and BLX DNA sequences were modelled with PAX6 using the 6PAX crystal structure[23]. A solvent accessibility surface area threshold of > 25% was used to classify variants as solvent accessible and < 25% as solvent inaccessible. Distances between PAX6 residues and DNA were calculated by measuring between each amino acid residue and the nearest non-hydrogen atom of the DNA duplex. A median fitness score across each position was calculated, mapped to either the PAX6:LE9 or PAX6:BLX structure, and visualised with PyMol.

Human variants and phenotypes

Human pathogenic PAX6 variants and their respective annotated phenotypes were extracted from a previous publication[29] using the HGUeye cohort (n = 67), and aggregated with more recently identified variants in ClinVar (n = 17) and the Human Gene Mutation Database (HGMD) (n = 7). Variants were subdivided into three groups that corresponded to different phenotypes, based on the phenotypic definitions defined in the aforementioned publication. Group one comprised variants associated with mild or classical aniridia, while group two comprised of individuals with non-aniridia (atypical) phenotypes that included anterior segmentation and MAC (microphthalmia, anophthalmia, coloboma) spectrum, or classical aniridia with any of the following: optic disc coloboma, morning glory disc anomaly, coloboma (iris/choroid/retina), microphthalmia, anophthalmia, or microcornea. The third group comprised variants corresponding to individuals where categorisation into group one or two could not be determined. Benign variants were taken from the gnomAD database (n = 32).

VEP benchmarking

VEP scores were mostly obtained from the dbNSFP database version 4.2[89]. To generate amino acid level scores from the database, the score for each VEP across all codons resulting in the same amino acid were averaged. DeepSequence, SIFT, ESM-1v and sequence_unet were run locally. Publicly available pre-computed scores were used for VESPAl, CPT, DeepSAV, EVE, PonP2, mutationTCN and LASSIE. Eigen is classified as supervised VEP due to its inclusion of PolyPhen-2 as a predictive feature. To calculate the ROC AUC for DMS assays and VEPs, we used the roc_auc_score function of the Python package sklearn, with pathogenic variants labelled as class '1' and benign variants as class '0'. Due to variability in predictor coverage of PAX6, a minimum of 10 pathogenic and 10 benign variant predictions were required for the inclusion of each VEP. Predictor scale orientations were normalised by inversion where appropriate to maintain comparability, by application of the following formula:

$$inverted\ predictions = \ min(predictions) \ - \ predictions \ + \ max(predictions)$$

The optimal classification threshold was calculated from the ROC curve by identifying the threshold that generated the point closest to the top left of the curve (the highest true positive rate - false positive rate). Correctly classified pathogenic variants produced scores greater than, or equal to these threshold values, while correctly classified benign variants scored less than the threshold. Accuracy was also calculated at this threshold using the following formula:

$$accuracy = \frac{(true\ positives\ +\ true\ negative)}{total\ classified\ variants}$$

Estimation of the evidence of pathogenicity and benignity

All variants with a pathogenic or benign label were ordered by descending fitness score using the absolute scores from the BLX assay with Geneticin. The prior probability of pathogenicity was calculated from this data as the proportion of pathogenic variants. Windows of dynamic size were then constructed around each variant in the labelled dataset using a similar method to Pejaver et al.[90]. Briefly, windows were required to contain at least 20 variants, and at least two from the minority class. If these conditions were not met then the window size was increased to encompass more variants until they were achieved. The posterior probability of pathogenicity was calculated within each window as the proportion of pathogenic variants. As per the method described by Pejaver et al.[90], the number of benign variants were weighted to calibrate for the prior probability of pathogenicity using the formula:

$$w = \frac{(1 - P1) * n_{pathogenic}}{n_{benign} * P1}$$

Where P1 is the prior probability of pathogenicity, $n_{pathogenic}$ is the total number of pathogenic variants and n_{benign} is the total number of benign variants.

The odds of pathogenicity (oddsPath) was then calculated within each window using the formula:

$$oddsPath = \frac{P2 * (1 - P1)}{(1 - P2) * P1}$$

Where P1 is the prior probability of pathogenicity and P2 is the posterior probability of pathogenicity.

We then applied the resulting oddsPath results to thresholds in a recent recommendation for an update to the American College of Medical Genetics and Genomic (ACMG)/ Association for Molecular Pathology (AMP) guidelines[56,57] to obtain score thresholds for different evidence strength boundaries. These thresholds were then applied to the full dataset to estimate the evidence strength. The process was then repeated for benign variants with the prior probability of benignity, posterior probability of benignity and windows being determined from the lowest DMS score rather than the highest. Thresholds for pathogenic evidence were calculated from the pathogenic data while thresholds for the benign evidence were calculated from the benign data. Data that did not meet the threshold of 'supporting' for either pathogenic or benign was labelled as indeterminate.

Sequence Alignment

The evolutionary sequence conservation of PAX6 was aligned with ConSurf[91] using MAFFT and homologues collected from UNIREF90 and HMMER search algorithm.

REFERENCES

1.   Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

2.   Karin, M. & Smeal, T. Control of transcription factors by signal transduction pathways: the beginning of the end. *Trends Biochem. Sci.* **17**, 418–422 (1992).

3.   Brivanlou, A. H. & Darnell, J. E. Signal transduction and the control of gene expression. *Science* **295**, 813–818 (2002).

4.   Zabidi, M. A. *et al.* Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).

5.   Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

6.   Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).

7.   Thompson, B. *et al.* Overview of PAX gene family: analysis of human tissue-specific variant expression and involvement in human disease. *Hum. Genet.* **140**, 381 (2021).

8.   Chi, N. & Epstein, J. A. Getting your Pax straight: Pax proteins in development and disease. *Trends Genet.* **18**, 41–47 (2002).

9.   Cvekl, A. & Callaerts, P. PAX6: 25th anniversary and more to learn. *Exp. Eye Res.* **156**, 10–21 (2017).

10.   Dohrmann, C., Gruss, P. & Lemaire, L. Pax genes and the differentiation of hormone-producing endocrine cells in the pancreas. *Mech. Dev.* **92**, 47–54 (2000).

11.   Manuel, M. N., Mi, D., Mason, J. O. & Price, D. J. Regulation of cerebral cortical neurogenesis by the Pax6 transcription factor. *Front. Cell. Neurosci.* **9**, (2015).

12.   Ochi, S., Manabe, S., Kikkawa, T. & Osumi, N. Thirty Years' History since the Discovery of Pax6: From Central Nervous System Development to Neurodevelopmental Disorders. *Int. J. Mol. Sci.* **23**, (2022).

13.   Terzic, J. & Saraga-Babic, M. Expression pattern of PAX3 and PAX6 genes during human embryogenesis. *Int. J. Dev. Biol.* (1999) doi:10.1387/IJDB.10610023.

14.   Sun, J. *et al.* Identification of in vivo DNA-binding mechanisms of Pax6 and reconstruction of Pax6-dependent gene regulatory networks during forebrain and lens development. *Nucleic Acids Res.* **43**, 6827–6846 (2015).

15.   Sansom, S. N. *et al.* The Level of the Transcription Factor Pax6 Is Essential for Controlling the Balance between Neural Stem Cell Self-Renewal and Neurogenesis. *PLOS Genet.* **5**, e1000511 (2009).

16.   Thakurela, S. *et al.* Mapping gene regulatory circuitry of Pax6 during neurogenesis. *Cell Discov. 2016 21* **2**, 1–22 (2016).

17.   Aota, S. *et al.* Pax6 autoregulation mediated by direct interaction of Pax6 protein with the head surface ectoderm-specific enhancer of the mouse Pax6 gene. *Dev. Biol.* **257**, 1–13 (2003).

18.   Bhatia, S. *et al.* Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. *Am. J. Hum. Genet.* **93**, 1126–34 (2013).

19.   Inoue, M. *et al.* PAX6 and SOX2-dependent regulation of the Sox2 enhancer N-3 involved in embryonic visual system development. *Genes Cells* **12**, 1049–1061 (2007).

20.   Hayashi, S., Goto, K., Okada, T. S. & Kondoh, H. Lens-specific enhancer in the third intron regulates expression of the chicken delta 1-crystallin gene. *Genes Dev.* **1**, 818–828 (1987).

21.   Narasimhan, K. *et al.* DNA-mediated cooperativity facilitates the co-selection of cryptic enhancer sequences by SOX2 and PAX6 transcription factors. *Nucleic Acids Res.* **43**, 1513 (2015).

22.   Epstein, J., Cai, J., Glaser, T., Jepeal, L. & Maas, R. Identification of a Pax paired domain recognition sequence and evidence for DNA-dependent conformational changes. *J. Biol. Chem.* **269**, 8355–8361 (1994).

23.   Xu, H. E. *et al.* Crystal structure of the human Pax6 paired domain-DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding. *Genes Dev.* **13**, 1263–75 (1999).

24.   Coutinho, P. *et al.* Discovery and assessment of conserved Pax6 target genes and enhancers. *Genome Res.* **21**, 1349–59 (2011).

25.   Chow, R. L., Altmann, C. R., Lang, R. A. & Hemmati-Brivanlou, A. Pax6 induces ectopic eyes in a vertebrate. *Dev. Camb. Engl.* **126**, 4213–4222 (1999).

26.   Halder, G., Callaerts, P. & Gehring, W. J. Induction of Ectopic Eyes by Targeted Expression of the eyeless Gene in Drosophila. *Science* **67**, 1788–1792 (1995).

27.   Cunha, D. L., Arno, G., Corton, M. & Moosajee, M. The Spectrum of PAX6 Mutations and Genotype-Phenotype Correlations in the Eye. *Genes* **10**, (2019).

28.   Kejun Tang, H., Chao, L.-Y. & Saunders, G. F. Functional Analysis of Paired Box Missense Mutations in The PAX6 Gene. *Hum. Mol. Genet.* **6**, 381–386 (1997).

29.   Williamson, K. A. *et al.* Recurrent heterozygous PAX6 missense variants cause severe bilateral microphthalmia via predictable effects on DNA–protein interaction. *Genet. Med. 2019 223* **22**, 598–609 (2019).

30.   Fowler, D. M. *et al.* An Atlas of Variant Effects to understand the genome at nucleotide resolution. *Genome Biol.* **24**, 147 (2023).

31.   Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci.* **108**, 7896–7901 (2011).

32.   Wei, H. & Li, X. Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes. *Front. Genet.* **14**, (2023).

33.   Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

34.   Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).

35.   Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).

36.   Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).

37.   Giacomelli, A. O. *et al.* Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387 (2018).

38.   Staller, M. V. *et al.* A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Syst.* **6**, 444-455.e6 (2018).

39.   Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–206 (2015).

40.   Reece-Hoyes, J. S. & Walhout, A. J. M. YEAST ONE-HYBRID ASSAYS: A HISTORICAL AND TECHNICAL PERSPECTIVE. *Methods San Diego Calif* **57**, 441–447 (2012).

41.   Ji, X., Wang, L., Zang, D. & Wang, Y. Transcription Factor-Centered Yeast One-Hybrid Assay. *Methods Mol. Biol. Clifton NJ* **1794**, 183–194 (2018).

42.   Reece-Hoyes, J. S. *et al.* Enhanced yeast one-hybrid (eY1H) assays for high-throughput gene-centered regulatory network mapping. *Nat. Methods* **8**, 1059–1064 (2011).

43.   Kamachi, Y., Uchikawa, M., Tanouchi, A., Sekido, R. & Kondoh, H. Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development. *Genes Dev.* **15**, 1272 (2001).

44.   Wrenbeck, E. E. *et al.* Plasmid-based one-pot saturation mutagenesis. *Nat. Methods 2016 1311* **13**, 928–930 (2016).

45.   Redden, H. & Alper, H. S. The development and characterization of synthetic minimal yeast promoters. *Nat. Commun.* **6**, (2015).

46.   Crawford, K. & Bloom, J. alignparse: A Python package for parsing complex features from high-throughput long-read sequencing. *J. Open Source Softw.* **4**, 1915 (2019).

47.   Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 1–15 (2017).

48.   Livesey, B. J. & Marsh, J. A. The properties of human disease mutations at protein interfaces. *PLOS Comput. Biol.* **18**, e1009858 (2022).

49.   Ferrer-Costa, C., Orozco, M. & de la Cruz, X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.* **315**, 771–786 (2002).

50.   King, J. L. & Jukes, T. H. Non-Darwinian evolution. *Science* **164**, 788–798 (1969).

51.   Xu, W., Rould, M. A., Jun, S., Desplan, C. & Pabo, C. O. Crystal structure of a paired domain-DNA complex at 2.5 å resolution reveals structural basis for pax developmental mutations. *Cell* **80**, 639–650 (1995).

52.   Treisman, J., Harris, E. & Desplan, C. The paired box encodes a second DNA-binding domain in the paired homeo domain protein. *Genes Dev.* **5**, 594–604 (1991).

53.   Gerasimavicius, L., Liu, X. & Marsh, J. A. Identification of pathogenic missense mutations using protein stability predictors. *Sci. Rep. 2020 101* **10**, 1–10 (2020).

54.   Li, C., Zhi, D., Wang, K. & Liu, X. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. *Genome Med.* **14**, 115 (2022).

55.   Pejaver, V. *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Commun.* **11**, 5918 (2020).

56.   Richards, S. *et al.* Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).

57.   Brnich, S. E. *et al.* Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* **12**, 3 (2019).

58.   Lima Cunha, D. *et al.* PAX6 missense variants in two families with isolated foveal hypoplasia and nystagmus: evidence of paternal postzygotic mosaicism. *Eur. J. Hum. Genet.* **29**, 349–355 (2021).

59.   Ilsley, M. D. *et al.* Corrupted DNA-binding specificity and ectopic transcription underpin dominant neomorphic mutations in KLF/SP transcription factors. *BMC Genomics* **20**, 417 (2019).

60.   Dansault, A. *et al.* Three new PAX6 mutations including one causing an unusual ophthalmic phenotype associated with neurodevelopmental abnormalities. *Mol. Vis.* **13**, 511 (2007).

61.   Lee, S. *et al.* Impaired DNA-binding affinity of novel PAX6 mutations. *Sci. Rep. 2020 101* **10**, 1–10 (2020).

62.   Chen, J. & Zhu, J. [Analysis of PAX6 gene mutations in a Chinese family affected with congenital aniridia]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi Zhonghua Yixue Yichuanxue Zazhi Chin. J. Med. Genet.* **33**, 523–525 (2016).

63.   Pedersen, H. R. *et al.* The Cone Photoreceptor Mosaic in Aniridia: Within-Family Phenotype–Genotype Discordance. *Ophthalmol. Retina* **3**, 523–534 (2019).

64.   Pedersen, H. R. *et al.* Color Vision in Aniridia. *Invest. Ophthalmol. Vis. Sci.* **59**, 2142–2152 (2018).

65.   Tarilonte, M. *et al.* Parental Mosaicism in PAX6 Causes Intra-Familial Variability: Implications for Genetic Counseling of Congenital Aniridia and Microphthalmia. *Front. Genet.* **9**, (2018).

66.   Chou, C. M. *et al.* Biochemical Basis for Dominant Inheritance, Variable Penetrance, and Maternal Effects in RBP4 Congenital Eye Disease. *Cell* **161**, 634–646 (2015).

67.   Gill, G. & Ptashne, M. Negative effect of the transcriptional activator GAL4. *Nat. 1988 3346184* **334**, 721–724 (1988).

68.   Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, 9380 (2020).

69.   Livesey, B. J. & Marsh, J. A. Updated benchmarking of variant effect predictors using deep mutational scanning. *Mol. Syst. Biol.* **n/a**, e11474 (2023).

70.   Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

71.   Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. in (2021).

72.   Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).

73.   Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).

74.   Livesey, B. J. & Marsh, J. A. Interpreting protein variant effects with computational predictors and deep mutational scanning. *Dis. Model. Mech.* **15**, dmm049510 (2022).

75.   Manuel, M., Pratt, T., Liu, M., Jeffery, G. & Price, D. J. Overexpression of Pax6 results in microphthalmia, retinal dysplasia and defective retinal ganglion cell axon guidance. *BMC Dev. Biol.* **8**, 59 (2008).

76.   Davis, J. & Piatigorsky, J. Overexpression of Pax6 in Mouse Cornea Directly Alters Corneal Epithelial Cells: Changes in Immune Function, Vascularization, and Differentiation. *Invest. Ophthalmol. Vis. Sci.* **52**, 4158–4168 (2011).

77.   Schedl, A. *et al.* Influence of PAX6 Gene Dosage on Development: Overexpression Causes Severe Eye Abnormalities. *Cell* **86**, 71–82 (1996).

78.   Jones, L., López-Bendito, G., Gruss, P., Stoykova, A. & Molnár, Z. Pax6 is required for the normal development of the forebrain axonal connections. *Dev. Camb. Engl.* **129**, 5041–5052 (2002).

79.   Chauhan, B. K., Zhang, W., Cveklova, K., Kantorow, M. & Cvekl, A. Identification of Differentially Expressed Genes in Mouse Pax6 Heterozygous Lenses. *Invest. Ophthalmol. Vis. Sci.* **43**, 1884–1890 (2002).

80.   Li, X., Lalić, J., Baeza-Centurion, P., Dhar, R. & Lehner, B. Changes in gene expression predictably shift and switch genetic interactions. *Nat. Commun.* **10**, 3886 (2019).

81.   Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).

82.   Zhang, T. *et al.* Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. *PLoS Genet.* **16**, e1009009 (2020).

83.   Puchta, O. *et al.* Network of epistatic interactions within a yeast snoRNA. *Science* **352**, 840–844 (2016).

84.   Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009).

85.   Cheng, J. *et al.* MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).

86.   Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc. 2007 21* **2**, 31–34 (2007).

87.   Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267 (2014).

88.   Babraham Bioinformatics - Trim Galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

89.   Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).

90.   Pejaver, V. *et al.* Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am. J. Hum. Genet.* **109**, 2163–2177 (2022).

91.   Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).
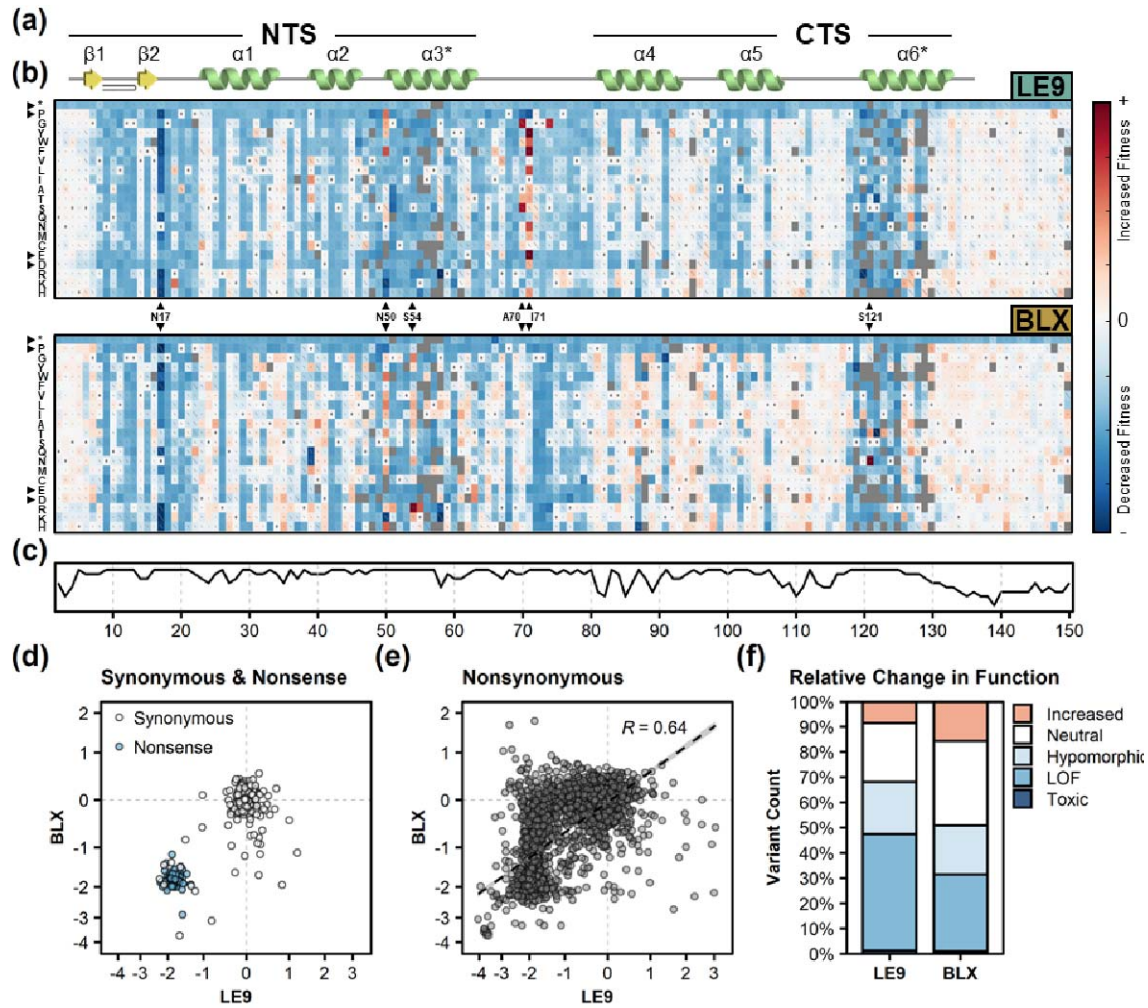
FIGURES

Fig 1. Experimental design of the deep mutational scan of PAX6 paired domain.

(a) A pool of barcoded variants of human PAX6 paired domain fused to the yeast activation domain (GAL4AD) was transformed into yeast containing a specific PAX6 DNA response element (bait) upstream of an antibiotic resistance gene (Geneticin^R). Transformants were grown together in a competitive growth assay in the presence of the antibiotic Geneticin, such that PAX6 variants with disrupted DNA binding were depleted in the population and those with unperturbed DNA-binding were conversely enriched. Changes in variant frequency were measured by counting barcodes using next-generation sequencing, and fitness scores were subsequently calculated using Enrich2[47]. (b) PAX6 bait sequences used in the DMS assays aligned using CLUSTALW. Each bait sequence consisted of three tandem repeats, indicated by horizontal broken lines. Sequence homology is highlighted in black. (c) (top) Linear structure of PAX6 depicting the paired domain (PD), homeodomain (HD) and proline-serine-threonine-rich transactivation domain (PSTD); the PAX6-GAL4 fusion protein (lower) constituting the "prey" component of the yeast one-hybrid assay. NLS, nuclear localisation sequence; PAX6 (1-150), PAX6 paired domain with short flanking sequence additions; GAL4AD; GAL4 activation domain.

Fig 2. Mutational Landscape of PAX6 paired domain

(a) Secondary structure of PAX6 paired domain depicting the N-terminal subdomain (NTS) and C-terminal subdomain (CTS) joined by a short flexible linker. α-helices and β-sheets are represented in green and yellow, respectively. A short β-turn is shown in the n-terminal portion of the NTS. Primary DNA-recognition helices in the NTS and CTS are denoted using asterisks (*). (b) Sequence-function maps generated following the deep mutational scan of PAX6 using LE9 (upper) and BLX (lower) as bait sequences in the presence of Geneticin, with the position in PAX6 (x axis) and amino acid variant (y axis) shown. Each tile represents a PAX6 variant, with fill colour indicating an increased (positive, red) or decreased (negative, blue) fitness score relative to the wild-type (score of 0, white). Variants not screened (no data) are shown as filled grey tiles. Grey dots within tiles depict the wild-type sequence at each position, and diagonal lines represent the standard error for the score, and are scaled such that a diagonal covering the entire square is the highest standard error on the plot. Highlighted are positions and amino acid substitutions with highly consistent or inconsistent fitness scores across DMS assays (black

triangles). (c) PAX6 evolutionary sequence conservation generated using ConSurf[91]. (d-e) Comparisons of fitness scores between the LE9 and BLX assays of synonymous (white) and nonsense (blue) variants (d), and all single nonsynonymous variants with Spearman correlation (e). Dashed lines represent the fitness score for wild-type PAX6. (f) Proportion of fitness scores in functional categories defined by comparison (FDR < 0.05) with wild-type and nonsense fitness score distributions (see methods). Increased, increased function; Neutral, wild-type-like; hypomorphic, intermediately perturbed function; LOF, complete loss of function; Toxic, toxic to yeast cells.
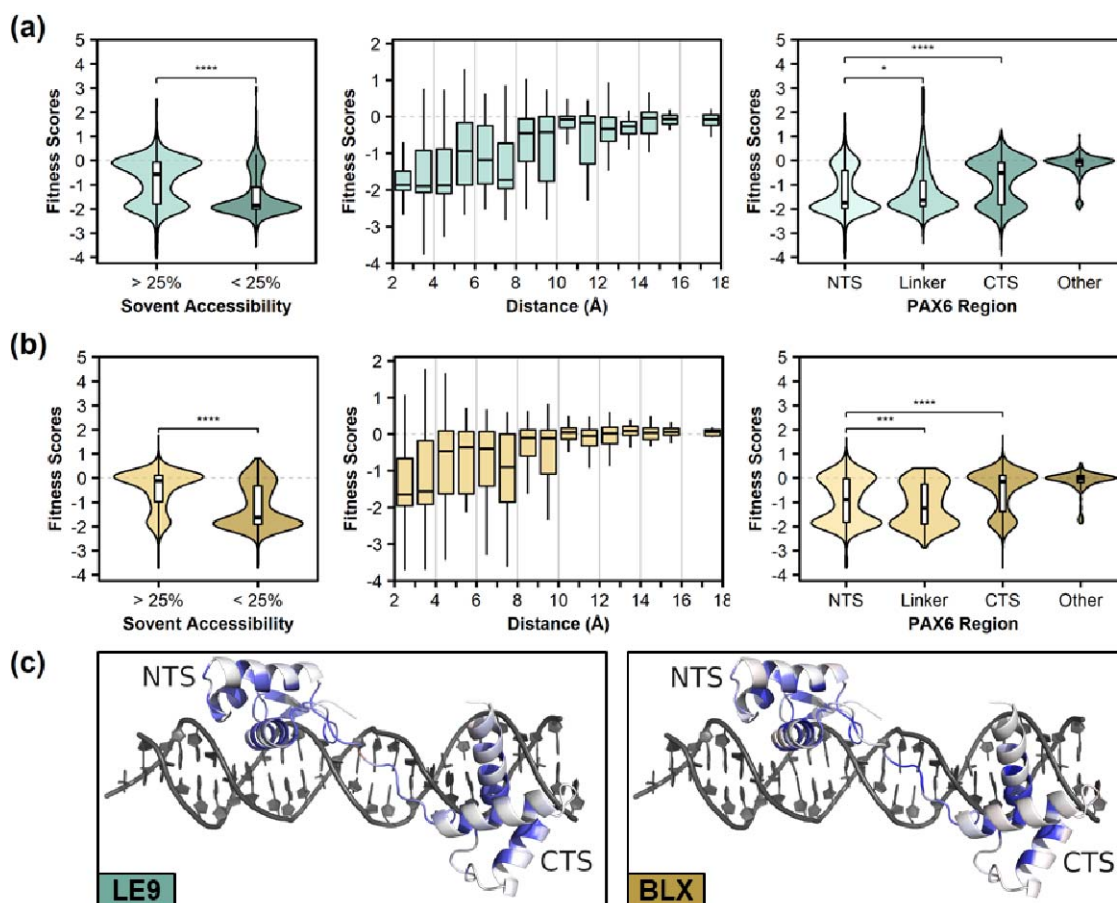
Fig. 3. Effects of mutations on protein structure

Distributions of variant fitness scores from the LE9 (a) and BLX (b) assays, grouped by solvent accessibility, distances from DNA, and region of PAX6. Dashed lines represent the fitness score for wild-type PAX6. Solvent accessibility was defined using a threshold of 25% solvent accessible surface area. Distances between each amino acid residue and the nearest non-hydrogen atom in DNA were measured and discretised into 0.5 Å bins. PAX6 regions were defined as NTS (4-64), Linker (64-80), CTS (81-133) and "other" (1-3 and 134-150). All p-values are calculated using the Wilcoxon test and adjusted using FDR. (c) PAX6:DNA heterodimers were generated by modelling LE9 (left) or BLX (right) in place of the DNA sequence crystallised in the 6PAX crystal structure[23]. Median fitness scores were calculated per residue and mapped onto each structure and coloured using a gradient where red and blue represent increased or decreased binding, respectively, and white as no change. NTS; N-terminal subdomain, CTS; C-terminal subdomain.
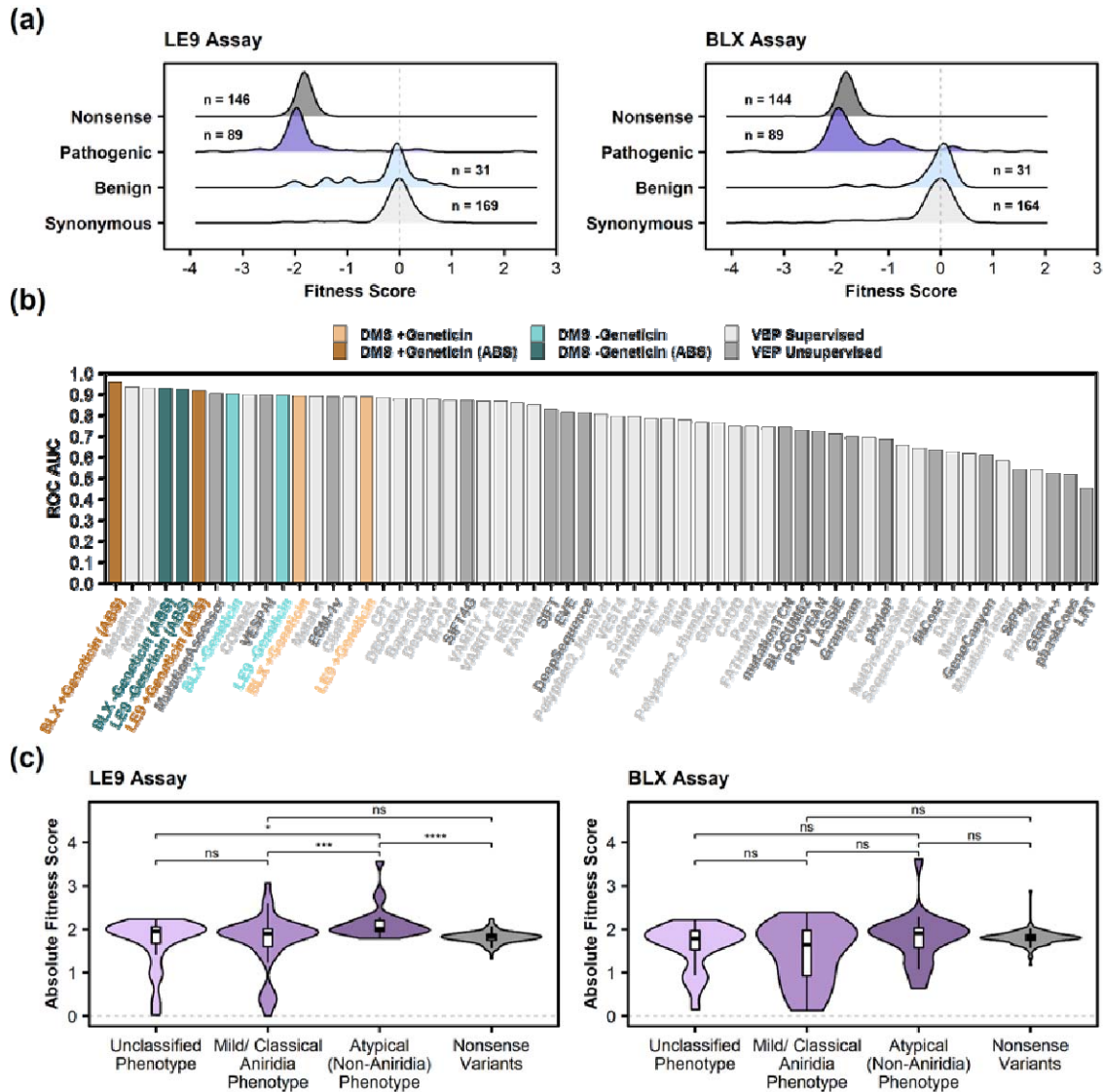
Fig 4. Classification of benign and pathogenic variants of PAX6 using DMS

(a) Human PAX6 variants (pathogenic: HGUeye, ClinVar, and HGMD; benign: gnomAD) and their corresponding fitness scores were compared for LE9 (left) and BLX (right) assays, using synonymous and nonsense variants to indicate neutral and loss of function, respectively. (b) Comparison of DMS assay data ability to accurately classify pathogenic and benign PAX6 variants against supervised and unsupervised variant effect predictors (VEPs), ranked by area under the receiver operating characteristic curve (ROC AUC). +/- Geneticin indicates the presence/absence of Geneticin in each DMS assay, respectively. "ABS" suffix denotes DMS datasets for which absolute variant fitness scores have been calculated and used in variant classification. (c) Absolute fitness scores for pathogenic PAX6 variants subdivided by the corresponding disease phenotype where accurate classifications could be made (see Methods)

for the LE9 (left) and BLX (right) assays. Dashed lines represent the fitness score for wild-type PAX6 in panels (a) and (c). All p-values are calculated using the Wilcoxon test and adjusted using FDR.
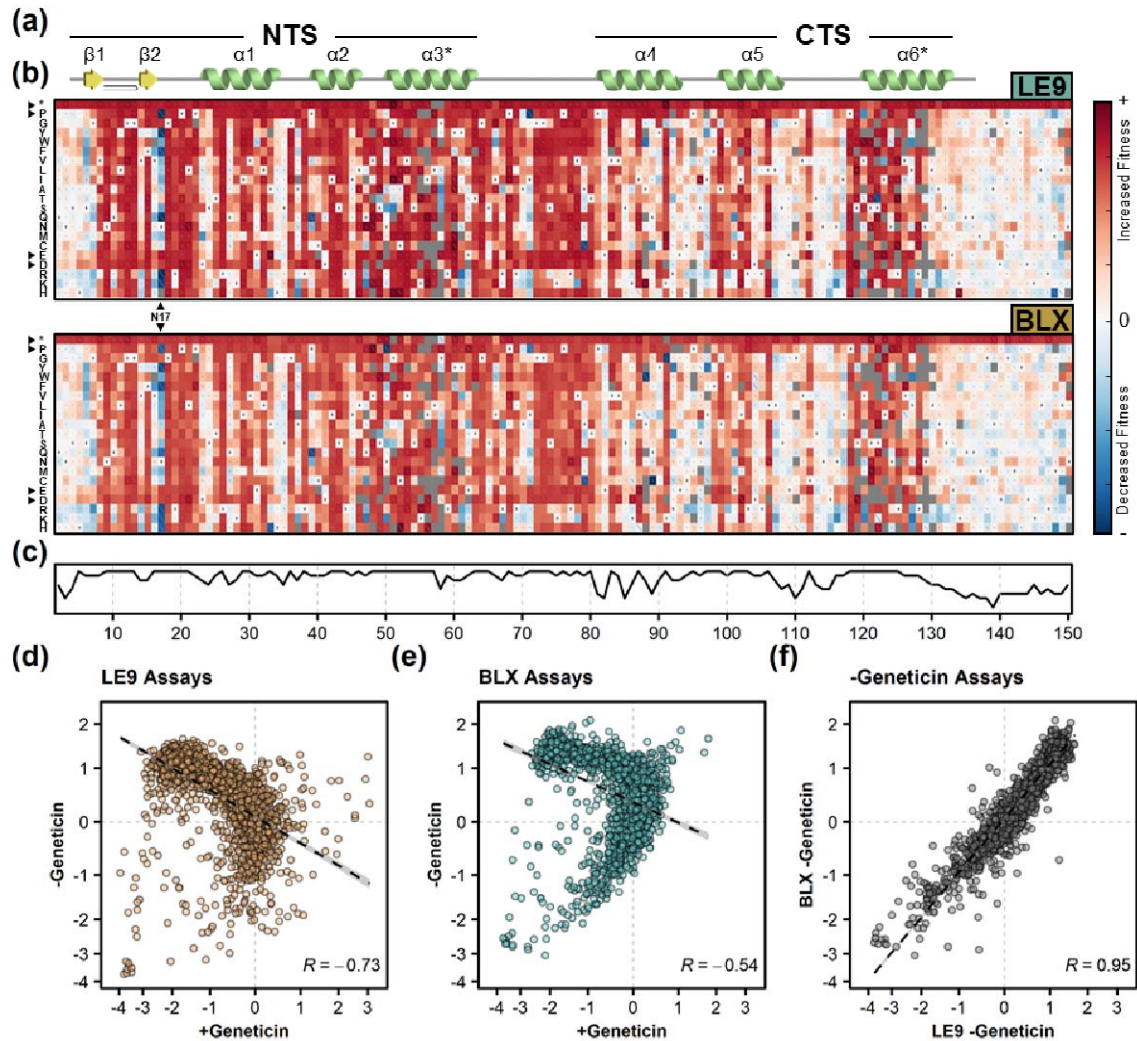
Fig 5. PAX6 DMS selection pressure without antibiotic

(a) Secondary structure of PAX6 paired domain, as in Figure 2. (b) Sequence-function maps generated following the deep mutational scan of PAX6 in the LE9 strain (top) and BLX strain (bottom) in the absence of Geneticin, with the position in PAX6 and amino acid variant plotted on the X and Y axes, respectively. Colours and other symbols as in Figure 2. (c) PAX6 evolutionary sequence conservation. (d-e) Correlations of variant fitness scores with (+Geneticin) and without antibiotic (-Geneticin) in the LE9 (d) and BLX (e) assays. (f) Correlation of variant fitness scores between LE9 and BLX assays performed in the absence of Geneticin.