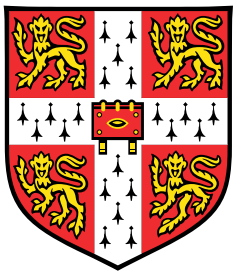
# Elucidating the mechanistic impact of single nucleotide variants in model organisms



**Omar Wagih**

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Sidney Sussex College

January 2018

I would like to dedicate this thesis to my loving parents Mohamed and Nagat, without whom none of my success would be possible.

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation does not exceed the specified length limit of 60,000 words as defined by the Biology Degree Committee.

Omar Wagih
January 2018

# Acknowledgements

This thesis has been a long and challenging journey and would not have been possible without the generous support and feedback of many people, for which I am deeply grateful.

First and foremost, I would like to extend my gratitude to my supervisor Pedro Beltrao for allowing me the opportunity to carry out my PhD studies in his research group. He has guided, challenged, and supported me through out the last four years, while giving me every chance to embark on my own endeavours. I would also like to thank all lab members of the Beltrao group, particularly Marta Strumiłło, David Ochoa and Marco Galardini. Marta, thank you for your enthusiasm and energy around the lab it has never been a dull day since you joined. David and Marco, thank you for all the work advice and the many fruitful conversations. I am greatly indebted to Brendan Frey, Daniele Merco and Andrew Delong for allowing me the opportunity to become part of the Deep Genomics family. I have learned an incredible deal under your supervision and I look forward to continuing doing so. I would also like to offer my profound gratitude and appreciation to Leopold Parts for being a remarkable mentor from my undergraduate through to my PhD. You have taught me much of what I know about academia.

To all my friends and learning partners who have made Cambridge a delightful place to be. Nils Eling, thank you for all the incredible adventures we have been on. I will miss the good times at the Blue Moon and late night marathons of Key & Peele, Workaholics, and the League. Alina Guna, thank you for the endless supply of laughs and always making me feel at home at times when I needed it the most. I would like to thank all other EBI and EMBL PhD students: Hannah Meyer, Dimitrios Vistos, Lara Urban, Daniel Elías Martín-Herranz, Jack Monahan, Maja Gehre and Niccolò Arecco that have made my experience both in Heidelberg and Cambridge a fantastic one. I would also like to thank my good friends back home. Christian Somody, thank you for the many trips we have been on and for always offering me advice and guidance. Melinda Schell, thank you for getting me through my first two years in Cambridge and always managing to lighten up my day.

I would like finally like to thank my family who have supported me unconditionally throughout my studies. To my mother, Nagat, thank you for the love, care and nourishment you have always provided me at all times. I would not have been able to get through this

without you. To my father Mohamed, you have always been my role model and someone I have looked up to. Your advice and guidance has helped me get to where I am today. To my sister Manar, thank you for your goofiness, wit and sense of humour. You have always made me smile and got me through many tough days. To my brother Ramy, thank you for all your love and support.

# Abstract

Understanding how genetic variation propagate to differences in phenotypes in individuals is an ongoing challenge in genetics. Genome-wide association studies have allowed for the identification of many trait-associated genomic loci. However, they are limited in their inability to explain the altered cellular mechanism. Genetic variation can drive disease by altering a range of mechanisms, including signalling networks, transcription factor (TF) binding, and protein folding. Understanding the impact of variants on such processes has key implications in therapeutics, drug development, protein engineering and more. This thesis aims to utilise computational predictors to shed light on how cellular mechanisms are altered in the context of genetic variation and better understand how they drive both molecular and organism-level phenotypes. Many binding events in the cell are mediated by short stretches of sequence motifs. The ability to discover these underlying rules of binding could greatly aid our understanding of variant impact. Kinase-substrate phosphorylation is one of the most prominent post-translational modifications (PTMs) which is mediated by such motifs. We first describe a computational method which utilises interaction and phosphorylation data to predict sequence preferences of kinases. Our method was applied to 57% of human kinases capturing known well-characterised and novel kinase specificities. We experimentally validate four understudied kinases to show that predicted models closely resemble true specificities. We further demonstrate that this method can be applied to different organisms and can be used for other phospho-recognition domains. The described approach allows for an extended repertoire of sequence specificities to be generated, particularly in organisms for which little data is available. TF-DNA binding is another mechanism driven by sequence motifs, which is key for the tight regulation of gene expression and can be greatly altered by genetic variation. We have comprehensively benchmarked current methods used to predict non-coding variant effects on TF-DNA binding by employing over 20,000 compiled allele-specific binding variants across 43 TFs. We show that machine learning-based approaches significantly outperform more rudimentary methods such as the position weight matrix. We further note that models for many TFs with distinct binding specificities were unable to accurately assess the impact of variants. For these TFs, we explore alternative mechanisms underlying TF-binding, such as methylation, co-operative binding, and DNA

shape that drive poor performance. Our results demonstrate the complexity of predicting non-coding variant effects and the importance of incorporating alternative mechanisms into models. Finally, we describe a comprehensive effort to compile and benchmark common sequence and structure-based predictors of mutational consequences and predict the effect of coding and non-coding variants in the reference genomes of *H. sapiens*, *S. cerevisiae*, and *E. coli*. Predicted mechanisms include the impact on protein stability, interaction interfaces, PTMs and TF-binding. These variant effects are provided through mutfunc, a fast and intuitive web tool by which users can interactively explore pre-computed mechanistic variant impact predictions. We validate computed predictions by analysing known pathogenic disease variants and provide mechanistic hypotheses for causal variants of unknown function. We further employ our predictions to devise gene burden scores in *S. cerevisiae* strains that are used to perform gene-phenotype association tests and uncover several known and novel associations.

# Table of contents

# List of figures

# List of tables

# Nomenclature

**Acronyms / Abbreviations**

ASB Allele-specific binding

AUROC/AUC Area under receiver operating characteristic

AUPRC Area under precision-recall curve

ChIP-chip Chromatin immunoprecipitation followed by microarray hybridization

ChIP-seq Chromatin immunoprecipitation followed by sequencing

CNV Copy number variation

DBD DNA-binding Domain

ELM Eukaryotic linear motif

EMSA Electrophoretic mobility shift assay

eQTL Expression quantitative trait loci

FDR False discovery rate

FN False negative

FP False positive

FPR False positive rate

GWAS Genome-wide association study

HapMap Haplotype map

HMM Hidden Markov model

IC Information content

indels Insertions or deletions

LD Linkage disequilibrium

MAF Minor allele frequency

mRNA Messenger RNA

MSA Multiple sequence alignment

MS Mass spectrometry

MSS Matrix similarity score

NGS Next-generation sequencing

NMD Nonsense-mediated decay

NSD Nonstop-mediated decay

PBM Protein binding microarray

PDB Protein data bank

PFM Position frequency matrix

phosphosite Phosphorylation site

PPI Protein-protein interaction

PPV Positive predictive value

PR Precision-recall

PTM Post-translational modification

PTV protein-truncating variant

PWM Position weight matrix

QTL Quantitative trait loci

RNA-Seq RNA sequencing

ROC Receiver operating characteristic

RSA Relative surface accessibility

SDR Substrate-determining residue

SELEX Systematic evolution of ligands by exponential enrichment

SLiM Short linear motif

SNP Single nucleotide polymorphism

SNV Single nucleotide variant

SVM Support vector machines

TF Transcription factor

TFBS Transcription factor binding site

TN True negative

TPR True positive rate

TP True positive

tRNA Transfer RNA

TSS Transcrpition start site

UV Ultraviolet

VCF Variant call format

VdW Van der Waals

VUS Variants of uncertain significance

# Chapter 1

# Introduction

The cell is an intricate system of interconnected components governing the phenotype through a combination of genetic and environmental factors. Changes to the genetic composition of an individual that are either inherited or arise during development are one of the major driving forces of phenotypic variability both at the organismal and molecular level. Even the simplest changes in the genome can alter protein levels or function, ultimately interfering with critical biological processes that then manifest as changes in molecular and organismal phenotypes. The advent of next-generation sequencing (NGS) and *omics* profiling technologies, has resulted in a substantial increase in the availability of genomes and corresponding phenotypic readouts [1, 2]. This has unravelled the era of modern genetics and has significantly propelled the bridging of the genotype-phenotype gap. The increased availability of such data has also led to the development of computational and statistical methods that are critical to implicating thousands of genetic loci to changes in phenotype [3]. Understanding the articulate relationship between genotype and phenotype is an ongoing challenge in genetics and one that is the basis of this thesis.

## 1.1 Genetic variation

Genetic variation refers to changes in the underlying DNA and range from single nucleotide variation (SNV) to larger chromosomal rearrangements, such as copy number variation (CNVs). SNVs involve a single substitution from one nucleotide to another, insertions or deletions occur when a stretch of sequence with one or more bases are either inserted or removed from the genome and CNVs occur when regions of the genome are repeated or deleted. SNVs are, by far, the most common form of genetic variation with over 14 million identified in the human genome alone [4]. As such, SNVs will be the primary focus of this

thesis and any reference to genetic variation henceforth can be assumed to be SNVs unless otherwise stated.

SNVs exist throughout the genome in both noncoding and coding regions and are responsible for a wide range of phenotypic consequences [5]. Those which are observed in coding regions can be categorized into *nonsynonymous* (or *missense*), *synonymous*, *nonsense* and *nonstop*. Nonsynonymous SNVs are those that result in an amino acid substitution, whereas synonymous SNVs, due to the redundancy of the genetic code, results in a change on the DNA level but not to the encoded amino acid. Although studies have shown that synonymous SNVs can alter messenger RNA (mRNA), splicing [6], mRNA stability [7] and protein folding [8] because no changes are made to the amino acid sequence they are generally often benign relative to their nonsynonymous counterparts [9]. The remaining categories include nonsense SNVs, which introduce a premature stop codon in the protein sequence and nonstop SNVs that results in the loss of a stop codon, both of which have detrimental effects on protein function [10].

Within a population, SNVs typically exhibit two alleles, where one is more frequently observed than the other. These are referred to as the *major* and *minor* alleles, respectively, and the frequency at which they occur in the population is referred to as the *allele frequency*. For instance, if an SNV with the minor allele, T, occurs in 10% of the population, it has a minor allele frequency (MAF) of 0.10. If the minor allele of an SNV is prevalent in the population, typically at a MAF of greater than 1%, it is often referred to as a single nucleotide polymorphism (SNP).

Genetic variation is acquired by an individual through one of two ways. *Germline* variants are inherited from maternal and paternal genomes during homologous recombination, whereas *somatic* variants arise throughout the lifespan of the individual. In diploid organisms such as human, the acquired SNV can be classified as *heterozygous* or *homozygous*, depending on whether the allele is observed on one or both copies of the genome, respectively.

## 1.2 Genotype-phenotype association

Over the past decade, there has been significant progress towards bridging the genotype-phenotype gap. This has been propelled by the development of sophisticated statistical methods aimed at associating genetic variation to complex phenotypic traits, both on the organismic and molecular level. Such methods have helped identify many causal variants in both human and model organisms, and the increasing availability of genotype and phenotype data is further enabling additional discoveries. Here, I briefly discuss the approaches taken by association-based methods, the advances made and challenges faced.

### 1.2.1 Genome-wide association studies

A key challenge of genetics is to identify the role risk factors play in Mendelian diseases that can be linked to a single gene, such as cystic fibrosis and muscular dystrophy, as well as common complex disorders that are instead governed by a more complex genetic architecture, such as heart disease and obesity. Linkage studies, a process involving the tracking of genetic variation through family lineages, have been pivotal at pinpointing causal genetic variants in Mendelian diseases such as cystic fibrosis [11] and tuberous sclerosis [12]. Such phenotypes are often highly penetrant, meaning the presence of the causal allele often implies a high likelihood of exhibiting the phenotype. However, these approaches often do not extrapolate well to phenotypes of a more complex genetic nature.

With the advent of sequencing and genotyping technologies, large-scale approaches employing statistical means to establish a map between genomic loci and phenotype have become increasingly abundant and key to identifying risk factors in human disease. Genome-wide association studies (GWAS) is the most common approach, which interrogates, on a genome-wide scale, millions of SNPs for associations to a trait of interest. Unlike linkage analysis, GWASs are able to dissect complex phenotypes due to the abundance of data involved and have been largely successful at identifying causal variants for a multitude of both Mendelian and complex diseases, those of which include sickle cell anaemia [13], type 2 diabetes [14], inflammatory bowel disease [15], multiple sclerosis [16, 17] and obesity [18, 19]. In addition to disease phenotypes, GWASs have also associated genetic variation with other physical traits, such as height [20], hair colour [21], and even facial morphology [22]. Another field to which GWAS is increasingly contributing to is that of pharmacogenomics where many genetic variants have been associated with drug efficacy, metabolism, and toxicity [23–25]. These results, in turn, can be used to modulate drug dosage to patients, giving rise to the era of personalised medicine. GWAS has also been applied to several plant species [26, 27] and bacterial species [28], shedding light on the genetic architecture of other organisms.

The design of a GWASs relies on the concept of linkage disequilibrium (LD), which is defined as the non-random co-occurrence of alleles of two SNPs in close proximity to one another and is caused by forces such as natural selection and random drift [29, 30]. Because genomic loci in close proximity are more likely to co-segregate during recombination, they are therefore more likely to co-occur. Regions in the genome with a group of co-occurring alleles are referred to as haplotype blocks, which can be leveraged to identify a set of representative or "tag" SNPs within each haplotype thereby avoiding redundancy and minimising the number of genotyped SNPs. The international haplotype map project (HapMap) has facilitated this process by cataloguing SNP allele frequencies and their associations with nearby SNPs in a

diverse set of individuals. The third and latest release from the HapMap project contains 1.5 million SNPs, genotyped across 1,397 individuals across 11 human populations [31]. The haplotype map streamlines the process of identifying tag SNPs and is key in designing an effective GWAS study.

GWASs are carried out by carefully defining two sets of individuals, those that harbour a particular trait (case) and those that do not or that harbour an alternative trait (control). Each group is then genotyped for tag SNPs, typically through SNP microarrays, and allelic frequencies of SNPs from each group are compared using contingency table methods, commonly the chi-squared test. Because of the high number of statistical tests carried out, rigorous multiple testing correction is often required with a stringent $p$-value selection criterion for identification of causal variants.

In contrast to more traditional approaches for surveying risk factors, such as linkage analysis, GWASs are able to comprehensively survey the genome in a hypothesis-free and unbiased manner, employing elegant statistical methods to uncover variant associations. GWASs are not limited to families and can instead leverage information in a large number of individuals from a diversity of outbred populations. In addition, GWASs are capable of identifying causal variants in low-penetrance phenotypes [32]. The endless possibilities of GWASs, combined with its robustness makes it a valuable tool for dissecting the complex underlying genetic architecture driving qualitative and quantitative traits.

### 1.2.2 Quantitative trait loci mapping

Quantitative trait loci (QTL) mapping is another association-based approach used for associating genetic variation with changes in phenotype. QTL studies are similar in nature to GWAS studies in that they both aim to identify genomic loci that correlate with changes in the phenotype. However, QTLs differ in that they typically underlie quantitative traits, which can range from morphological features such as height and weight to molecular-level phenotypes such as protein levels and gene expression. QTLs can be applied to outbred populations, such as human populations or experimental crosses carried out in model organisms, where variation in recombinant offspring is inherited from either the maternal or paternal genome. They have contributed significantly to further the interpretation of GWAS-derived variants [33] and the genetics of several model organisms including *S. cerevisiae* [34] and *D. melanogaster* [35] in which crosses can be easily achieved.

Expression QTLs (eQTLs) is one of the most commonly studied types of QTLs in which an allele is able to explain differences in gene expression. The genotype of an individual is first obtained through SNP genotyping or whole genome sequencing followed by a measure of the individuals mRNA levels, typically through RNA sequencing (RNA-Seq). For each

SNP, individuals are grouped based on the allele they carry and mRNA expression levels for each group are tested against each other using statistical methods, such as linear regression, to assess the significance of difference amongst the groups and the magnitude of the difference (effect size).

Brem et al. [34] carried out one of the first comprehensive eQTL studies in *S. cerevisiae*, in which microarrays were used to identify over 1,500 differentially expressed genes amongst recombinant offspring, over a third of which were associated with variation at one or more genomic loci. This set a precedent for and propelled the field of gene regulation. Today, QTL studies are carried out on a much larger scale. For instance, the GTEx consortia measured genotype and expression for 7,051 samples across 44 different human tissues carrying out millions of association tests and identifying tens of thousands of significant eQTLs [36].

New derivatives of QTL studies are constantly being developed, such as those that survey protein levels (pQTL) [37], histone modifications (hQTLs) [38] and methylation (mQTL) [39]. Such studies provide the ability to dissect the complex genetic architecture underlying the human genome and will allow us to better understand complex diseases and help develop targeted therapies.

### 1.2.3 Limitations of association-based methods

Despite the successes of both GWAS and QTL-based studies, they are not without limitations. For instance, rare variants require a significantly large sample size for appropriate statistical power, which may not always be feasible [40]. The phenotype being assessed must also display a substantial degree of variation within the population to be considered. Furthermore, due to the extremely large number of statistical tests being carried out combined with multiple testing, the increasing false positive rates further hinder the discovery of causal variants [40]. More importantly, a fundamental drawback of association-based approaches is their inability to pinpoint the underlying biological mechanisms driving the phenotypic change. Through assessing quantitative molecular phenotypes, QTL-based studies focusing on molecular phenotype offer a step closer to identifying the altered mechanism but are nevertheless unable to provide the exact altered mechanisms driving this change. For example, a GWAS-identified casual variant may also be identified as an eQTL, suggesting that alters a transcriptional regulatory mechanism. However, the exact mechanism by which this occurs remains unknown and could be due to a number of altered mechanisms such as changes in epigenetic marks or altered binding of regulatory factors. Thus, a complete understanding of how genetic variation is propagated to change in a phenotype requires the interpretation of molecular consequences of mutations.

## 1.3 Molecular phenotypes impacted by single nucleotide variants

GWAS and QTL-based studies have significantly accelerated our understanding of how genetic variation ties into phenotypic diversity. Yet, a complete understanding of the role of variation in phenotype requires going beyond what is contributed by such approaches. The lack of biological information provided by GWASs limits our ability to understand disease origins and develop treatments accordingly.

A causal variant can affect protein function in a number of ways. For example, it can alter gene expression through modifying key regulatory regions in the genome, affect the protein stability, disrupt interactions between proteins, and affect other key physiologically properties of proteins. Here, I describe a number of commonly-studied biological mechanisms by which genetic variation can alter protein function. I further discuss advances in the computational methodologies used to model these mechanisms and describe how they are employed to assess the impact of genetic variation.

### 1.3.1 Transcription factor binding

Transcription is a key process in the central dogma, governing the creation of mRNA from DNA the and thereby regulating the rate at which a gene is expressed. Regulatory sequences encode a series of instructions that direct transcription. A key challenge in genetics is to understand how the instructions encoded within sequence give rise to complex patterns of transcription and how this is affected by sequence context (e.g. tissue or cell type). Understanding this will allow us to understand the role of transcription in differentially expressed patterns of genes.

The intricate process of transcription revolves around DNA-binding proteins known as transcription factors (TFs), along with other factors that control chromatin structure. Together, they have a central role in controlling the initiation of transcription through the recruitment of the transcriptional machinery to core promoters of a gene. TFs carry this out by recognising short stretches of regulatory sequences, typically between 6-18 bp, that exist in close proximity to the promoter (*cis*-regulatory elements), or distal sequences, such as enhancers (*trans*-regulatory elements), that are brought into close proximity of the promoter with the aid of chromatin looping. This leads to the assembly of complexes that are responsible for carrying out transcription initiation.

### High-throughput approaches for assaying transcription factor binding

Characterization of TF binding sites (TFBSs) was initially carried out using techniques such as electrophoretic mobility shift assay (EMSA), where shifts in molecular weights through a polyacrylamide or agarose gel are used to determine the occupancy of the TF [41]. Although effective, and applicable both *in vivo* and *in vitro*, such low-throughput approaches do not scale. High-throughput approaches based on microarrays, chromatin immunoprecipitation (ChIP) and NGS [42] were soon after developed to facilitate large-scale identification of TF-occupied regions *in vivo* [42].

Developed in the early 1990s, systematic evolution of ligands by exponential enrichment (SELEX) was the first high throughput approach to assay protein-DNA interactions *in vitro* [43]. The process begins with the generation of a large fixed-width double-stranded oligonucleotide library, which is then incubated with the immobilised TF. Unbound oligonucleotides are then purged, typically using affinity chromatography [44], and bound oligonucleotides are eluted, subjected to PCR, and reincubated with the TF for subsequent rounds of binding and selection to ultimately identify the consensus binding sequence [45]. This process is repeated a number of times, where each round is referred to as a cycle. One of the inherent limitations of SELEX is that it promotes over-selection of high-affinity binders. Medium and low-affinity binders are critical to understanding the full extent of binding specificity and shed light on the position-specific variability [46]. This poses an issue since, in the classical SELEX methodology, such sequences are eventually purged out. Modern SELEX derivatives, known as SELEX-seq, have mitigated this effect by incorporating massively parallel sequencing of bound DNA after each round of selection [47, 48] (Figure 1.1a). This, in turn, reduces the number of cycles required and allows for identification of sequences with varying degree of binding affinity.

With the advances of methods for production and fluorescent detection, modern DNA microarrays were born in the late 1990s. Protein binding microarrays (PBM) is an approach to detect *in vitro* TF-bound DNA sequences [49]. Contrary to SELEX, the DNA array containing a library of immobilised double-stranded oligonucleotides are exposed to a purified and expressed TF of interest, which is modified to carry an epitope tag. After multiple rounds of washing away unbound material, the array is exposed to a fluorophore-conjugated antibody specific to the epitope (e.g. glutathione *S*-transferase), resulting in the fluorescing of bound sequences [50] (Figure 1.1b). To eliminate any biases related to the probed double-stranded oligonucleotides, a parallel control experiment is often carried out in which DNA is directly stained using a second fluorescence signal, such as the fluorescent dye Sybrgreen I, which is specific to double-strand DNA [51]. Fluorescent signals from both the control and treatment

experiments are used to obtain a normalised score indicating TF protein-binding. PBMs have since been applied to systematically successfully identify TFBSs [52, 51] (Figure 1.1b).

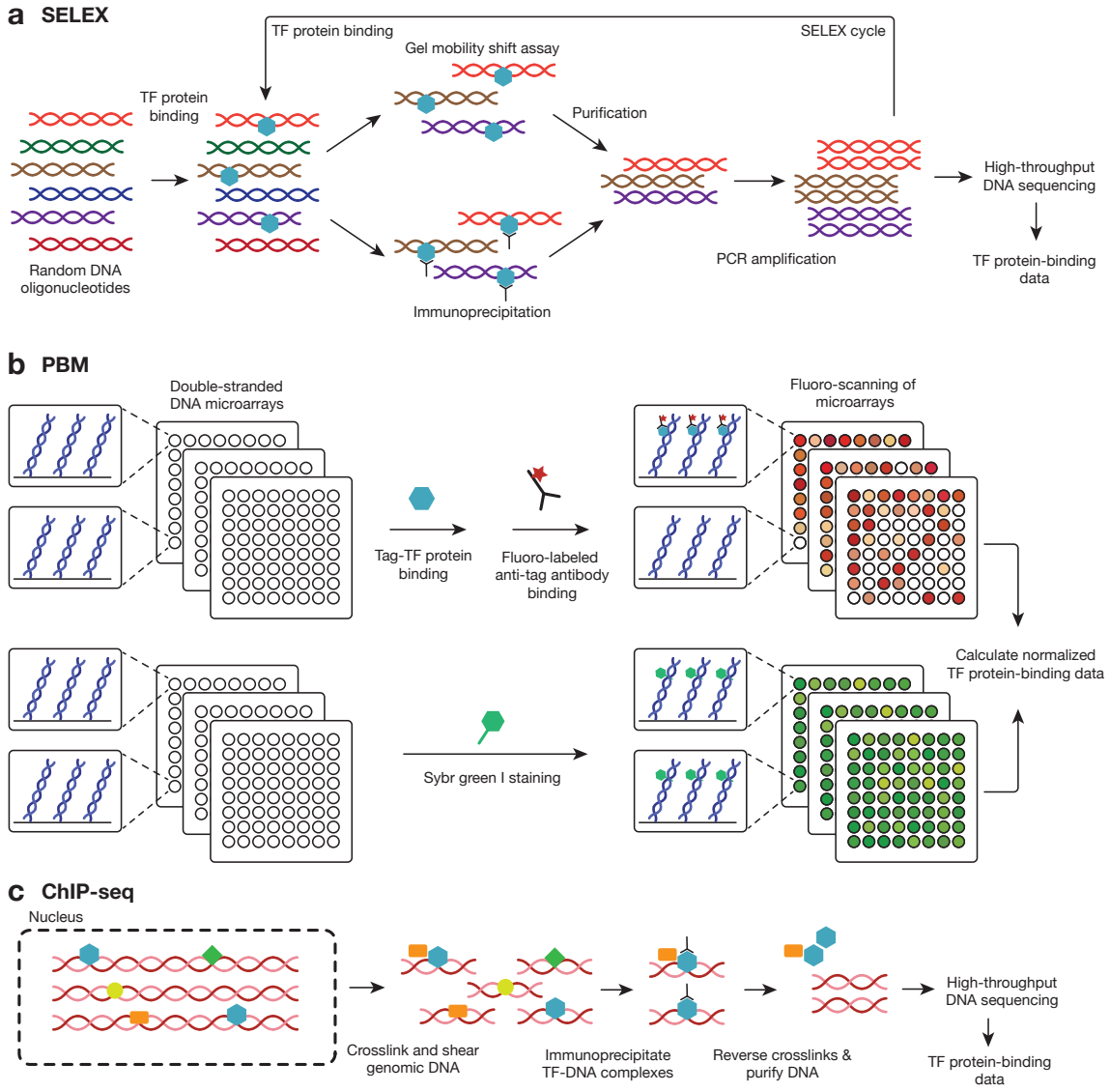



Fig. 1.1 Diagrams of common high-throughput approaches for assaying TFBSs, including (a) SELEX (b) PBMs and (c) ChIP-seq.

Methods that utilise ChIP were later developed that were capable of assaying TF-binding both *in vitro* and *in vivo*. ChIP followed by microarray hybridization (ChIP-chip) is one such approach that utilises DNA microarrays. Here, the TF is first cross-linked to the genomic DNA using formaldehyde fixation. Cells are then subject to lysis and DNA is sheared by sonication or enzymatic digestion into fragments of 150-500 bp [53]. A TF-specific antibody is then used to precipitate the TF-DNA complexes and heat can be used to reverse

the generated formaldehyde-generated cross-links. The remaining DNA is purified and labelled with a fluorescent probe that is then exposed to a DNA microarray containing a library of potentially complementary single-stranded DNA. To increase the signal-to-noise ratio, dynamic range and leverage advances in sequencing technology, the ChIP sequencing (ChIP-seq) protocol was developed. Instead of purified DNA being labelled and exposed to a microarray, it is instead subject to high throughput DNA-sequencing [54] (Figure 1.1c). It is noteworthy to mention that the performance of ChIP-based methods largely hinges on the availability and quality of the TF-antibody, crosslinking efficiency and the native abundance of the TF [54].

### Computational methods for modelling transcription factor specificity

The mode by which TFs recognise their binding sites *in vivo* is a non-trivial and ongoing challenge in the field of gene regulation. The DNA-binding domain (DBD) of many TFs exhibit a high degree of sequence specificity towards their binding sites, which is driven primarily by a unique chemical signature displayed by the base pairs of a binding site [55]. This sequence specificity is thought to be driven primarily by favourable hydrogen bonds, Van der Waals (VdW) contacts and electrostatic interactions between residue side chains in the DBD and the bases in the binding site [56]. The abundance of TFBS profiling data made available through methods like ChIP-seq and SELEX have propelled the systematic discovery of sequence specificity through computational approaches aimed at extracting underlying *de novo* sequence motifs for binding [57].

Computational approaches have utilised known binding sites of a TF to model the properties of its sequence specificity, thereby allowing for potential binding sites in the genome to be predicted *in silico*. Consensus sequences were the first approach to modelling TF specificity. Given a set of fixed-width aligned sequences, the consensus sequence simply reflected the most common base at each position (Figure 1.2a-b). To account for position-specific variability, position weight matrices (PWMs) were introduced in 1982 by Gary Stormo [58] as an alternative approach to consensus sequences, and have been one of the first approaches to quantitatively model TF-binding [59]. For a given set of fixed-width binding sites for a TF, a PWM is a matrix reflecting the log likelihood of observing each given base (A, T, C and G) at a position.

PWMs are constructed by tallying up the position-specific counts for each base, termed position frequency matrix (PFM). The probability of observing a base at a position is computed and a PWM is constructed by computing the log ratio of observed to expected probability (Figure 1.2c-d). The PWM can then be used to provide a quantitative score reflecting the likelihood of binding (Figure 1.2e).

The TF sequence specificity is often graphically represented as a sequence logos, which reflect the position-specific conservation of bases. The conservation at a given position is computed as the Shannon entropy, reflecting the information content (IC). Each position in a sequence logo is depicted as a stack of letters, where the height of each letter is relative to its relative frequency of the IC (Figure 1.2f).

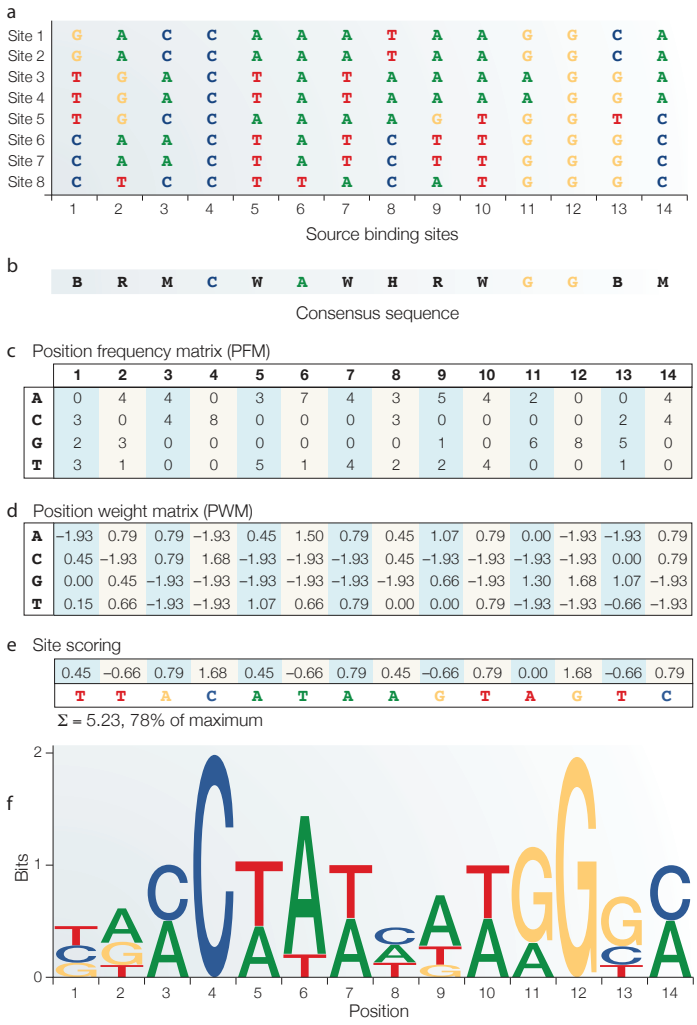



Fig. 1.2 Modelling TF-binding using PWMs. (a) Given a set of binding sites for a TF, (b) the consensus sequence provides a qualitative approach of describing the TF specificity. (c-d) PFMs and PWMs instead model specificity quantitatively, which can then (e) be used to score potential binding sites. (f) The specificity of a TF can be visualised using the sequence logo. Figure adapted from [60].

Today, PWMs remain the *de facto* standard for predicting TFBSs, offering a flexible, and intuitive approach to modelling TF sequence specificity. However, they do possess several drawbacks. First, a PWM assumes positional independence, meaning that each

position independently contributes independently to the final score. Many studies have, however, shown that co-dependent positions do exist and play a significant role in binding affinity [61–63]. Second, and more importantly, not all regions predicted as TFBSs sites will correspond to functional regulatory sites, particularly in short length motifs where matches are more likely to occur by random chance. This results in a high false positive rate. Other limitations of PWMs include that they do not account for other mechanisms of binding such as the sequence context around the immediate motif, which has been shown to facilitate binding [64, 65].

Adapted versions of PWMs have been developed in an attempt to reduce the effects of some of these drawbacks. For example, the dinucleotide weight matrix was developed to account for co-dependency between positions [66]. Predictions from PWMs have also been limited to regions more likely to be functional, such as those with regions of high sequence conservation, or phylogenetic footprints, in attempt to increase identification of functional binding sites. More recently, sophisticated approaches employing hidden Markov models (HMMs), and machine learning including support vector machines (SVMs) [67], deep learning [68, 69], and random forests [70] have been used to model TF specificity and predict TFBSs. Such approaches have been trained on both regulatory and random genomic sequences to learn far more complex patterns underlying the TFBSs, beyond what is capable by PWMs.

Computational predictors of TFBSs have significantly advanced over the past several decades. However, much work is yet to be done for improved sensitivity and accuracy. Sequence specificity alone is not sufficient to capture all variability observed in TF-binding. Specifically, many other factors have been shown to dictate binding for many TFs such as geometric complementarity between DNA and the DBD [71], epigenetics [72], chromatin accessibility [73] and cofactor availability [74]. For instance, a recent study that incorporated three-dimensional features of DNA shape was able to bring a substantial improvement to the performance of PWMs [71]. Incorporating a larger number of such mechanisms into TF-binding models will be key to accurate modelling of TF binding in the future.

### Genetic variation in TFBSs

Given the structured patterns encoded within TFBSs, genetic variation can alter sequence-specific binding sites and either abolish or create a novel binding site. Known disease variants have been found to be enriched in both *cis* and *trans*-regulatory regions [75, 76] and specifically in TFBSs [77]. Understanding how variation alters binding is therefore crucial to understand the exact mechanisms underlying aberrant gene regulation.

| Phenotype | Variant | TF | Downstream gene | Effect | Reference |
|---|---|---|---|---|---|
| Haemophilia B Leyden | chrX:138612890A>T | ONECUT1 | *F9* | Loss | [78] |
| Colorectal cancer | chr10:90749256G>A | SP1 | *FAS* | Loss | [79] |
| Delta-thalassemia | chr11:5255790A>G | GATA1 | *HBD* | Loss | [80] |
| Breast cancer | chr11:69331642C>G | ELK4 | *CCND1* | Loss | [81] |
| Melanoma | chr11:111957523C>T | ELF1 | *SDHD* | Loss | [82] |
| Congenital erythropoietic porphyria | chr10:127505271A>G | GATA1 | *UROS* | Loss | [83] |
| Maturity-onset diabetes of the young | chr12:121416289A>T | HNF4A | *HNF1A* | Loss | [84] |
| Bernard-Soulier syndrome | chr22:19710933C>G | GATA1 | *SFN* | Loss | [85] |
| Alpha-thalassemia | chr16:209709T>C | GATA1 | *HBA1* | Gain | [86] |
| Familial combined hyperlipidemia | chr8:19796725T>C | POU2F1 | *LPL* | Loss | [87] |
| Systemic lupus erythematosus | chr1:172627498C>T | CEBPB | *FASLG* | Loss | [88] |
| Osteoarthritis | chr20:34026064G>T | YY1 | *GDF5* | Loss | [89] |
| Haemophilia B Brandenburg | chrX:138612869G>C | HNF4A | *F9* | Loss | [90] |
| Coagulant factor VII deficiency | chr13:113760095T>G | HNF4A | *F7* | Loss | [91] |
| Insulin resistance | chr3:12386337C>T | PRRX1 | *PPARG* | Loss | [92] |
| Prostate cancer | chr8:128531689A>T | FOXA1 | *MYC* | Gain | [93] |
| Hereditary persistence of fetal haemoglobin | chr11:5271262A>G | GATA1, TAL1 | *HBG1* | Gain | [94] |
| Pancreatic agenesis | chr10:23508446A>C | PDX1 | *PTF1A* | Loss | [95] |
| Type 2 diabetes | chr11:72432985G>A | PAX6 | *ARAP1* | Gain | [96] |
| Asthma and autoimmune diseases | chr17:38029120G>C | CTCF | *ZPBP2* | Loss | [97] |
| Pierre Robin syndrome | chr17:68676303T>C | MSX1 | *SOX9* | Loss | [98] |
| Hirschsprung disease | chr22:38412781G>C | SOX10 | *SOX10* | Loss | [99] |
| Beta-thalassemia | chr11:5248372G>A | GATA1 | *HBB* | Loss | [100] |
| Treacher Collins syndrome | chr5:149736964C>T | YY1 | *TCOF1* | Loss | [101] |
| Nonsyndromic cleft lip | chr1:209989270G>A | TFAP2A | *IRF6* | Loss | [102] |

Table 1.1 Examples of variants linked to variable TF binding.

Early approaches used mutagenesis combined with EMSA to assay the impact of sequence variation. The first of such studies appeared in the early 1990s which investigated the role disease variants in the gain or loss of TF binding, for a limited number of TFs with well-characterized DNA binding specificities at the time. One of the first was that by Martin et al., who attributed the increased expression of the *HBG* gene in hereditary persistence of fetal haemoglobin to a variant in its core promoter that altered binding specificity to create a novel binding site for GATA1 TF [94] (Table 1.1). Another example is that of Crossley et al. where a haemophilia-associated variant directly upstream of the *F9* gene transcription start site (TSS) disrupts the binding site for androgen receptor (*AR*) (Table 1.1). A larger list of examples is listed in (Table 1.1).

Today, the extensive availability of TF-binding specificity models has allowed for *in silico* prediction of variants impacting TFBSs. A typical way of approaching this is to the assess the difference between the PWM score of two alleles, where a larger difference indicates a larger impact on TF-specific binding [77]. Other methods such as gkmSVM [103], DeepSEA [69] and DeepBind [68] have devised similar approaches using machine learning-based predictors, instead of PWMs. Such approaches are limited in their interpretability since they employ black box approaches, the direct impact of a variant cannot be graphically represented. They do, however, harness the ability to model far more complex binding mechanisms.

A relatively novel experimental approach to characterising the impact of genetic variation on TF binding is the analysis of allele-specific binding (ASB) in ChIP-seq data [104, 105]. Given ChIP-seq data measuring the occupancy for a particular TF in a sample (tissue or cell-line), along with the corresponding genotype for the sample, one can identify heterozygous variants and count the number of ChIP-seq reads that map to the reference and alternate alleles. Loci exhibiting significant imbalances in mapped read counts are then classified as ASB events, in which one allele exhibits a significantly lower number of mapped reads compared to the other; in particular, significance here can be tested using a binomial test [104, 106]. Identified ASB events are not necessarily causal and the observed effect can instead be caused by another variant in LD, epistatic effects, whereby multiple variants contribute to the effect [107] or epigenetic effects influencing binding [108]. Nevertheless, ASB mapping provides an elegant alternative to prioritizing causal variants and assessing the impact of variants on TF binding.

TF binding is a complex mechanism that is collectively driven by a number of factors including sequence specificity. Altered TF binding has the ability to alter gene expression and future improvement to binding predictors will allow us to more accurately pinpoint causal variants.

### 1.3.2 Post-translational modifications

Post-translational modifications (PTMs) are reversible or irreversible chemical modifications made to a protein and are involved in a multitude of biological functions such as cellular signalling [109], protein folding and degradation [110] and metabolism [111] to name a few. PTMs act as molecular switches by introducing conformational changes to the protein's structure, promoting or disrupting protein interactions or altering protein localization [112, 113] ultimately extending the functional repertoire of the proteome. Over 200 different types of PTMs have been characterized [114] and amongst the most commonly occurring include phosphorylation, acetylation and ubiquitination (Table 1.2). Modified sites have been well-characterised and collated into resources such as UniProt [115], dbPTM [116] and PhosphoSitePlus [117]. Commonly, PTMs are reversible and involve two enzymes a writer, responsible for attaching the modification and an eraser responsible for removing the modification. PTMs are often attached to side chains of certain amino acids that act as strong (S, T, Y, R, K, H, D, E, M, C) or weak (N, Q) nucleophiles. For instance, acetylation involves the addition of an acetyl group to lysine residues (Table 1.2).

| Modification | Number of proteins | Number of sites | Commonly modified residues |
|---|---|---|---|
| Phosphorylation | 19,655 | 231,160 | S, T, Y |
| Ubiquitination | 9,022 | 63,729 | K, C |
| Acetylation | 7,555 | 22,762 | K, A, M |
| Methylation | 5,607 | 14,807 | R, K, C |
| Sumoylation | 2,155 | 6,987 | K |
| *O*-linked Glycosylation | 764 | 4,738 | T, S, K |
| *N*-linked Glycosylation | 1,134 | 2,976 | N, K, D |
| Disulfide bond | 167 | 1,136 | C |
| *S*-nitrosylation | 448 | 749 | C |
| Hydroxylation | 38 | 274 | P, K, N |
| Palmitoylation | 93 | 179 | C |
| Pyrrolidone carboxylic acid | 82 | 83 | Q |
| Gamma-carboxyglutamic acid | 10 | 82 | E |
| *S*-Nitrosylation | 70 | 82 | C |

Table 1.2 Identified PTM sites in human collected from public databases dbPTM [118] and PhosphoSitePlus [117]. For each PTM, the number of proteins with at least one site and the total number of sites are shown. Only the top 15 PTM types, as defined by the number sites are shown.

Given the functional relevance of PTMs, it is evident that variants affecting PTM sites can alter the modification through disrupting the modifiable site and affecting downstream

function. Indeed there have been many reported cases where mutations of PTM sites were directly associated with disease phenotypes. For instance, in the androgen receptor (AR), loss of acetylation has been associated with spinal and bulbar muscular atrophy, a neurodegenerative disorder affecting muscle movement. In the wild-type AR, mutations of residues K630, K632 and K633 to alanine disrupt acetylation sites and have been shown to significantly impede nuclear translocation. Furthermore, K632A and K633A mutants aggregate and co-localise with other proteins (HSP70/HSP90) resulting in loss of proteasome function [119]. Another example is that of the prion protein (PRNP), where the substitution T183A was implicated in spongiform encephalopathy through the loss of *N*-glycosylation [120]. More systematic studies have shown that PTM sites are under strong negative selection and are significantly enriched in disease mutations [121]. The study of PTMs in the context of genetic variation can, therefore, offer insight into altered mechanisms and help identify causal mutations.

### Kinase-substrate phosphorylation

Phosphorylation is the most prominent PTM and involves the transfer of a phosphate group ($PO_4^{3-}$) to different amino acids, including serine, threonine or tyrosine residues of proteins. In humans, this process is catalysed by over 500 protein kinases [122], which regulate a spectrum of function. This process is crucial for the regulation of a diverse range of cellular process, including growth, metabolism, cell cycle progression, differentiation, and apoptosis [122].

The mode by which protein kinases recognize specific target residues depends on numerous factors such as co-expression, residue surface accessibility, and other PTMs [123]. More importantly, kinases have been shown to have preferences for certain amino-acids flanking the central phospho-acceptor residue [123]. These preferences define the kinase-substrate specificity often referred to as the kinase-substrate 'motif', which were initially defined by searching for consensus patterns among a set of known target phosphosites. For example, the cyclin-dependent kinase (CDK2) is known to preferentially target the motif [ST]PX[RK] (a proline at position +1, any amino-acid at position +2 and an arginine or lysine at position +3) [123].

Kinase-substrate motifs were initially employed to predict potential targets for kinases [124]. However, to allow for the modelling of kinase-substrate specificity and quantitative prediction of potential targets, PWMs have been employed. Similar to TFs, using a set of target phosphosites sequences a PFM is constructed, where the observed amino acid frequencies are tallied up for each position. This matrix can then be used to generate the PWM by calculating the log ratio of observed versus expected relative frequencies. Generated

PWMs can be used to provide a quantitative score for a peptide describing the likelihood of kinase-specific phosphorylation. Sophisticated methods that employ machine learning approaches such as neural networks [125], SVMs [126] and random forests [127] have since been developed that similarly, model kinase sequence specificity and successfully predict putative target phosphosites with a higher degree of accuracy [125].

### The role of genetic variation on kinase-substrate phosphorylation

Disease variants occurring in kinase-substrate sequence motifs have been shown to rewire phosphorylation networks and potentially drive disease [148, 146]. Variants falling within specificity determinants could alter kinase-substrate phosphorylation either by diminishing an existing site (loss-of-binding) or introducing a novel binding site (gain-of-binding) (Figure 1.3a). By altering phosphorylation, these variants individually or collectively interfere with underlying signalling networks that may contribute to disease properties.

Several studies have described disease mutations altering key residues in kinase-substrate phosphorylation motifs. For instance, mutations in protein-tyrosine phosphatase 1B (PTPN1) have frequently been associated with type 2 diabetes. Specifically, the rare P387L substitution disrupts phosphorylation of the phosphosite S386, resulting in altered phosphatase activity [131] (Figure 1.3b). Another example is that of KCNH2, a member of the potassium voltage-gated channel family. Here, the substitution K897T introduces a novel phospho-acceptor residue for the Akt protein kinase, resulting in altered protein regulation that contributes to cardiac arrhythmias (Figure 1.3b). A larger set of examples are summarised in Table 1.3. In addition to these examples, recent analyses of cancer driver mutations in regions flanking the phospho-acceptor have implicated phosphorylation as an altered mechanism in disease [148].

### Predicting the impact of variants on kinase-substrate phosphorylation

There have been a number of studies systemically exploring the impact of phosphorylation-associated variation and numerous resources have been developed to categorize likely causal variants in phosphorylation. The PTMVar dataset provided within the PhosphoSitePlus database collates disease mutations curated in the UniProt resource [149] that occur in the vicinity of experimentally identified phosphosites, identifying over 19,000 variants falling within seven residues of a phosphosite [117]. The PTM-SNP database more generally identifies ~180,000 variants occurring in proximity to any PTM-modified residues including phosphorylation [150].

While such documented variants are more likely to play a role in altering signalling networks, additional information on whether they impact underlying mechanisms driving

| Phenotype | Gene | Variant | Phosphosite | Affected kinase(s) | Effect | Reference |
|---|---|---|---|---|---|---|
| Obesity | *PPARG* | P113Q | S112 | MAPK1; MAPK8 | Loss | [128] |
| Mycobacterial diseases | *STAT1* | L706S | Y701 | SRC | Loss | [129] |
| Advanced sleep phase disorder | *PER2* | S662G | S662 | CSNK1D; CSNK1E | Loss | [130] |
| Type 2 Diabetes | *PTPN1* | P387L | S386 | CDK1 | Loss | [131] |
| Prostate Cancer | *NKX3-1* | R52C | S48 | PRKCA | Loss | [132] |
| Multiple cancer types | *TP53* | P47S | S46 | MAPK14 | Loss | [133] |
| Multiple cancer types | *OGG1* | S326C | S326 | | Loss | [134] |
| Multiple cancer types | *CCND1* | T286R | T286 | GSK3B | Loss | [135] |
| Multiple cancer types | *CDKN1A* | D149G | S146 | AKT1; PRKCA | Loss | [136] |
| Cardiac arrhythmias | *KCNH2* | K897T | T897 | AKT1 | Gain | [137] |
| Dilated cardiomyopathy | *PLN* | R14C | S16 | PRKACA | Loss | [138] |
| Masculinization | *AR* | R407S | S407 | | Gain | [139] |
| Osteoporosis | *BDNF* | V66M | T62 | CHEK2 | Loss | [140] |
| Rett syndrome | *MECP2* | R306C | T308 | PRKACA | Loss | [141] |
| Breast cancer | *GAB1* | T387N | T387 | | Loss | [142] |
| Cholelithiasis | *ABCB4* | T34M | T34 | PRKACA; PRKCA | Loss | [143] |
| De Vivo disease | *SLC2A1* | R223W | S226 | PRKCA | Loss | [144] |
| Fine-Lubinsky syndrome | *MAF* | P59H | T58 | GSK3A; GSK3B | Loss | [145] |
| Multiple cancer types | *TP53* | R213Q | S215 | PRKCD | Loss | [146] |
| Multiple cancer types | *TP53* | R282W | T284 | AURKA; AURKB | Loss | [146] |
| Multiple cancer types | *CLIP1* | E1012K | S1009 | CSNK2A1 | Loss | [146] |
| Multiple cancer types | *CTNNB1* | S37C | S33 | GSK3A; GSK3B | Loss | [146] |
| Multiple cancer types | *CTNNB1* | G34R | S37 | CAMK2A | Gain | [146] |
| Autism | *UBE3A* | T508A | T508 | PRKACA | Loss | [147] |

Table 1.3 Examples of disease variants linked to altered kinase-substrate phosphorylation.

phosphorylation is not provided. As such, several resources were developed that take into account the modelling of kinase-binding preferences. For example, Ryu et al. modelled kinase-binding preferences using SVMs and used them to assess the impact of over 33,000 nonsynonymous mutations obtained from Swiss-Prot [151, 149]. Impact predictions were then provided through the PhosphoVariant database [151]. Similarly, Ren et al. used the kinase-specific phosphorylation site predictor GPS 2.0 [152] to assess the impact of over 64,000 nonsynonymous variants from the dbSNP resource [4] and provide results into the PhosSNP database [153].

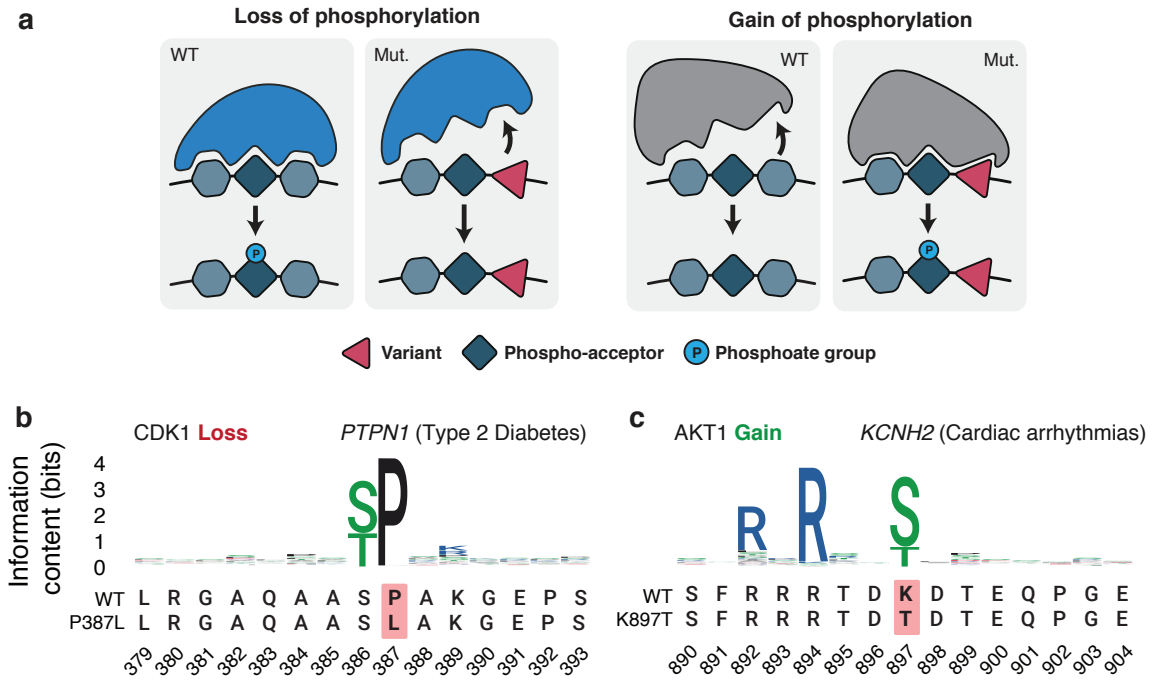



Fig. 1.3 Genetic variation in kinase-substrate phosphosites. (a) Substitutions of residues critical to the sequence specificity of a can either result disrupt existing phosphosites or create new ones. (b-c) Two examples of disease variants altering specificity determinants. (b) The loss of a proline at position +1 disrupts a key motif for the CDK1 kinase, resulting in the loss of phosphorylation at S386 [131]. (c) The substitution of a lysine to threonine introduces a novel phosphorylation site for the AKT1 kinase [137].

The described databases contain pre-computed predictions on a static set of variants. To facilitate flexibility in predicting any mutations, user-friendly tools were then developed, the first of which was the machine learning approach MIMP [146]. Briefly, MIMP employs PWMs to model binding specificities for 124 kinases. Peptides are scored before and after a mutation and a Bayesian approach is used to identify variants impacting the kinase site [146]. ReKINect utilises the NetPhorest [125] and NetworKIN [154] algorithms, which model

kinase specificity using neural networks. The magnitude of loss or gain of phosphorylation is then assessed through the predicted probabilities of kinase-binding to the wildtype and mutant peptides [155]. PhosphoPICK-SNP is another tool that allows for prediction of phosphosite-altering variants [156] by employing PhosphoPICK, a method which utilises HMMs to model kinase-specificity. Unlike MIMP and ReKINect, PhosphoPICK incorporates additional contextual information such as protein-protein interaction (PPI) networks and protein abundance allowing to allow for more accurate kinase-substrate and variant impact prediction. The described approaches have associated web servers and/or software libraries that further facilitate the interpretation of protein-coding variants.

### 1.3.3 Short linear motifs

Short linear motifs (SLiMs) are short, conserved, stretches of sequences (typically 3-11 amino acids long) following a particular pattern that is able to mediate transient protein interactions [157]. SLiMs are linear meaning that they are bound without the need of three-dimensional structure to bring residues in close proximity to one another. The linear pattern is often represented as a regular expression to reflect the positions and of residues critical for binding. SLiMs typically occur in intrinsically disordered regions, which lack structure and allow for the plasticity often required for binding [157]. It is characteristic that within a SLiM, only a limited number of residues are responsible for mediating the interaction. SLiMs are typically bound at low affinities meaning interactions are transient in nature, allowing for the formation of highly rapid and dynamic interaction networks [157]. By recognising and binding SLiMs, protein domains are able to carry out a variety of cellular functions including, but not limited to, proteasomal degradation, apoptosis, ligand binding and PTMs [157]. Since sequence specificity underlying PTMs, which was previously discussed, can be thought of as a SLiM, I generally refer to SLiMs as those not involving PTM functions throughout this thesis.

Both experimental and predicted SLiMs identified in the literature have been manually curated and compiled into publicly accessible resources. These include the eukaryotic linear motif (ELM) database, the LMPID database [158], Minimotif Miner (MnM) [159], and Scansite [160]. The ELM resource is, by far, one of the largest and most commonly used. It contains a total of 2,972 experimentally validated instances of SLiMs of 264 classes that mediate 1,429 interactions across 188 organisms [161]. Such resources provide comprehensive and systematic datasets on identified SLiMs.

Variants altering critical residues in a SLiM can interfere with the underlying interaction and thus function. There have been numerous examples linking mutations in SLiMs to diseases. For instance, the 14-3-3$\zeta$ protein (YWHAZ) is responsible for regulating the

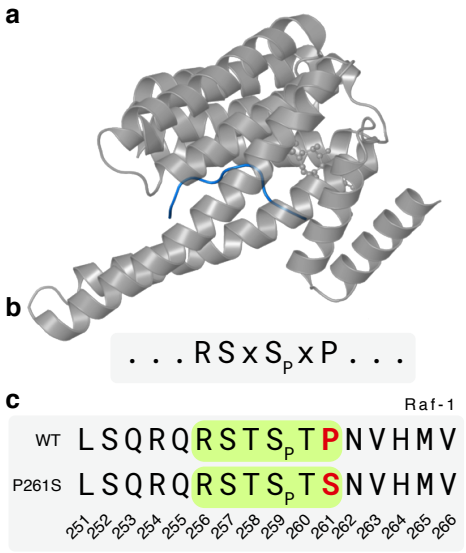



Fig. 1.4 An example of a SLiM affected by a disease variant. (a) 14-3-3$\zeta$ protein (grey) in complex with the RAF1 proto-oncogene peptide containing the SLiM (PDB:4IHL), represented as a regular expression in (b). (c) A variant (red) affecting a key residue in the SLiM (green) results in loss of binding by 14-3-3$\zeta$.

activity of proto-oncogene RAF1 through binding at its motif RSXS$_p$XP [162], where S$_p$ is a phosphorylated residue. The P261S substitution has been implicated in Noonan syndrome and abolishes the SLiM mediating the interaction, thereby deregulating its activity [163] (Figure 1.4). There have been cases where substitutions of non-critical residues in SLiMs are associated with disease. Although such positions are not directly involved in binding, they can contribute to the stability of the interaction, for example by forming hydrogen bonds with the interactors [164].

### 1.3.4 Protein Stability

The three-dimensional configuration of a protein is key to its function. Upon protein folding, a linear string of amino acids is converted into a functional three-dimensional structure that is maintained and stabilised by forces and interactions formed between atoms of the residues. The folds and interactions formed by a protein are driven primarily by its amino acid sequence [165]. Throughout this process, the protein traverses an energy landscape whereby unstable secondary, tertiary and quaternary folds and interactions are formed until a stable and minimal energy structure is achieved (Figure 1.5). How proteins traverse this landscape and reach a stable folded state ever so rapidly while avoiding misfolding and aggregation is very much an open question in science. An improved understanding of mechanisms underlying the

stability of protein intermediates during this process will allow us to better grasp how protein misfolding drives proteopathic diseases and how disease variation can alter this landscape.

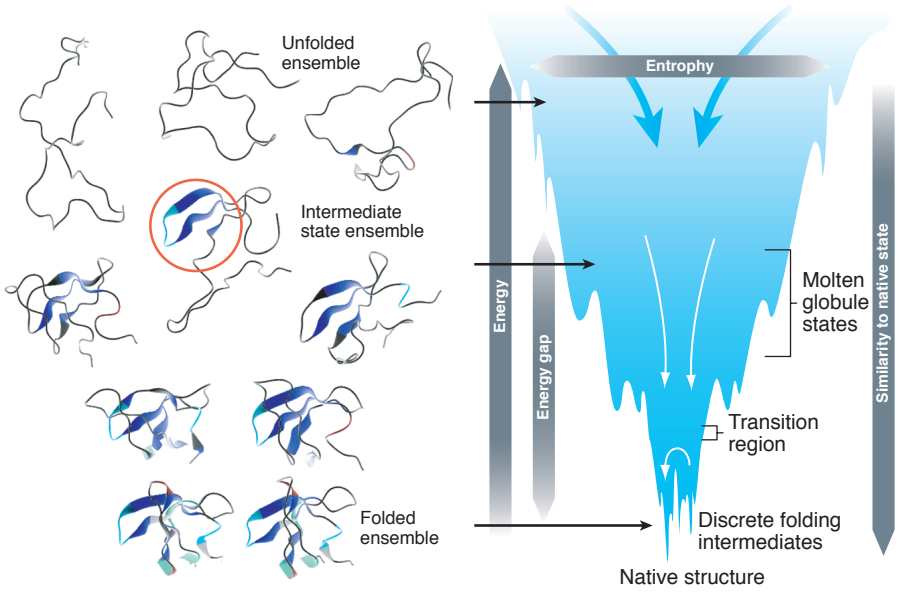



Fig. 1.5 A schematic of the energy landscape traversed by a protein during folding. Figure adapted from [166].

### Forces driving protein stability

There are several forces that govern the stability of the native conformational state of the protein. Commonly, these include ionic bonds, disulphide bridges, hydrogen bonds, hydrophobic interactions and VdW forces [167, 168] (Figure 1.6). Ionic bonding or salt bridges are forces that rely on the charge of the amino acid side chain. Naturally, the side chain of some amino acids are either fully protonated or fully deprotonated making them acidic (e.g. aspartic acid, glutamic acid) or basic (arginine, lysine). When oppositely charged side chains of residues are in close proximity (typically within 5Å) the bond formed is often highly favourable thereby contributing to the stability of a protein [169]. Hydrogen bonds are another force that largely contributes to protein stability. They form when a hydrogen atom covalently bonded to a highly electronegative atom (donor e.g. nitrogen or oxygen) interacts with another electronegative atom (acceptor) through electrostatic interactions. This will often occur when the hydrogen atom is less than 2.5Å away from the acceptor and the donor-hydrogen-acceptor angle is between 90-180°[170]. Hydrophobic interactions are thought to be one of the largest contributing forces to protein stability [168]. They help reduce the surface area of the protein to avoid unnecessary interactions with polar

solvents and occur when hydrophobic nonpolar residues in the protein (alanine, valine, leucine, isoleucine, phenylalanine, tryptophan and methionine) interact with one another in the presence of a polar solvent (typically water). VdW forces occur through electrostatic interactions between any two or more atoms in close proximity to one another (typically less than 4.5Å). They form between a temporary dipole (a molecule with uneven distribution of electrons, causing slightly positive and negative poles) and another atom that, upon contact, becomes dipolar (an induced dipole). Although VdW forces are much weaker in nature than that of chemical covalent bonds, an accumulation of VdW forces can result in significant stability contributions.

Overall, such forces and others, help provide the stability and allow conformational flexibility required to carry out protein function.

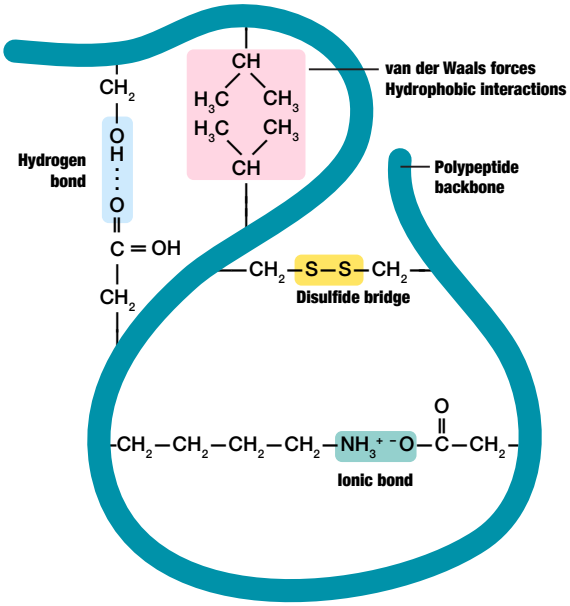



Fig. 1.6 Schematic representing common forces and interactions driving protein stability.

### The thermodynamic stability of a protein

The thermodynamic stability of a protein that is constantly between folded and unfolded states can be used to represent the overall stability of a protein (Figure 1.5). The stability of a protein can then be calculated as the difference in Gibbs free energy, $\Delta G$, between folded ($G_f$) and unfolded ($G_f$) states of the protein:

$$\Delta G = G_f - G_u \tag{1.1}$$

$$G = H - TS \tag{1.2}$$

where $G$ is the Gibbs free energy defined in terms of the enthalpy (H), temperature (T) and entropy (S) of the system. Typically, $\Delta G$ is measured in kcal/mol and a negative value indicates the protein is stable.

Experimental approaches to identify the stability of a protein typically involve first denaturing a native-state protein using heat or denaturing agents (e.g. urea, guanidinium chloride). The choice of denaturing agent often depends on the protein as some proteins are resistant to denaturing by certain agents. A readout for how denatured the protein is measured at different temperatures or denaturing agent concentrations. Examples of these readouts include ultraviolet (UV) light absorbency, fluorescence, or the catalytic activity of the protein. UV light absorbency and fluorescence utilise the presence of aromatic residues like tryptophan, tyrosine and phenylalanine, which are often partially buried in the core of the protein. As such, these residues absorb and fluoresce light differently when the protein is folded or unfolded [171], serving as an ideal indicator of the denatured level. Similarly, the catalytic activity of a protein often directly corresponds to the folding state, i.e. denatured proteins lose their catalytic activity [172]. A curve is then drawn where the readout measured is plotted as a function of the varying temperature or denaturant concentration (Figure 1.7a). This can then be used to calculate the rate at which the protein folds and unfolds, defined as $k_f$ and $k_u$, respectively. The ratio between these values is defined as the equilibrium constant $K_{eq}$ and is used to calculate the stability of the underlying protein, $\Delta G$. Given the gas constant ($R$) and the absolute temperature in kelvins ($T$), the stability is calculated as follows:

$$K_{eq} = \frac{k_u}{k_f} \tag{1.3}$$

$$\Delta G = -RT \ln K_{eq}. \tag{1.4}$$

**Genetic variants altering protein folding and stability**

Given the role of the forces formed between residues in the stability of a protein, mutations substituting these critical residues can lead to a decrease in stability, and therefore function. The impact a mutation has on protein stability is calculated by computing the difference between the free energy of the wildtype ($\Delta G_{wt}$) the mutant ($\Delta G_{mut}$), which is referred to as $\Delta\Delta G$:

$$\Delta\Delta G = \Delta G_{wt} - \Delta G_{mut}. \tag{1.5}$$

This value quantifies the magnitude of effect a mutation has on stability. A high value here (>1.8) indicates that the mutation is highly destabilising, whereas a low value (<-1.8) indicates the mutation introduces further stability [173]. Over the past several decades,

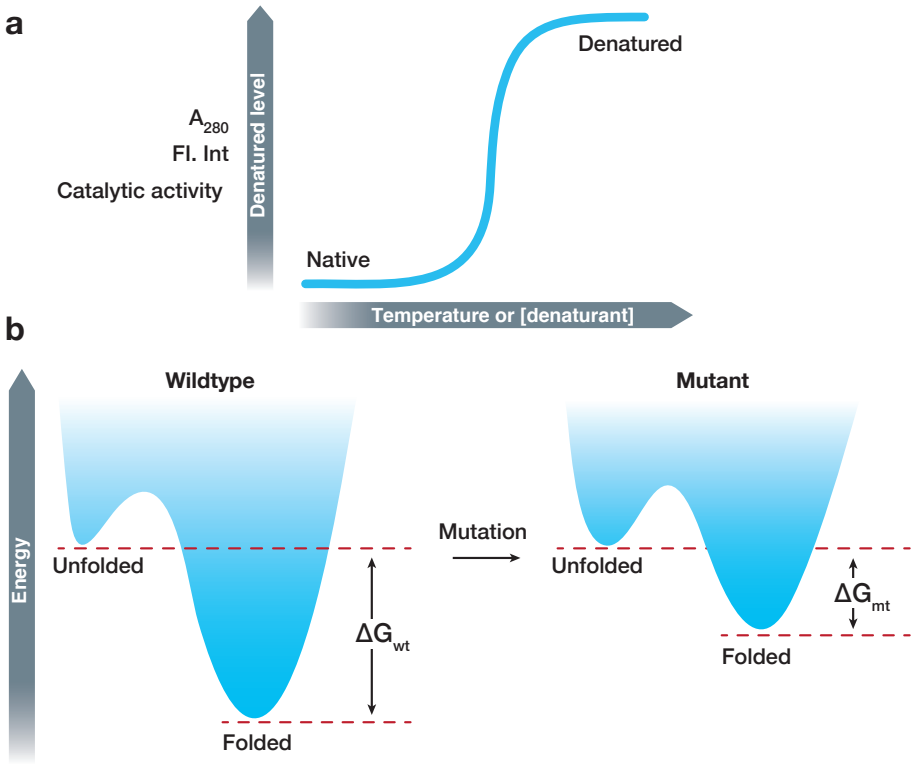



Fig. 1.7 The impact of mutations on protein stability. (a) To experimentally identify $\Delta G$ values, temperature or a denaturant are used to denature a protein while a readout, such as catalytic activity, measures the denatured level. The generated curve aids the calculation of $\Delta G$. (b) Energy landscapes of proteins before and after a destabilising mutation. The mutated protein has a low $\Delta G$ value compared to the wildtype and therefore, the computed $\Delta\Delta G$ is highly positive.

studies have carried out thousands of mutagenesis experiments to investigate the role of mutations in protein stability. The majority of identified destabilising mutations occur in the core of the protein compared to the surface. However, many surface mutations can still alter function by, for instance, affecting the active site of a protein or a PPI interface [174] (see section 1.3.5). The results of these experiments have been manually curated and collated in numerous online resources such as the ProTherm database [175], protein mutant database (PMD) [176], and the human genome mutation database (HGMD) [177]. ProTherm is one of the largest database and most comprehensive resource listing experimental $\Delta\Delta G$ values for 15,437 documented mutants in 311 protein structures [175].

It is no surprise that disease mutations have also been shown to alter protein stability. There has been a wide range of diseases associated with either the stabilization or destabilization of proteins such as prion [178, 179], muscular [180], retinal [181] and neurodegenerative [182, 183] diseases, to name a few. For instance, the M114T substitution in profilin 1 (PFN1)

is associated with amyotrophic lateral sclerosis through creating a cavity in the core of the protein, thereby destabilising it [180]. Conversely, the H101Q substitution in the CLIC2 protein, which is associated with X-linked intellectual disability, further stabilises the protein. Due to the added stability, the likelihood of conformational changes in CLIC2 is dramatically reduced, which impedes its transport to the cell membrane [184, 185]. These examples highlight the significance of disease-causing variants on protein stability.

### Computational methods for predicting the impact of genetic variants on protein stability

To facilitate identification of both stabilising and destabilising variants, computational approaches have been developed that predict the impact of a variant on protein stability ($\Delta\Delta G$). Stability predictors can be classified into two primary classes: sequence-based predictors and structure-based predictors. The performance of these methods is typically measured with respect to experimentally identified $\Delta\Delta G$ values obtained from the ProTherm database [175]. A quantitative value describing the performance the stability of the underlying protein as the correlation coefficient between experimental $\Delta\Delta G$ values and those predicted by the method.

Sequence-based predictors such as MuStab [186], EASE-MM [187], and iPTREE-STAB [188] utilise machine learning methods such as decision trees [188] or SVMs [186, 187] to predict $\Delta\Delta G$ values. Sequence features include biochemical features of amino acids, such as the hydrophobicity, polarity, molecular weight, and acidity ($pK_a$), as well as sequence-based structural features such as the tenancies for an amino acid to form different secondary structures and proportions of a residue estimated to be buried within the core [186]. Other biological features have also been used such as the number of codons per amino acid, average residue flexibility, refractivity and the relative mutability of an amino acid [186].

Alternatively, structure-based approaches such as FoldX [189], MAESTRO [190], Eris [191], SDM [192] and PoPMuSiC [193] exploit three-dimensional features available in protein structures to predict $\Delta\Delta G$ values. These approaches often utilise a set of physics-based energy functions to estimate the overall free energy of folding, $\Delta G$, of the wildtype and mutant. The contribution of stabilizing forces, such as electrostatic interactions, covalent bonding, and VdW forces, to the total energy, is identified and combined in a manner that accurately reflects the free energy of folding [194].

Sequence-based predictors benefit from higher coverage since they do not depend on the availability of protein structures yet often suffer from lower prediction accuracy. The bottleneck of structural data availability in structure-based approaches is often mitigated by hybrid approaches that incorporate sequence-based information [195]. Furthermore, recent studies have shown that even predictions made with medium to low-quality homology models

based on structures from closely related organisms show virtually no loss in prediction power [196].

In summary, protein stability is a mechanism that plays a major role in maintaining regular protein function. Mutations altering the biochemical properties of amino acids can disrupt or create stabilising interactions, and potentially alter protein function. Stability is also commonly altered in a wide variety of diseases. Understanding how mutations alter protein stability is, thus, of considerable interest for genetically engineering proteins for industrial, environmental and pharmaceutical applications, as well as better understanding the role of mutations in disease. The developed methods for predicting the impact of variants on protein stability is key for variant interpretation. However, much work is to be done to improve the accuracy of these algorithms, which will partly be aided by increased availability of structural data.

### 1.3.5 Protein-protein interaction interfaces

Rather than acting individually, proteins typically form macromolecular structures that carry out biological function through cooperative interactions with one another. With the majority of human proteins involved in at least one complex [197], PPIs are at the centre of a large variety of biological processes. Much like protein stability, these interactions are stabilised by forces within the interaction interfaces of both proteins such as hydrogen bonds, salt bridges, hydrophobic effects and VdW forces [198]. Residues within an interface that are crucial for binding are often referred to as 'hot-spots' and contribute to the Gibbs free energy of binding ($\Delta G_{bind}$). Understanding of the molecular determinants underpinning the stability of PPIs will help us better understand how these interactions form and how genetic variation can alter them.

#### Properties of protein interaction interfaces

PPIs are often mediated by structured globular domains that interface with the partner. The complex formed as a result of the interaction between two different proteins is referred to as *heterodimers*, whereas a self-interacting complex is a *homodimer*. The duration of the interaction can also vary, ranging from short-lived interactions (e.g. those mediated by SLiMs, see section 1.3.3) to more permanent ones (e.g. the ribosomal machinery).

Interaction interfaces often exhibit characteristic features that distinguish them from the rest of the protein. For instance, the amino acid composition at interaction interfaces tends to be more hydrophobic relative to other surface residues and less hydrophobic relative to the core of the protein [199–201]. The geometric shape of interfaces also plays an important role

in ensuring a complementarity between both proteins [202]. Computational predictors often leverage this information to predict interface residues [203, 204], however, the increasing availability of protein structures is simplifying the identification of *bona fide* interfaces. The protein data bank (PDB) currently holds 123,388 protein structures, 6,363 (5%) of which represent protein complexes of one or more PPIs. Resources such as Interactome3D [205] and eppic [206] have readily made available the structures of binary PPIs from multimeric protein complexes in PDB. Identifying interface residues in these structures is typically carried out by computing the relative surface accessibility (RSA) of residues [199, 200, 207], a measure that reflects how exposed or buried a residue is relative to its maximum solvent accessible surface area. The RSA is computed for residues in both monomeric structures and the interaction complex then subtracted to identify residues that are exposed in the monomeric structure but buried in the complex (Figure 1.8a). Although this is effective at identifying interface residues, not all residues will contribute equally to binding stability [208], and additional measures must be taken to identify those that do.

### Experimental and computational approaches to identify the contribution of interface residues to binding

Many experimental techniques aim to identify the precise contribution of interface residues to binding. For instance, alanine-scanning is a process by which surface residues are systemically mutated to alanines. The binding affinity can then be assayed for each mutant to assess the impact of the variant [209]. However, such approaches are often low-throughput and time-consuming. Instead, techniques such as deep mutational scanning provide a more high-throughput approach at assaying the contribution of interface residues to binding [210–212]. While this is approach is exhaustive and highly effective at identifying residues implicated in binding, routine application remains challenging and time-consuming.

Alternatively, computational approaches, similar in nature to those used for protein stability, have been developed that predict the impact of mutations on the free energy of binding, $\Delta\Delta G_{bind}$ (Figure 1.8b). The commonly used algorithms include FoldX [189], BindProfX [213, 214] and MutaBind [215]. MutaBind and FoldX use empirical energy functions to estimate $\Delta\Delta G_{bind}$. BindProfX uses information on conserved residues within homologous binding domains combined with $\Delta\Delta G_{bind}$ predictions from FoldX to improve accuracy. These predictors are often benchmarked against experimentally derived $\Delta\Delta G_{bind}$ values, which are deposited in resources like the SKEMPI database, which holds $\Delta\Delta G_{bind}$ values for over 3,000 mutations across 85 PPIs [216] and more recently the PROXiMATE database, with over 6,200 mutations across 175 complexes [217].

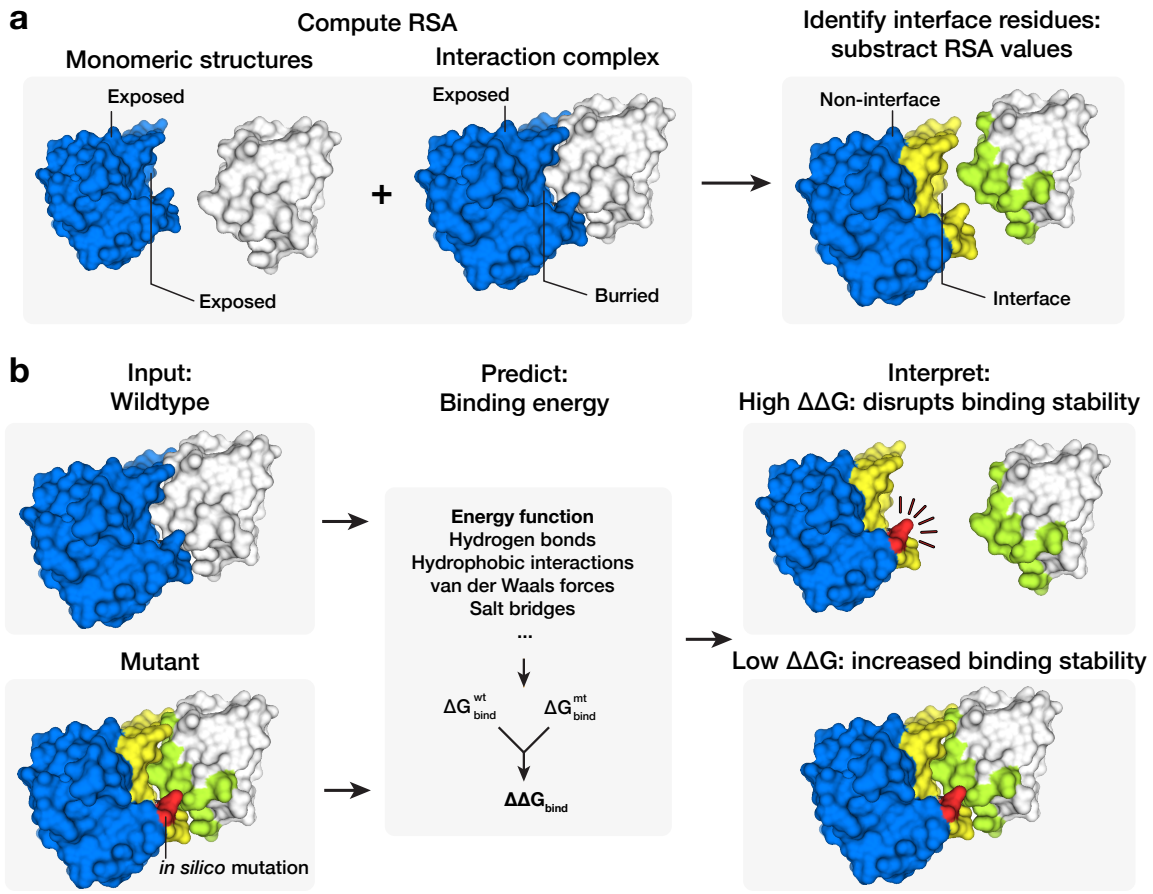



Fig. 1.8 (a) To identify interface residues the RSA of residues in monomeric structures is subtracted from that of the interaction complex. This identifies residues that are exposed in the monomeric protein but buried within the complex. (b) Predicting the impact of a mutation on a PPI involves predicting the binding energy, $\Delta G_{bind}$, of both wildtype and mutated interaction complex. The difference $\Delta\Delta G_{bind}$ between those two values typically indicates whether the mutation destabilises, stabilises or has no impact on binding.

There still exists much room for improvement, with respect to the performance of predictors, partially due to the relatively limited number of experimentally derived $\Delta\Delta G_{bind}$ values and PPI structural data. These predictors do, however, provide a quick and inexpensive alternative to experimental approaches and have been used to uncover valuable insights into developing drug PPI inhibitors [218], PPI-mediated viral infections [219, 220], and engineering of novel PPIs [221].

### Disease variants in interaction interfaces

PPIs are commonly modified in disease. Disease variants have been known to occur in interface residues, resulting in altered binding and leading to disease phenotypes. For

instance, the substitutions K228E and N374H in T-box, brain, 1 (TBR1) are associated with sporadic autism and have been shown to abolish both homodimerization of TBR1 and interaction with the TF FOXP2, which has been previously implicated in speech and language disorders [222]. Another example is that of the proliferating cell nuclear antigen (PCNA), which has been associated with multiple diseases including DNA repair diseases such as xeroderma pigmentosum, Cockayne syndrome, and ataxia telangiectasia [223]. The substitution S228I was shown to cause a large deformation of a binding pocket, abolishing binding with multiple partners [224].

More systematic studies have been carried out that demonstrate, on a large-scale, the role of disease variation in PPIs. For instance, a recent study analysed mutation data from 5,989 tumour samples across 23 different cancer types in the context of interfaces in over 10,000 proteins [225]. They identified over 100 PPI interfaces significantly enriched in somatic cancer mutations, one-third of which were interfaces in known cancer driver genes, which were also shown to serve as central hubs in PPI networks. The remaining proteins showed literature implicating them in cancer. These findings were additionally coupled with clinical data such as survival curves to highlight the implication of PPI interface mutations. More structure-based studies have utilised predictions of $\Delta\Delta G$ of binding to assess the role of disease variants in PPIs. For example, using FoldX, Billur Engin et al. analysed over 1.2 million somatic cancer mutations in the context of over 4,800 experimentally derived interaction interfaces and found that over 20% of mutations on the surface of a protein were predicted to affect binding affinity [226]. These mutations commonly fall within interaction interfaces of known tumour suppressors and oncogenes to alter binding.

Understanding the molecular underpinnings of PPI specificity is key to understanding the impact of mutations on protein interfaces. Advances in both physics and statistical-based algorithms have allowed for computational tools, such as FoldX, to estimate the impact upon which a mutation will have on the binding. While such approaches provide a convenient way by which mutational impact can be assessed, their performance hampered by the scarcity of experiential $\Delta\Delta G_{bind}$ values and structural data. Increasing availability of such data, along with advances in NGS and proteomics, will propel a complete understanding of the molecular basis upon which mutations alter PPIs.

### 1.3.6 Initiation and termination of translation

Translation is an intricate system by which ribosomes convert transcribed mRNA into a folded, functional protein (Figure 1.9a). Initiation and termination of translation are guided by start and stop codons, respectively. Start codons indicate the start of translation and in eukaryotes it is almost always the AUG codon, which encodes for a methionine. Unlike

eukaryotes, prokaryotes have a selection of primarily three start codons: AUG (73%), GUG (14%) and UUG (3%) [227]. Stop codons, encoded by UAG, UAA, and UGA, indicate a signal for the termination of translation (Figure 1.9b). This occurs because transfer RNAs (tRNAs) do not harbour any anticodons that are complementary to stop codons, thus forcing the release from the polypeptide. Mutations disrupting start or stop codons can have drastic effects that hamper translation and result in mRNA degradation, protein truncation or misfolding. Accounting for start and stop codons is therefore insightful to understanding altered mechanisms and their role in disease.

### 1.3.7 Mechanisms for quality control of variants altering start and stop codons

Mutations can have one of three effects on start and stop codons: disrupt a stop codon, introduce a premature stop codon or disrupt a start codon.

A point mutation disrupting a stop codon, also known as a nonstop mutation, and results in the stalling of the ribosome upon reaching the 3' end of the mRNA. This triggers a process termed nonstop-mediated decay (NSD). In *S. cerevisiae*, this process involves the recruitment of the Ski complex, containing cofactors SKI2, SKI3, SKI8, and SKI7, as well as the exosome complex. Together, the SKI and exosome complexes are responsible for degrading the mRNA in a 3' to 5' fashion, while also repressing further translation of the transcript [228] (Figure 1.9b). *S. cerevisiae* E3 ubiquitin ligases LNT1 and NOT4 then promote polyubiquitination-mediated degradation of the produced polypeptide [229, 230]. Genes involved in the SKI and exosome complexes are highly conserved across eukaryotes [231] and their homologs also play key roles in NSD [232].

Nonsense mutations introduce premature stop codons resulting in aberrant mRNA, which, if translated, will produce shortened polypeptides that will likely lead to deleterious effects on protein function [233, 234] (Figure 1.9c). The severity of nonsense mutations primarily depends on their location in the protein, where variants closer to the N-terminal are more likely to have a deleterious impact. Processes in place such as nonsense-mediated decay (NMD) aim at reducing this likelihood by purging mRNA transcripts carrying nonsense mutations (Figure 1.9c). In *S. cerevisiae*, when the ribosome reaches the termination codon it is often unable to release an incomplete polypeptide [235]. Termination factors SUP34 and SUP45 are recruited, which in turn recruit three additional factors UPF1, UPF2, and UPF3. This promotes the disassociation of the ribosome, GTP hydrolysis of the generated peptide and recruitment of the DCP1-DCP2 decapping enzyme complex to the 5' of the mRNA [235]. DCP1-DCP2 triggers hydrolysis of the 5' cap, which quickly results in degradation

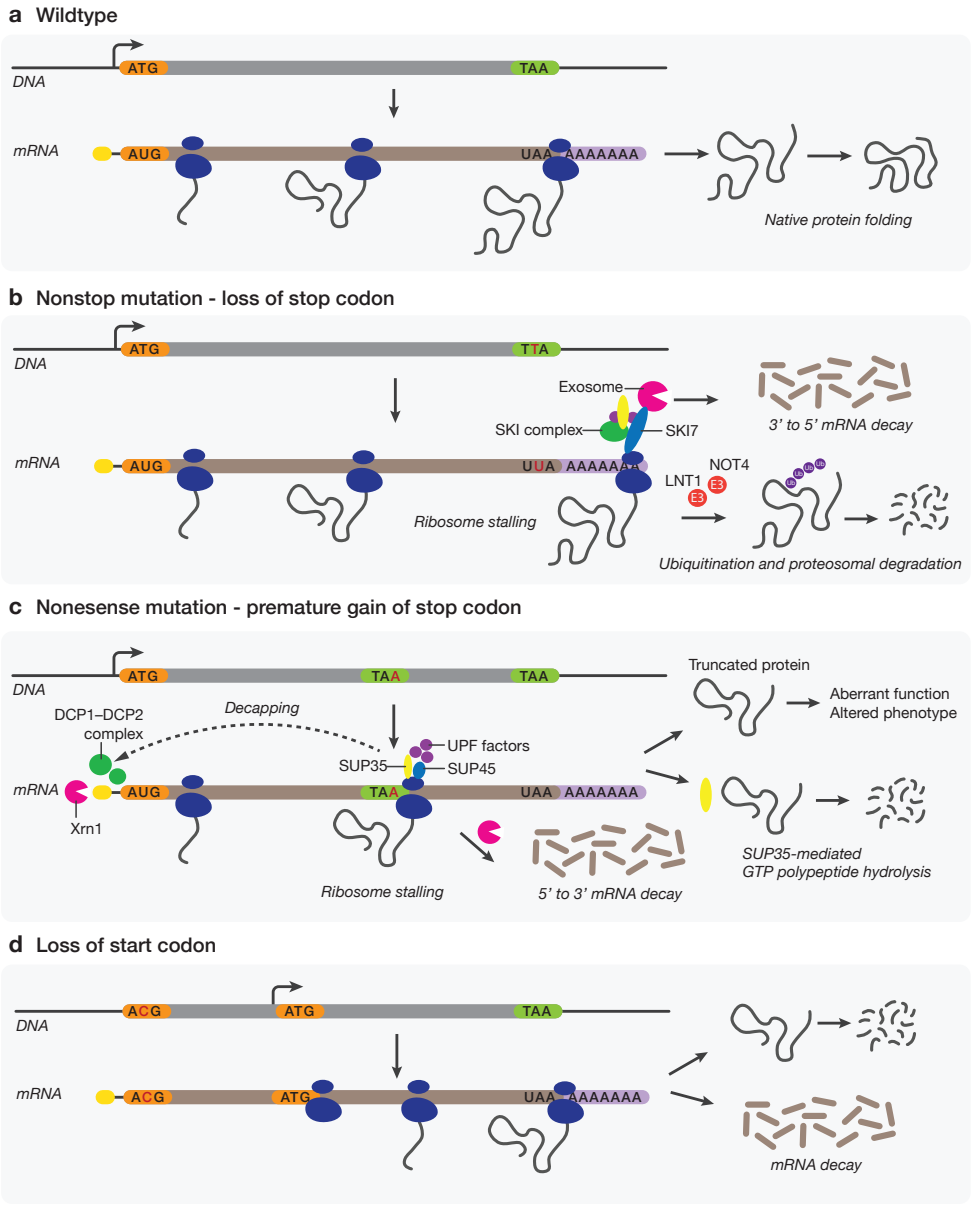



Fig. 1.9 (a) Translation involves the ribosome machinery (blue) sliding across generated mRNA transcribed from DNA and producing a polypeptide which folds into a fully functional protein structure (b-d) mRNA surveillance pathways to deal with aberrant transcription or translation. These include (b) the loss of a stop codon (c) premature gain of a stop codon and (d) loss of a start codon. Such effects either result in a non-functional misfolded or truncated protein or the triggering of respective pathways to degrade generated polypeptides and mRNAs.

of mRNA in a 5' to 3' fashion by the exonuclease XRN1 [236]. Such fail-safe mechanisms reduce misfolding of protein and aberrant functions.

Finally, mutations disrupting the AUG start codon in eukaryotes significantly hamper translation efficiency of mRNA. In many cases, either the mRNA or generated protein is degraded (Figure 1.9d). A recent study in *S. cerevisiae* shows that the degree of efficiency of non-AUG start codons depends on the sequence context. Specifically, if there is another AUG codon upstream of the mutated start codon, translation initiation can be rescued.

#### Disease mutations in start and stop codons

Disease-causing mutations that alter start and stop codons have been well documented in the literature. For instance, the nonsense mutation G542X in the *CTFR* gene results in significant loss-of-function and has been implicated in cystic fibrosis [237]. Another example is the nonstop mutation X420Y in the microphthalmia-associated TF (*MITF*), which is associated with Waardenburg syndrome, a genetic disorder characterised by deafness and pigmentation changes [238]. The mutation causes a 33 residue extension of the protein product, resulting in the reduced transcriptional activity of *MTIF* [238]. A final example is the M1I substitution identified in encephalomyelitis (inflammation of the brain and spinal cord) patients, causing the loss of the start codon in *CMT1X*. This results in abolished gene expression, even though the mutation had no effect on mRNA levels [239]. These examples highlight how disease mutation can alter initiation and termination of translation. Disease mutations altering start and stop codons have also been the subject of a number of therapies that aim to suppress their effects [240, 241], further elucidating their importance.

## 1.4 Sequence conservation of proteins

Although not a biological mechanism *per se*, sequence conservation plays a crucial role in identifying key functional regions in proteins. Regions of a protein that remain unchanged over time despite mutational pressure are often due to natural selection and suggest that any variation would be deleterious to the organism's fitness [242]. Indeed, conserved elements include protein domains, which often mediate key proteins function. Some of the most conserved proteins include essential cellular machinery such as the 16S and 23S ribosomal proteins, as well as binding domains of ATP-binding cassette transporters and yet these sequences, remain virtually indistinguishable across organisms that separated by billions of years of evolution [243]. Therefore, assessing similarities between protein sequences across organisms offers significant insight into functionally relevant protein regions.

Computational methods have leveraged sequence conservation to predict whether an SNV would impact protein function [244–246]. These commonly include SIFT [244], GERP++ [245], and PolyPhen2 [246], to name a few. These algorithms in one form or another utilise

sequence conservation similarly. In brief, for a protein of interest a set of related homologous protein sequences are identified, typically using programs like PSI-BLAST [247] and used to construct a multiple sequence alignment (MSA). Given a nonsynonymous SNV, the algorithm then uses a scoring metric to determine how frequently the mutated residue is observed at that position within the MSA. Algorithms also often account for the chemistry of amino acids when identifying how conserved a position is, rather than rely on sequence identity alone. For instance, a position that exhibits an abundance of glycines, alanines, valines and leucines does not necessarily lack conservation and could instead indicate the position is tolerant to amino acids with similar chemical properties, in this case, a hydrophobic side chain.

Algorithms utilising sequence conservation have become a fundamental component of variant analysis and prioritization pipelines. They have helped identify many causal variants across human disease [248, 249] and other organisms [250, 251]. Although they do not directly provide insight into affected mechanisms, they can often instead be combined with mechanistic predictors to allow for an increased confidence of variant impact prediction.

## 1.5 Aims of the thesis

Not all genetic variation will influence the phenotype of an organism. I have described here the approaches taken to identify causal genetic variants, such as GWAS and QTL-based association methods. Furthermore, I discuss some of the biological mechanisms which these variants can impact, how they can be computationally modelled and used to provide mechanistic insight into variant effects. This thesis is primarily focused on the latter.

Having described the importance of kinase-substrate sequence specificity in the context of disease variants, Chapter 2 describes a computational approach I have developed a method which utilizes functional interaction data and phosphorylation data to predict specificities of kinases. This method was applied to human kinases and was able to predict substrate sequence preferences for over half of all known kinases, capturing known well-characterised kinase specificities, as well as novel ones. Several specificities were additionally validated experimentally and were shown to closely resemble predicted ones.

In Chapter 3, I compile a comprehensive set of ASB variants from numerous studies and use them as a gold-standard to assess how TFBS predictors perform at assessing the impact of variants on TF binding. Results suggest that TF specificity models exhibit variable levels performance at predicting the impact of variants, that is independent of the performance of the TF specificity model at predicting TFBSs. This suggests the mechanism underlying TF binding for poorly performing TFs is one that is far more complex and robust than expected. I further compare the performance of different scoring schemes across different methods

(both PWMs and more sophisticated machine learning approaches) used to assess the impact of TF binding. For TFs that are unable to accurately predict the impact of variants, I explore alternative mechanisms such as methylation, DNA shape and binding co-factors that may explain the poor performance.

Chapter 4 describes a comprehensive effort to compile and benchmark state-of-the-art sequence and structure-based predictors of mutational consequences and predict the effect of all possible amino acid and nucleotide variants in the reference genomes of *H. sapiens*, *S. cerevisiae* and *E. coli*. Predicted mechanisms include protein stability, interaction interfaces, PTMs and TFBSs. These variant effects are provided through mutfunc, a fast and intuitive web tool by which users can query pre-computed predictions by providing amino acid or nucleotide-level variants. I apply computed predictions to analyse known causal disease variants as well as provide mechanistic hypotheses for causal variants of unknown function.

Chapter 5 involves the analysis of natural variants harboured by wild *S. cerevisiae* isolates in the context of predictions generated in Chapter 4. The mechanistic impact predictions are used to generate gene-level functionality scores, which are then used to perform gene-phenotype associations. This provides a number of advantages over traditional GWASs, namely, by identifying dysfunctional genes through analysis of rare genetic variants, it alleviates the requirement of a large number of samples required to identify associations in GWASs and uncovers many novel associations.

# Chapter 2

# Uncovering phosphorylation-based specificities through functional interaction networks

*In this chapter, we describe a novel computational method, which utilizes functional interaction data and phosphorylation data to predict sequence specificities of kinases. I conceived the method and carried out all computational analysis under the supervision of Pedro Beltrao. I was not involved in the generation of experimental data. Mass-spectrometry experiments were carried out by collaborators Naoyuki Sugiyama and Yasushi Ishihama at the University of Kyoto. The work in this chapter includes published material from the following article:*

Omar Wagih, Naoyuki Sugiyama, Yasushi Ishihama and Pedro Beltrao (2016). Uncovering phosphorylation-based specificities through functional interaction networks. *Molecular & Cellular Proteomics*, *15*(1), 236-245

## 2.1 Introduction

Over the past decade, there has been an ever-increasing quantity of phosphorylation site (phosphosite) data, typically identified using mass spectrometry (MS), with over 100,000 phosphosites identified in human [252–254]. Despite this, a large number of these phosphosites are without a known upstream regulatory kinase. Compendiums of kinase-substrate relationships curated from the literature [252–254] currently associate only 6% (6,320/107,444) of known phosphosites to one or more kinases. While the experimental characterization of

kinase target phosphosites allows for the discovery of many kinase specificities, they are typically expensive, time-consuming, and are not feasible for kinases that are difficult to work with. The ability to uncover specificities driving kinase-substrate phosphorylation will allow us to better model cellular signalling networks and understand how genetic variation can rewire such networks.

The only available computational method aimed at predicting kinase-substrate specificity without any prior knowledge of its substrates is Predikin [255], which does so by employing substrate-determining residues (SDRs) in the catalytic domain of the kinase. These are residues that have been identified as driving the kinase's specificity towards particular positions in the substrate phosphosite. In Predikin, SDRs are identified by aligning sequences of kinase catalytic domains using HMMER [256] to identify semi-conserved elements. Residues surrounding these elements are then mapped to three-dimensional structures of kinases in complex with their substrate peptide and manually curated for interactions with positions on the peptide (Figure 2.1a). To predict specificity of a kinase given its sequence, Predikin first analyses the sequence for a kinase catalytic domain. If identified, the domain sequence is scanned for previously reported SDRs (Figure 2.1a). Substrate phosphosites of kinases with similar SDRs are then combined and used to construct a predicted specificity model (Figure 2.1b). In this way, Predikin makes use of known kinase target sites. While effective, Predikin exhibits several shortcomings. First, it depends on the availability of protein structures and therefore cannot be easily scaled to kinase families without three-dimensional structures nor to other post-translational modification (PTM) recognition domains. Second, it depends on the availability of kinase-substrate data to construct final specificity models, further limiting scalability. Lastly, SDRs for tyrosine kinases have not been identified, and therefore predictions can only be made for serine/threonine kinases.

We decided to take an alternative approach to predicting kinase specificity. Previous studies have shown that it is possible to use information regarding the interaction partners of a peptide-binding protein to identify potential motifs mediating these interactions [257]. As such, putative interaction partners of a kinase should be more likely than random proteins to be phosphorylated by that kinase. Phosphosites occurring on interaction partners of kinases should, therefore, confer a bias in amino acid composition toward the kinase's specificity, which could be revealed by motif enrichment (Figure 2.1c). We applied our method using the STRING functional interaction network [258] and publicly available phosphoproteomic data to predict specificities for 57% (282) of all human kinases. These included kinases with previously known specificities, as well as other understudied kinases. To validate the proposed method, we experimentally determined kinase-substrate phosphosites for four understudied kinases. Predicted models were shown to closely resemble the specificity of

experimentally identified phosphosites. The analysis was extended to show that specificities can be predicted, not only for kinases but also for other phospho-residue binding domains, such as 14-3-3 proteins and the acetyl-lysine binding bromodomain. Finally, this method was applied to mouse kinases to explore conservation of sequence specificity.

We demonstrate that it is possible to combine large-scale PTM data with protein network information to derive the specificity of PTM regulators and believe that this approach can be widely applicable to different PTM types.

## 2.2 Results

### 2.2.1 Network-based prediction of kinase-substrate specificity

We hypothesized that the interaction network of a protein kinase should be enriched in its target proteins. This hypothesis was confirmed by the observation of a significant enrichment of known kinase targets in the functional interaction or physical interaction partners of kinases. Of phosphosites with an annotated kinase, 5.5% show a functional interaction with the kinase as revealed by STRING, which was significantly larger than that of random pairs of proteins (Figure 2.2, $p<2.22\times10^{-16}$). Similarly, 7.61% show a physical interaction in BioGRID [261], which was also significantly higher than that of random kinase-substrate pairs (Figure 2.2, $p<2.22\times10^{-16}$). In order to predict kinase specificities, information on human protein interaction data and phosphorylation data derived from large-scale MS studies were combined. Functional interactions derived from STRING [262] were used as a source for potential kinase interactors. STRING reports a score ranging from 0 to 1000 reflecting the confidence of the interaction, which is a combination of scores from across multiple evidence sources [263]. A total of 2,425,314 interactions in 22,523 proteins were collected along with experimentally determined phosphosites from three public databases (PhosphoSitePlus [117], PhosphoELM [253] and HPRD [264]). Phosphosites were mapped back to proteins with STRING interactions resulting in 107,444 sites across 12,207 proteins (Methods, section 2.3.1). A total of 493 kinases were identified within this reference proteome (81% serine/threonine and 19% tyrosine kinases) using the Kinomer prediction tool [265] (Methods, section 2.3.2). For a given kinase, enrichment was carried out on all phosphosites occurring on the STRING partners using the motif-x algorithm [266] (Methods, section 2.3.3, Figure 2.1c). A random sample of 10,000 unphosphorylated S/T/Y residues were used as the background for enrichment. Phosphosites matching the most significant extracted motifs were then used to build a position weight matrix (PWM) highlighting the predicted specificity of the kinase, which could then be used to predict phosphosites (Figure 2.1c).

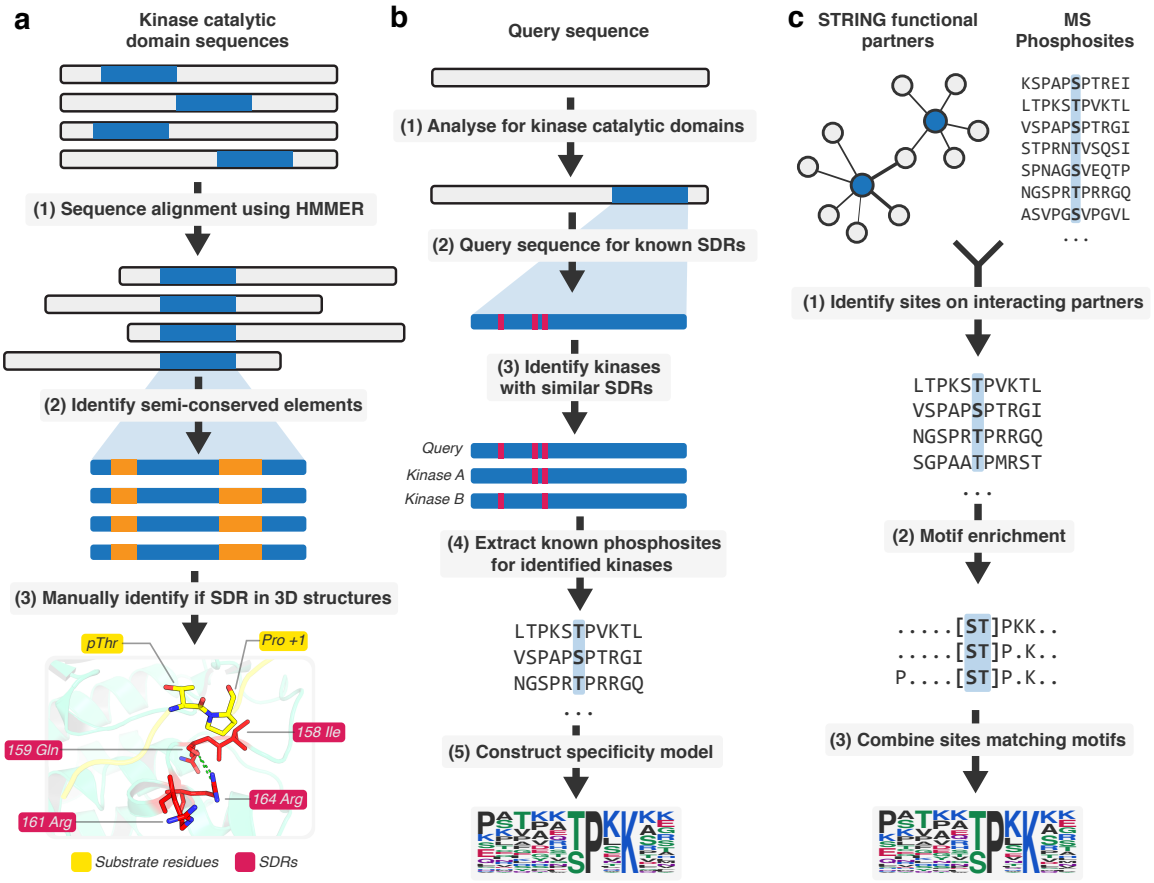



Fig. 2.1 (a) Identification of SDRs in Predikin is carried out by (1) aligning sequences of kinase catalytic domains to (2) identify semi-conserved elements. (3) Residues proximal to these elements are manually curated to identify if they interact with specific positions the peptide. The structure shows literature defined SDRs (magenta) interacting with proline at position +1 of the substrate (yellow) [259, 260]. (b) Predicting kinase specificity with SDRs using Predikin. (1) A kinase query sequence is first analyzed for a kinase catalytic domain. (2) If identified, this domain is annotated with previously identified SDRs. (3) Kinases with similar SDRs are then identified. (4) Their known target phosphosites are collected and (5) used to construct a predicted specificity model (c) Overview of the proposed method. (1) Experimentally identified phosphosites on functional interaction partners of a kinase are collected. (2) The phosphosites are then subject to motif enrichment to identify over-represented motifs in the flanking sequence, likely reflecting the kinase's specificity. (3) Phosphosites matching the top five significant motifs are then retained and used to construct a specificity model.

A survey of all known phosphosites revealed a strong bias of prolines at position +1 (P+1) (Figure 2.3a-b). This results in consistent enrichment of P+1 motifs (Figure 2.3c). The abundance of P+1 sites results in proline-directed motifs masking the true underlying motif of non-proline-directed kinases. To circumvent this, we require prior knowledge on whether

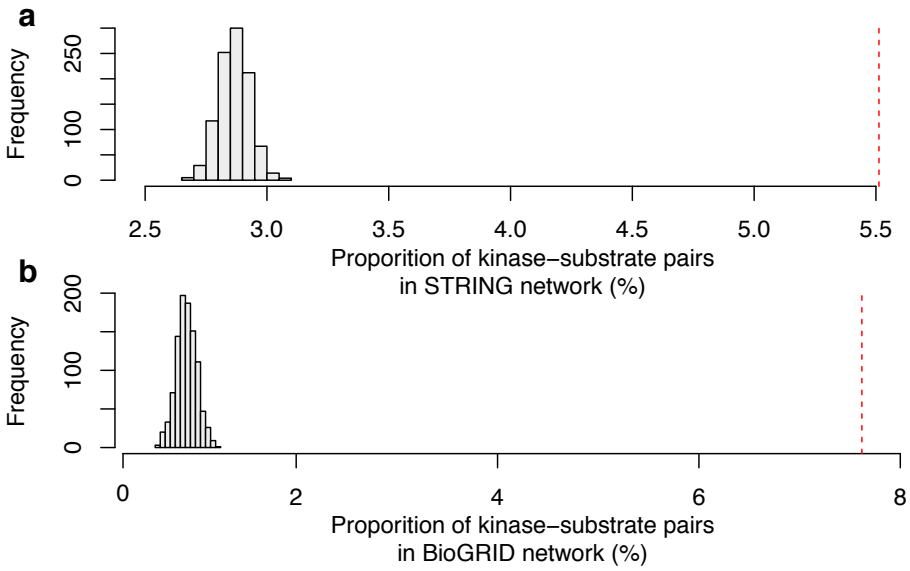



Fig. 2.2 Enrichment of kinase-substrate pairs in protein interaction networks. (a) In the STRING interaction network for the human kinases, 5.5% of the interactions correspond to known kinase-substrate interactions ($p<2.22\times10^{-16}$). (b) Similarly, BioGRID contains 7.61% known kinase-substrate interactions ($p<2.22\times10^{-16}$). The histograms show the proportion of 1,000 random kinase-substrate pairs in STRING and BioGRID.

a kinase is proline-directed (i.e. prefers P+1) is required. We found that in almost all cases, kinases belonging to the CMGC family, including CDKs, MAPKs, GSKs and CDK-like kinases harbour P+1 motifs, as shown in their experimental binding sites (Figure 2.3d). Thus, if a kinase was not predicted as belonging to the CMGC family, it was assumed not to be proline-directed and P+1 phosphosites were removed from the foreground and background sets prior to motif enrichment.

There are two variable parameters in the method (1) the threshold for the functional interaction prediction score from STRING and (2) the top $k$ number of significant motifs extracted during the enrichment. To determine the optimal thresholds, we assessed the performance of predicted kinase specificity models against a set of 9,595 gold standard kinase-substrate relationships. We carried out benchmarking using a set of nine well-studied kinases from a diverse set of kinase families (ABL1, AKT1, ATR, AURKB, CDK2, CSNK2A1, GSK3B, MAPK1, and PRKACA) with well-recognized specificities in the literature. The STRING score threshold and the top $k$ motifs extracted were varied and the performance of the resulting PWM in each case was evaluated using the Receiver Operating Characteristic (ROC) and area under the ROC curve (AUROC or AUC) (Methods, section 2.3.4). Increasing the STRING score threshold, on average, resulted in higher AUCs up to a value of 400, after which no significant increase in performance was observed (Figure 2.4). Furthermore, more

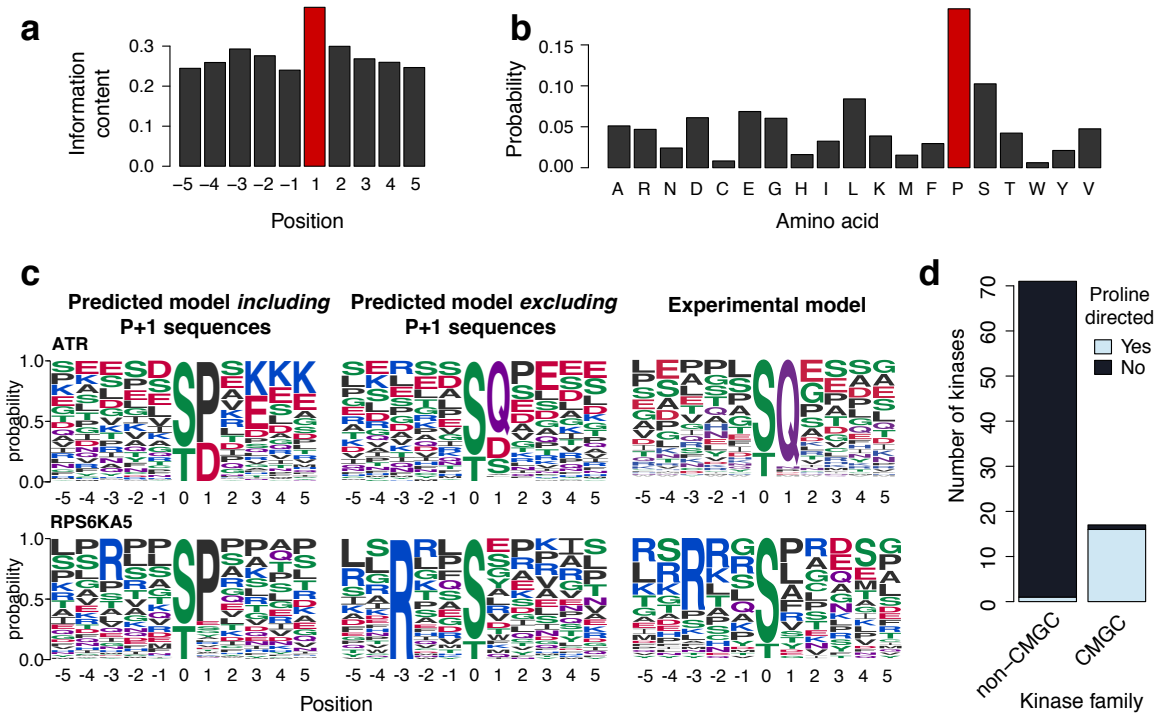



Fig. 2.3 Proline bias across all phosphosites. (a) Information content for each position flanking the central residue for all known phosphosites. The highest information content position is shown in red. (b) Amino acid frequencies for each position of all phosphosites. Proline is highlighted in red. (c) Two examples of non-proline-directed kinases showing the effect of removing P+1 phosphosites. Each example shows the predicted specificity before (left) and after (right) removal of P+1 phosphosites. Consistent enrichment of proline is observed for these cases if P+1 phosphosites are not removed, masking the true specificity of the kinase. (d) Bar plot showing CMGC vs. non-CMGC kinases with at least 20 substrates and the proportions of each class, which are proline-directed. Kinases were considered proline-directed if a predominance of Proline was observed at position +1.

stringent STRING thresholds result in an insufficient number of sites for motif enrichment thereby reducing the coverage of kinases for which we can predict specificity. Thus, a STRING score threshold of 400 is used throughout. Additionally, the top five motifs resulting from motif enrichment were selected for model construction for two reasons. First, varying the top $k$ motifs beyond a value of five did not show considerable improvement in performance. Second, over selecting motifs was shown to mask the true predicted specificity of the kinase (Figure 2.5).

The STRING score used here is based on a combination of multiple evidence sources (e.g. text mining, co-expression, interaction data). STRING also provides source-specific scores. To ensure certain evidences were not affecting performance, functional interactions were defined using the inclusion or exclusion of individual evidences. However, on average,

using alternate evidences did not provide a significant increase in performance, and resulted in fewer numbers of kinases with predicted specificity (Figure 2.6).

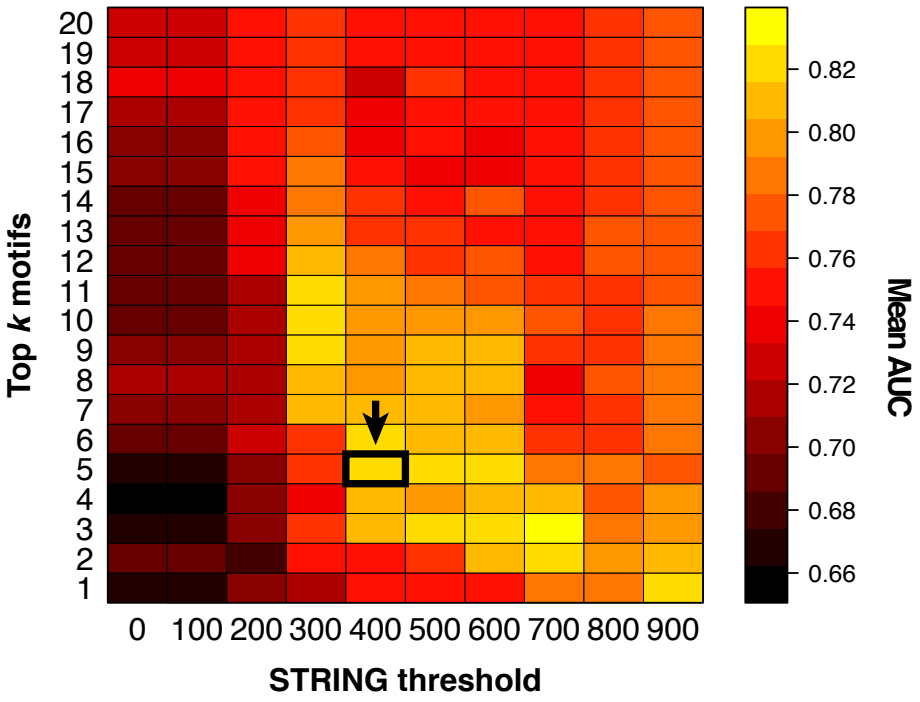



Fig. 2.4 Benchmarking of the method on nine kinases with well-defined specificities. Optimal parameters of the method were computed by varying the STRING score threshold and the top $k$ significant motifs used in constructing the model. The performance of kinases is measured in each case. The arrow corresponds to selected thresholds.

The collection of all phosphosites contain over-represented motifs. To ensure predicted models were not due to random motif enrichment, we compared the performance of predicted models against random models constructed without the STRING network information. If a given kinase has $n$ STRING interactions, and among those interactors, there are $m$ phosphosites $(s_1, s_2, ..., s_m)$, then $m$ random phosphosites are selected from all known phosphosites $(r_1, r_2, ..., r_m)$. Specificity is then predicted, as previously described using these sites. Random models were compared against the predicted models in their ability to discriminate the gold standard sequences, as measured by the AUC. Predicted models for most of the nine kinases, with the exception of AKT1 and MAPK1, performed significantly better than random (Figure 2.7a). This does not imply that the predicted AKT1 model is incorrect since it performs well at predicting known AKT1 target sites (AUC = 0.90, Figure 2.7a). However, some kinases such as AKT1 have specificities that are well modelled by the common motifs across all phosphosites. In such cases, the network information appears to provide almost no

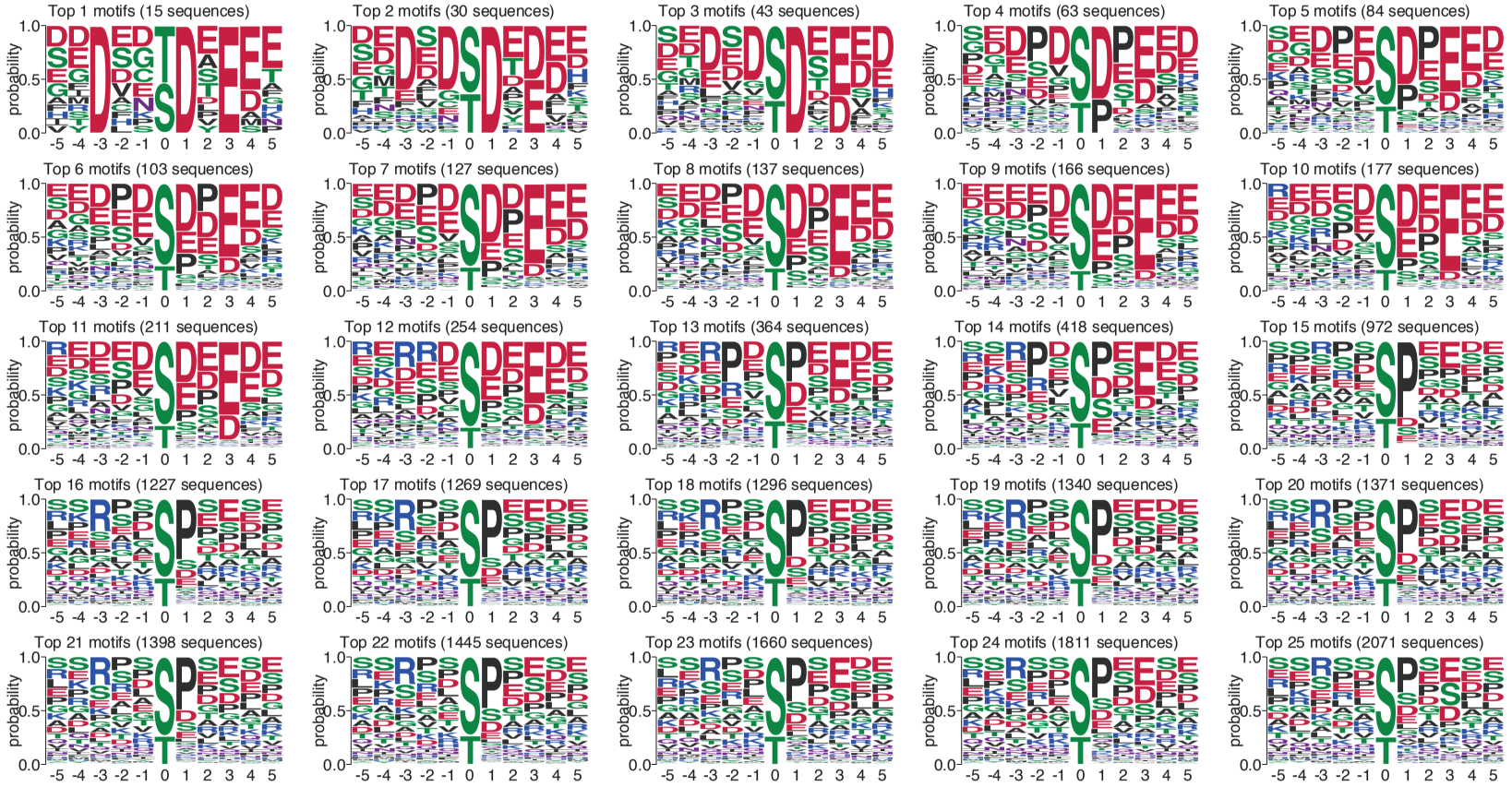



Fig. 2.5 Masking predicted specificity by over-selecting enriched motifs. Specificity predictions for CSNK2A1 (CK2) resulting from different top $k$ significant motifs. Over-selecting enriched motifs can result in less specific predictions and sometimes the inclusion of other contaminant motifs.

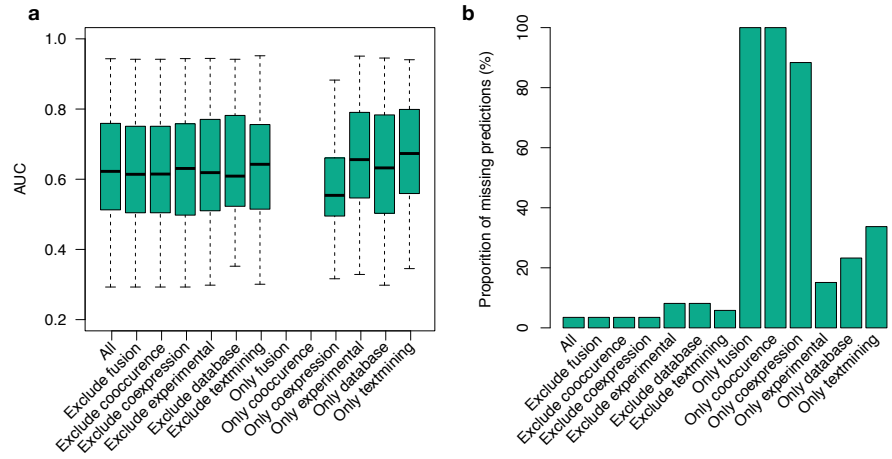



Fig. 2.6 Performance using inclusion or exclusion of different STRING evidences. (a) Distribution of AUCs for predicted models of all kinases with $\geq$20 known substrates by either excluding a particular string evidence or using only that evidence to generate the prediction. (b) The proportion of kinases with no prediction resulting from lack of interactions when restricting evidences.

gain compared to random sampling. In contrast, ATR has an atypical specificity with a Q+1 preference that is well recovered by this approach (Figure 2.7a,f) but highly unlikely to be observed in a random pool of phosphosites.

These results demonstrate the ability to integrate protein interaction information with large-scale data on protein phosphorylation to derive kinase specificity models.

### 2.2.2 Prediction of kinase-substrate specificity across all human kinases

Our method was applied to all human kinases, resulting in predictions for 282/493 (57%) of kinases. Kinases that did not yield a prediction either had a low number of partners or a scarcity of phosphosites on partners. Performance of predicted models for 85 kinases with at least 20 literature-defined phosphosites was measured by how well they performed at discriminating the literature-defined phosphosites from that of other kinases (Methods, section 2.3.4, Figure 2.8). The average AUC across all kinases was 0.64 with 32% (27/85) of kinases having an AUC greater than 0.7 (Figure 2.7b). On average, CMGC, PIKK, and AGC families performed best, whereas TKL, STE and TK kinases had a larger fraction of poorly performing models. When excluding the TKL, STE, and TK kinases, the average AUC increases to 0.68 with 44% of kinases (27/61) scoring higher than 0.7 (Figure 2.7b). Differences in performance across kinase families could reflect different degrees of sequence

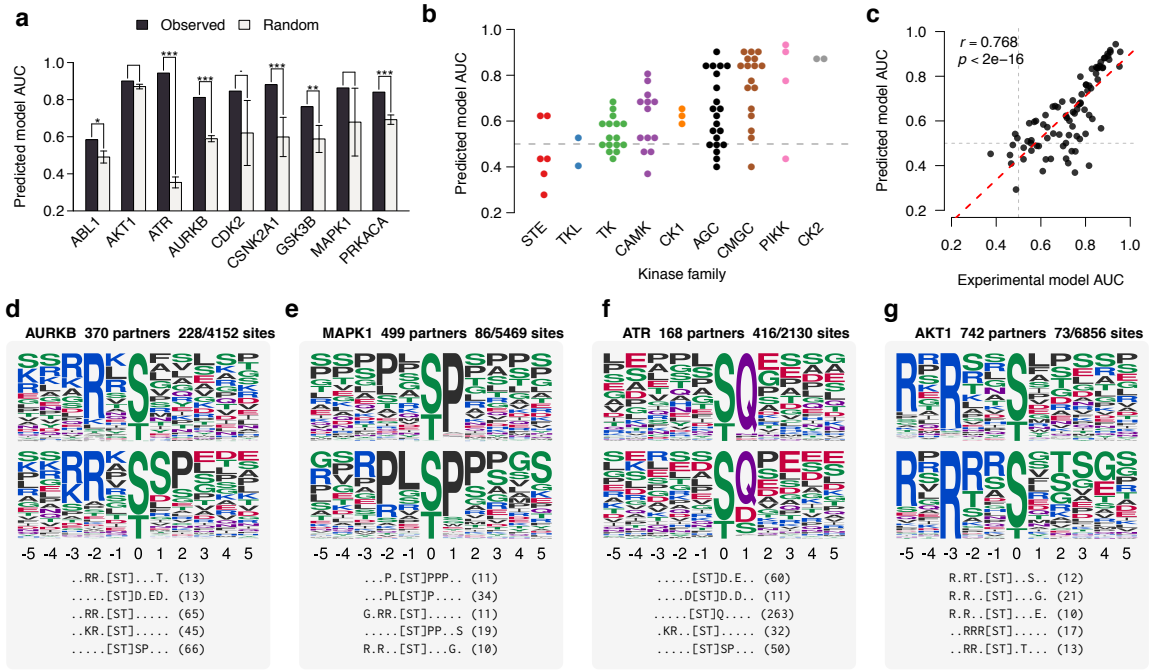



Fig. 2.7 Benchmarking of the method. (a) The performance of each predicted model compared with models predicted using random phosphosites. Seven of nine cases perform better than random (. $p$<.01, * $p$<0.05, ** $p$<0.01, *** $p$<0.001, one-sided $Z$-test). Error bars represent the median absolute deviation for 1,000 random models. 85 kinases with $\geq 20$ known substrates were used as the gold standard. (b) Performance of predicted models by kinase family. The grey line denotes near-random performance. (c) Performance of models constructed using experimental phosphosites is compared with that of the predicted models. A strong correlation suggests a relationship between the specificity of the kinase and predictability of a specificity model. (d-g) Examples of predicted specificity models. The top and middle panel of each example shows the specificity of the kinase as constructed from known phosphosites and as predicted by the described approach, respectively. The bottom panel shows the top five extracted motifs and the number of phosphosites matching them.

specificity in the kinase-substrate recognition. For example, many tyrosine kinases have additional targeting domains (i.e. PTB and SH2 domains) and several STE kinases are known to have an additional interaction surface known as "docking motifs" [267, 268]. For these kinases, targeting is achieved by multiple interfaces or often aided by other mechanisms and, therefore, may be less specific in the recognition of sequences flanking the phosphosite. In line with this reasoning, kinases that harboured additional protein domains also had poorer performing models ($p$=$1.88\times10^{-3}$, Figure 2.9). This notion was tested more explicitly by comparing predicted models with a proxy for kinase promiscuity. For the same set of kinases, experimental specificity models were constructed using the literature-defined phosphosites

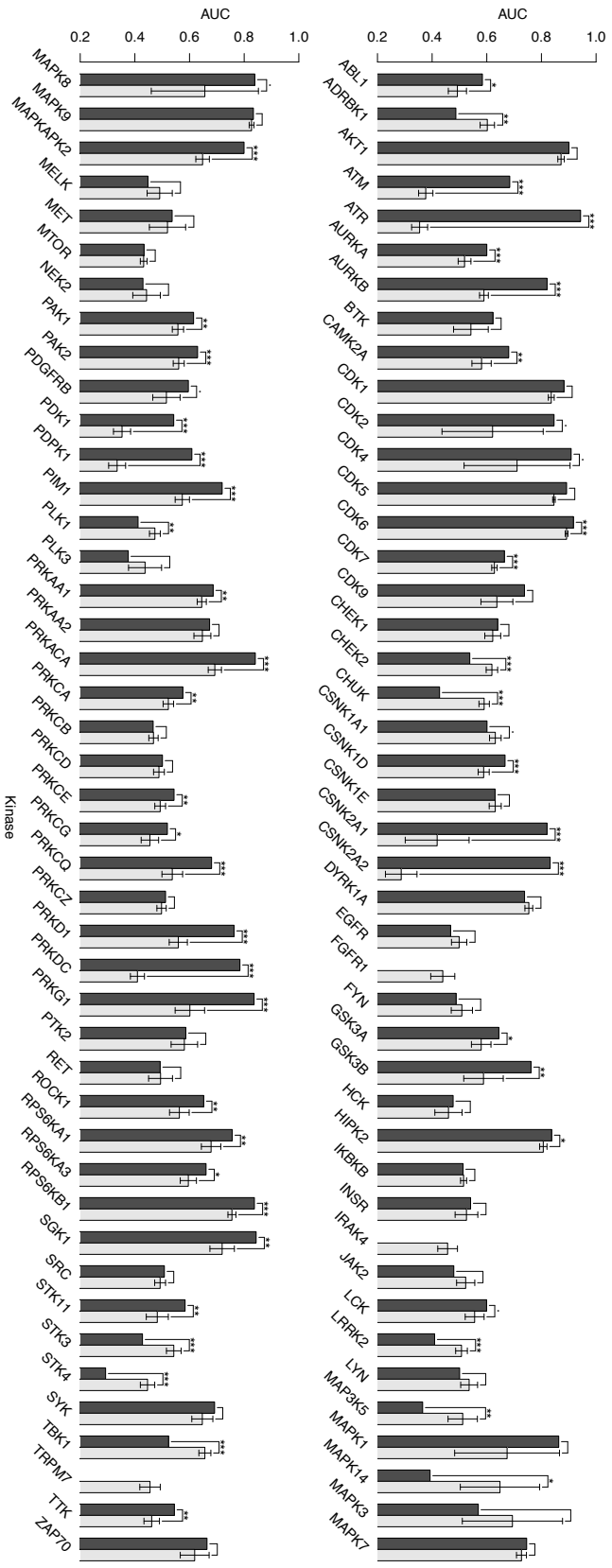



Fig. 2.8 Performance of 85 kinases with $\geq$20 known targets compared to that of random models (. $p$<0.01, * $p$<0.05, ** $p$<0.01, *** $p$<0.001, $Z$-test). Error bars represent the median absolute deviation of 1000 random models.

and performance was measured using 10-fold cross validation (Methods, section 2.3.4). Interestingly, a strong correlation was observed between the performance of predicted and experimental models ($r$=0.757, $p$=$2\times10^{-16}$, Figure 2.7c), suggesting that kinases with higher sequence specificity are more likely to have high predictability.

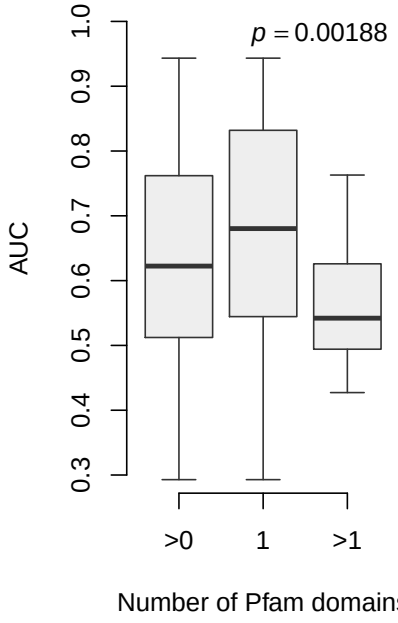



Fig. 2.9 Impact of the number of domains on predictions. Distribution of AUCs for kinases, separated by the number of Pfam domains they harbour. Kinases with multiple domains, overall, demonstrate significantly lower performance. Significance is measured using a one-sided Wilcoxon signed-rank test.

In attempt to identify features comprised by better performing models, the performance of the predicted models were correlated with several features including (1) the number of functional interacting partners, (2) the number of phosphosites on interacting partners, (3) the distribution of information content, and (4) the number of extracted motifs. Weak correlations ($r$=0.361, Figure 2.10) were observed for each of the individual features. A higher correlation was achieved by combining a number of features using a linear regression model ($r$=0.542, $p$=$8.37\times10^{-8}$, Figure 2.10). This model could thus be used to assign a quantitative measure of confidence related to the truth of predicted specificity, which is used here to rank predictions.

Several additional adaptations to the method were explored. For example, using different background sets for motif enrichment or restricting to high confidence phosphosites for motif enrichment.

In the current implementation, we exclude all P+1 phosphosites when predicting the specificity for kinases that are not proline-directed. This requires prior knowledge regarding

the different kinase families, which might not always be available for different PTM types or species. Rather than having to specify P+1 kinases, one can use all phosphosites as a background for motif enrichment to decrease the importance of P+1 motifs. However, this approach results in a moderate decrease in the mean AUC to 0.61 (Figure 2.11a) and fewer predictions (32 vs. 85). Furthermore, the importance of prolines and arginines at certain positions is decreased and in most cases are not enriched for, likely resulting in incorrect enrichments and ultimately models (Figure 2.11b).

To test if phosphosite quality impacted performance, we restricted phosphorylation data using two criteria. First, only phosphosites that were annotated to at least two PubMed articles were retained. Second, since MS methods are biased towards highly abundant proteins, phosphosites that occurred in the top 10% abundant proteins as defined in PaxDB [269] were removed. Models were predicted using both sets of filtered phosphosites and the performance was assessed. Overall, restricting phosphosites did not appear to improve the performance of the models (Figure 2.11a).

### 2.2.3 Mass spectrometry-based validation of kinase specificity

To validate predictions, four kinases with few literature-defined phosphosites spanning different kinase families were selected for experimental validation. These included CMGC kinases SRPK2 and HIPK2, AGC kinase AKT2, and PEK kinase EIF2AK4. For each of these kinases, *in vivo* target phosphosites were identified using the phosphoproteomic approach described by Imamura et al. [270] (Figure 2.12a). Briefly, HeLa cell extracts were treated with phosphatase to remove any existing phosphosites, and kinases were added in separate experiments. The phosphorylated extract was then subjected to trypsin digestion, phosphopeptide enrichment, and nanoLC-MS/MS (Methods, section 2.3.5, Figure 2.12a). This resulted in a total of 483 novel phosphosites being identified for these kinases (AKT2, $n$=248; EIF2AK4, $n$=91; HIPK2, $n$=106; SRPK2, $n$=38). The performance of predicted models for these kinases was then assessed against these sites (Figure 2.12b-e). All predicted models performed significantly better than random, and three of the four displayed an AUC $\geq 0.7$ at classifying the experimentally identified sites (Figure 2.12f). These results are in line with the benchmarks performed and further support the validity of the approach described here. We note that the SRPK2 kinase was predicted to have a strong preference for serines and arginines at several positions. This motif was unusual given previously described models, though several elements of this motif are confirmed by the experimental sites (Figure 2.12d). Furthermore, the kinase specificity of SRPK2 was recently determined using a chemical genetic approach [271] that identifies the conserved RXXSP motif for SRPK2. This provides further validation of the predicted specificity model of this kinase.

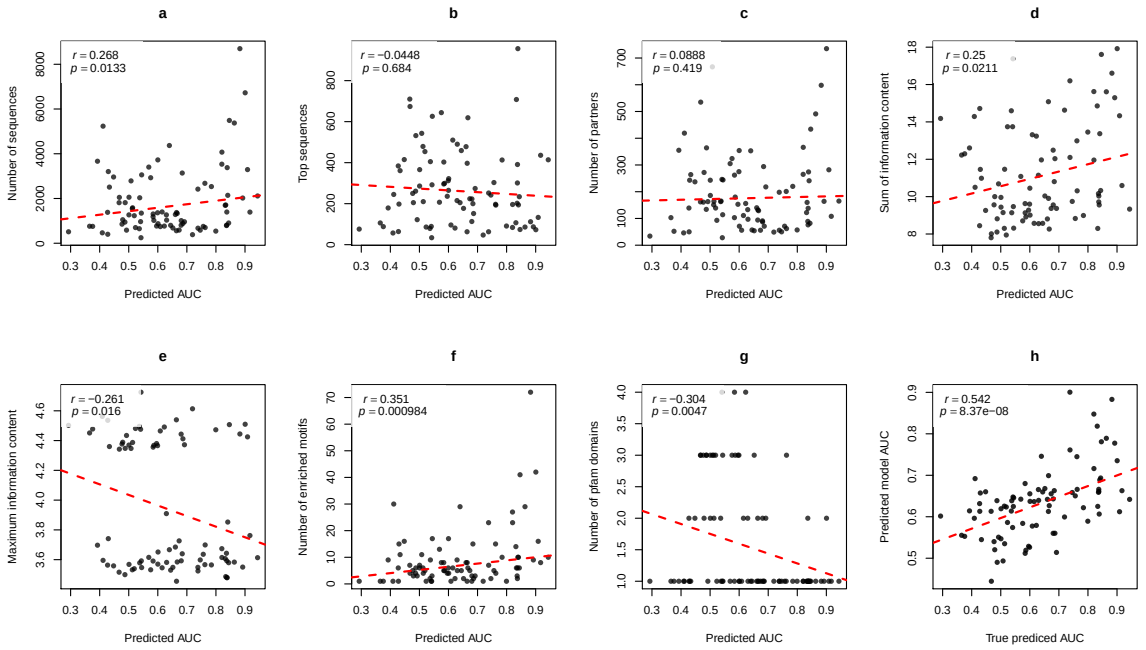



Fig. 2.10 Feature correlations. Performance of predicted models correlated with (a) number of phosphosite sequences on functional partners, (b) number of phosphosite sequences matching the top five enriched motifs, (c) number of functional partners, (d) sum of information content across positions of models (e) maximum information content amongst different positions, (f) total number of enriched motifs and (g) number of annotated Pfam domains. (h) A linear regression model built using a combination of features (a,d,e,f) is used to predict the AUC of predicted specificity models, which are correlated against the true AUC. The line of best fit is shown in red.

### 2.2.4 Prediction of post-translational modification binding specificities

To demonstrate the extensibility and application of the proposed method, it was applied to other types of linear motif specificities, such as that of 14-3-3 proteins. 14-3-3 proteins are conserved single domain proteins capable of binding a phospho-serine or threonine and are responsible for tight regulation of several important pathways including cell death, cell cycle control, and signal transduction [272]. Previous studies have shown that 14-3-3 proteins demonstrate distinct specificities towards their target phosphosites [273] (Figure 2.13a). We applied the method to human 14-3-3 proteins and similarly show that the recovered models are good predictors of known binding sites (AUC > 0.80) while performing significantly better than random (Figure 2.13b). Well-known determinants such as arginine at position -3 and some preference for proline at position +2 are recovered. Furthermore, little to no overlap was found between sites used to construct individual models of the 14-3-3 proteins, despite showing similar predicted specificity (Figure 2.13c-d). This suggests that the same motif is

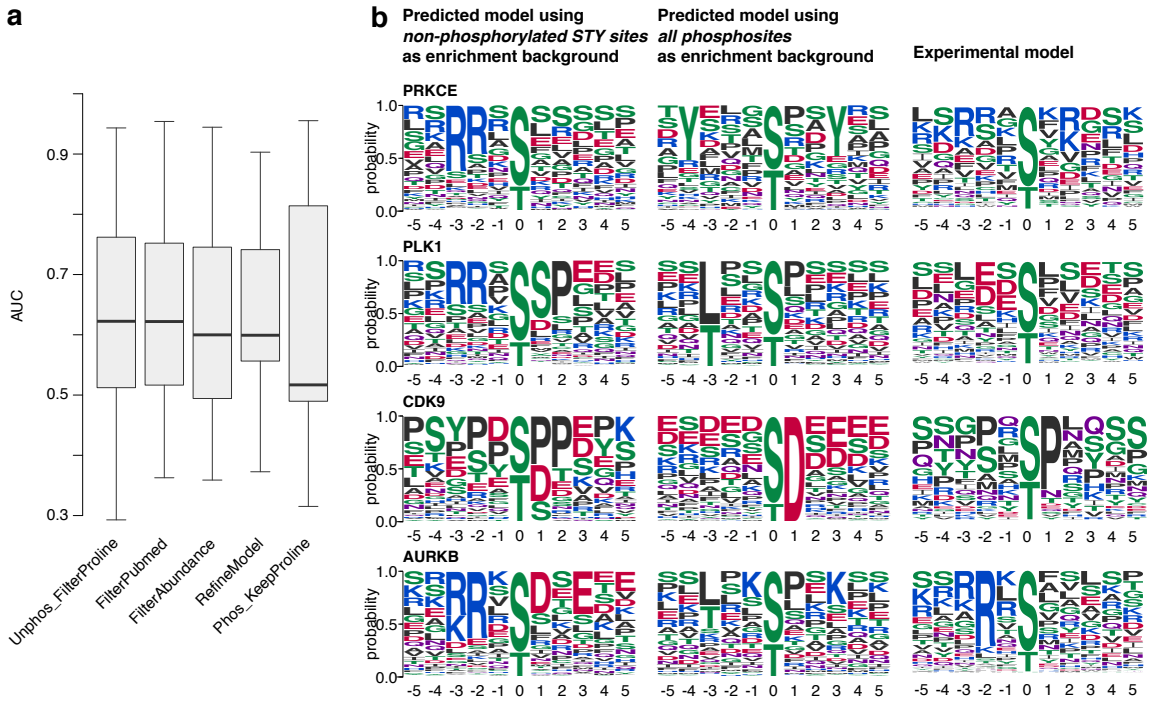



Fig. 2.11 Alternate motif enrichment background sets. (a) Performance of predictions using different background sets: (1) using unphosphorylated STY sites as background for motif enrichment, while filtering P+1 phosphosites versus, (2) restricting to phosphosites having two or more associated PubMed IDs, (3) removing phosphosites occurring in highly abundant proteins, (4) refining function partner phosphosites using the method described in Reimand et al. [148], and (5) using all phosphosites as a background while retaining P+1 phosphosites in non-proline-directed kinases. (b) Using phosphorylated sequences as background for motif enrichment, while retaining P+1 sequences for non-proline-directed kinases. Examples of predicted models for non-proline-directed kinases, using top five significant motifs. The left predicted model is using unphosphorylated sequences as the background for motif enrichment while filtering out P+1 sequences. The right predicted model is using all phosphosites as background for motif enrichment while retaining P+1 sequences.

recovered in each case, from a different source of partner sites, adding to the confidence of the recovered models.

The described method was also applied to the bromodomain-containing histone acetyltransferase p300. p300 has crucial roles in chromatin remodelling [274] and binds acetylated lysines with a well-characterized specificity [275]. A collection of 12,149 human lysine acetylation sites were obtained from dbPTM [116] and used along with the same network-based motif enrichment to predict the specificity of p300's bromodomain. The predicted specificity (Figure 2.13e) is very similar to the known preference for KXXK or KXXXK (where X is any amino acid and both lysines are acetylated).

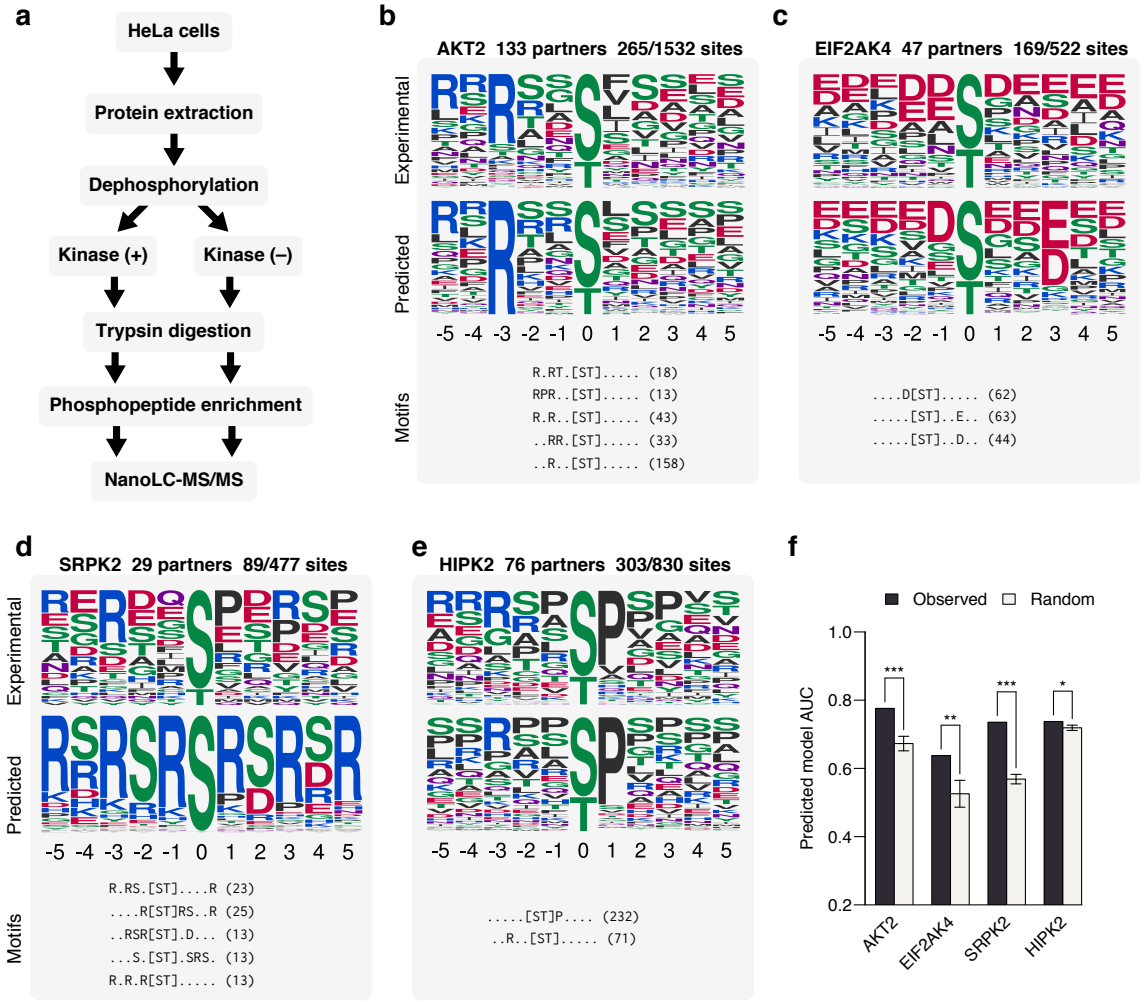



Fig. 2.12 Experimental validation. (a) Workflow for identifying phosphosites. (b-e) Predicted specificity models of four kinases that were selected for experimental validation of their target phosphosites. The top and middle panel of each example shows the specificity of the kinase as constructed from the experimental target phosphosites and as predicted models, respectively. The bottom panel shows the top five extracted motifs and the number of phosphosites matching them. (f) The performance of each predicted model compared with models predicted using random phosphosites.

These results recapitulate the benefit that PTM recognition specificity can be predicted by combining network information with PTM data.

### 2.2.5 Conservation of kinase-substrate specificity

Catalytic domains of many kinases, particularly those within the same families, are highly conserved across species [276], suggesting that their sequence specificity is likely conserved

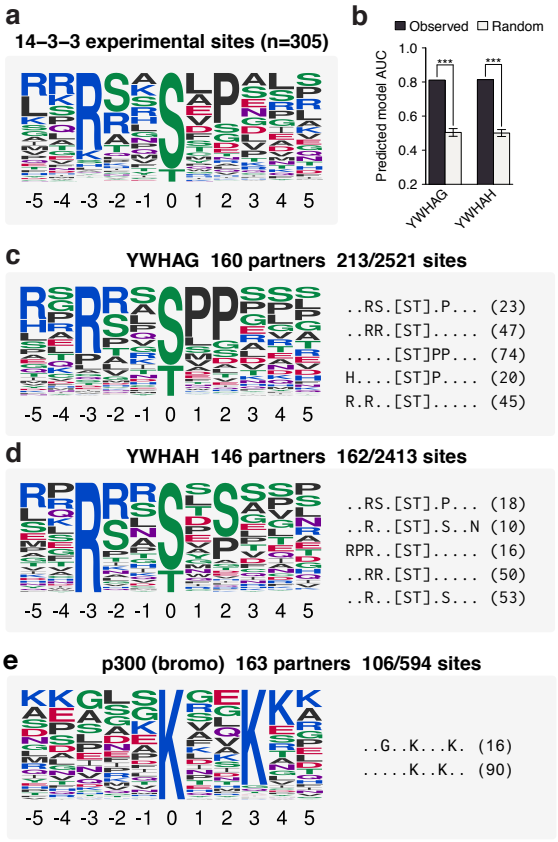



Fig. 2.13 Prediction of 14-3-3 domain specificities. (a) The specificity of 14-3-3 domains as constructed by experimentally verified substrates as reported by Johnson et al [273]. (b) The performance of each predicted model compared with models predicted using random phosphosites. Both cases perform better than random sampling of phosphosites (. $p$<0.1, * $p$<0.05, ** $p$<0.01, *** $p$<0.001, one-sided $Z$-test). Error bars represent the median absolute deviation of 1,000 random models.(c-d) Prediction of specificities for two 14-3-3 proteins. Each example shows a logo representing the predicted specificity (left) and the top five extracted motifs and the number of phosphosites matching them (right). (e) Prediction of acetylation-based specificities for the bromodomain-containing protein p300.

too. To test this hypothesis, the proposed model was applied to mouse (*Mus musculus*), which contained 29,732 phosphosites and 2,425,424 STRING interactions. Using human kinases with an AUC above 0.6, a total of 56 one-to-one ortholog kinases in mouse were identified, using the InParanoid resource [277]. Results displayed a close resemblance between the specificity determinants of human kinases and their corresponding mouse orthologs (Figure 2.14). Over 34% (19/56) of predicted mouse kinases demonstrated similar or better performance at predicting known human kinase sites than the orthologous human model. This suggests that at least these 19 kinase pairs have very conserved kinase preferences. For the remaining cases, one cannot confidently say that there is a divergence in specificity since we cannot rule an incorrect prediction.

### 2.2.6 The kpred resource for predicted kinase-substrate specificities

To facilitate visualization of results, predictions for all kinases are provided through a web resource (Figure 2.15a), kpred, which is available at https://evocellnet.github.io/kpred. Users can explore predicted specificity models through the sequence logo, investigate enriched

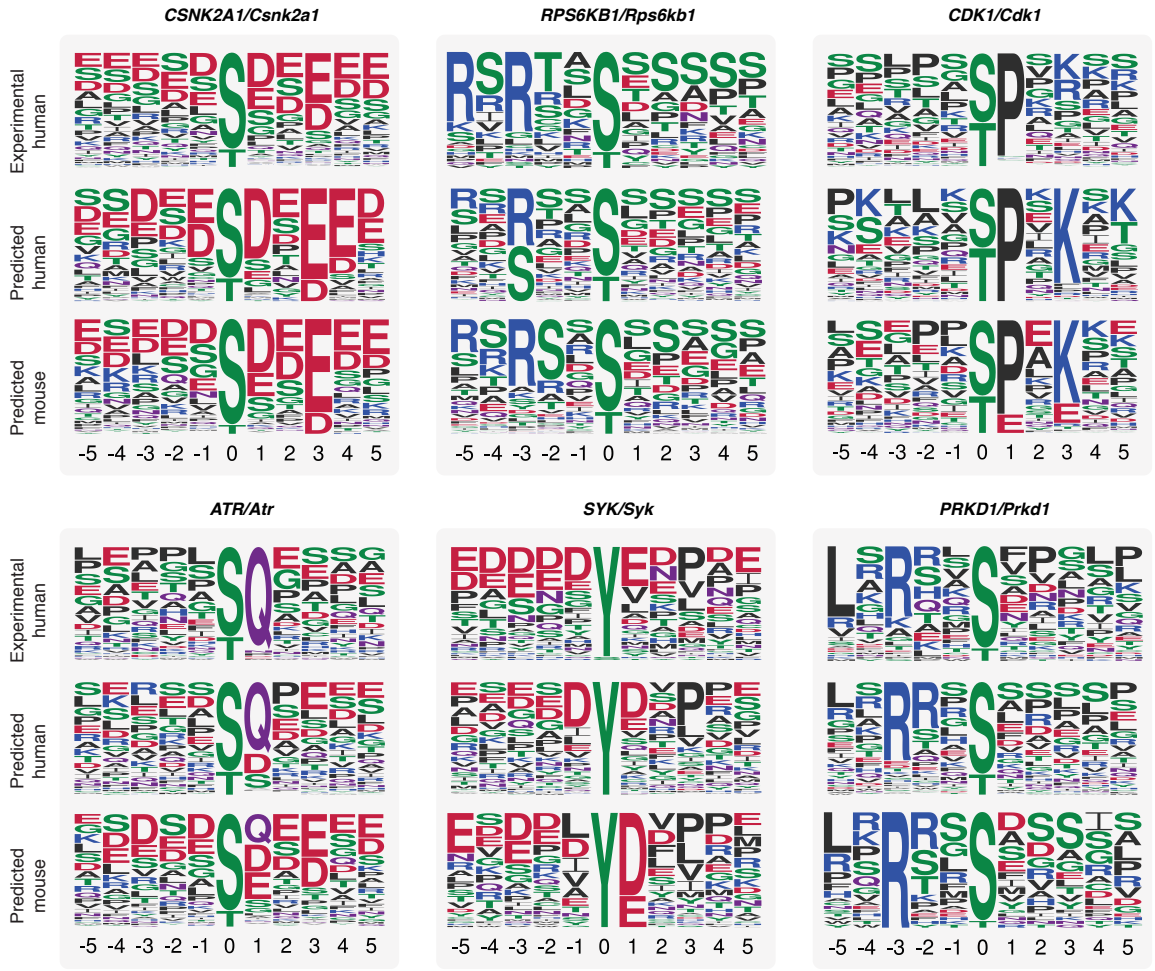



Fig. 2.14 Conservation of kinase specificity. (a-f) Six examples showing the comparison of predicted human versus mouse models. Each example shows logos for human gold standard specificity (top) and the predicted specificity model in human (middle) and mouse (bottom).

motifs and phosphosites and bulk-download prediction models. This result page is split into three main panels (Figure 2.15b-d), described in detail below.

The first panel highlights information on the kinase (Figure 2.15b). This includes the kinase family and group it belongs to and the kinase class (serine/threonine or tyrosine). Further information resulting from the method is also displayed such as whether P+1 sequences were removed prior to enrichment, the number of STRING partners and the number of phosphosites found on partners. The predicted model is interactively visualized as a sequence logo.

The second panel shows a table containing the top five enriched motifs used to identify phosphosites for construction of the predicted model (Figure 2.15c). The table also

includes additional motif information resulting from the motif-x [266] enrichment such as the enrichment score, the number of sequences matching phosphosites in the foreground and background datasets, and the fold increase of foreground over the background.

The final panel contains a table of phosphosites matching the enriched motifs and used to construct the predicted specificity model (Figure 2.15d). The table reports details of the phosphosites such as the sequence context of the site, its position, kinases known from the literature to phosphorylate this site (if any) and the source database of the phosphosite. The table also includes the STRING score between the kinase and the protein containing the phosphosite.

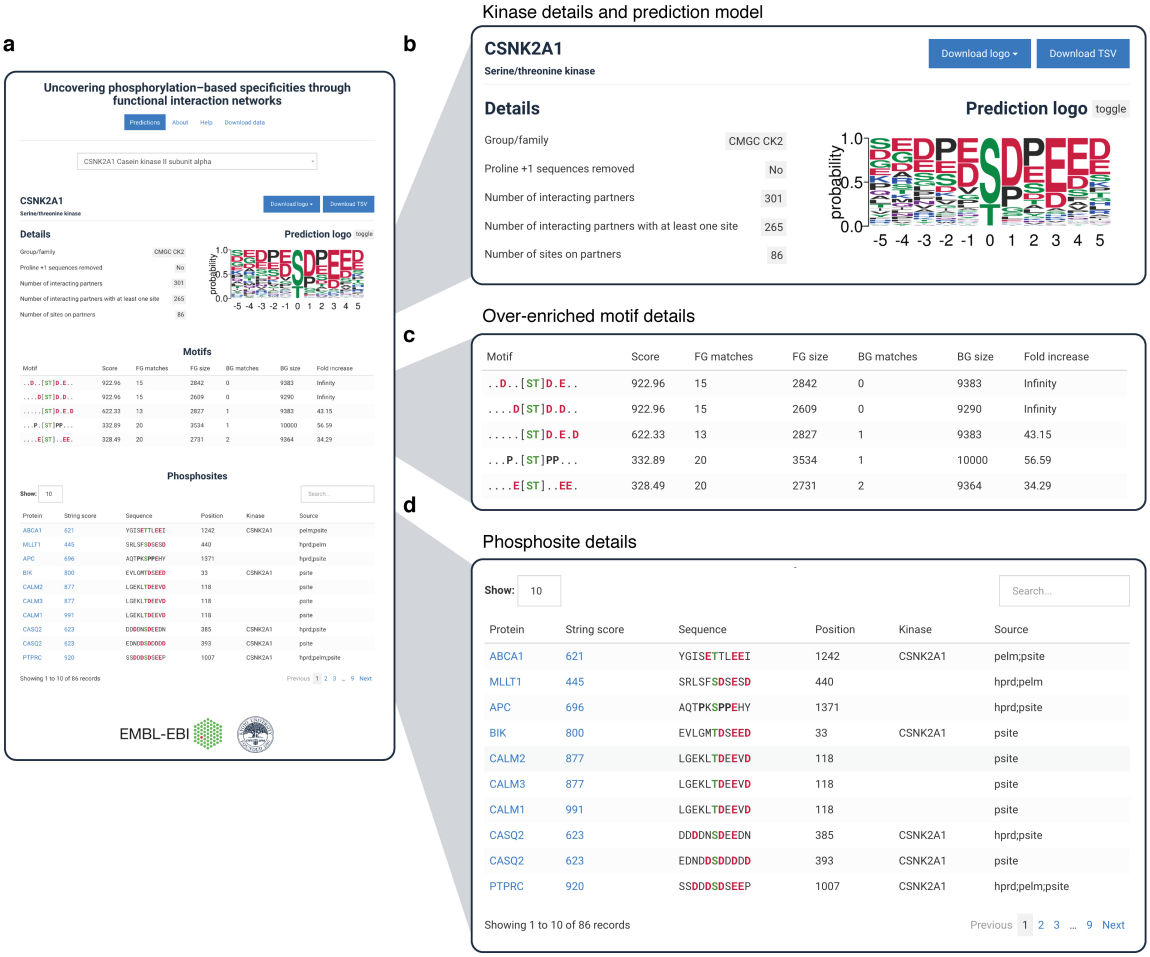



Fig. 2.15 The kpred resource. (a) Overview of a result page for the CK2 kinase CSNK2A1. The result page is split into three panels highlighting (b) the kinase and prediction model (c) the enriched motifs and (d) phosphosites used to construct the model.

Users can choose to download the logo in multiple formats or results as a flat file from the top of the result page (Figure 2.15b). The flat file is provided in a tab-delimited format

and similarly structured to the result page, where each panel is separated by //. Alternatively, logos and flat files for all kinases can be downloaded in batch from the download page.

## 2.3 Methods

### 2.3.1 Phosphorylation and functional interactions data collection

Functional interaction data were collected from STRING (v9.1). Phosphosites were collected from public databases, including PhosphoSitePlus [252], PhosphoELM [253], HPRD [254] and from a study of mouse tissues [278]. Phosphosites were then mapped to protein sequences provided by STRING. Kinase orthologs for 471/493 (95%) human kinases were obtained from InParanoid v8.0 [277].

### 2.3.2 Kinase domain prediction

Given a protein sequence, we used Kinomer [265], which uses multilevel HMMs and HMMER [256] to identify protein kinases and classify them into their appropriate kinase family. *E*-value cutoffs for each family were used as defined in Martin et al. [265]. If a kinase was predicted more than one family, the one with the highest *E*-value was retained. These families were also used to determine if the kinase is serine/threonine-specific or tyrosine-specific. A kinase is assumed to be either serine/threonine-specific or tyrosine-specific and does not account for dual specificity kinases.

### 2.3.3 Phosphorylation-based motif enrichment

The motif-x algorithm [266] was used to identify motifs enriched within a set of phosphosites, compared with a background (Figure 2.16). The background set used here was 10,000 11-mers centred on non-phosphorylated serine, threonine or tyrosine residues, depending on if the kinase is serine/threonine-specific or tyrosine-specific.

Given a foreground and background set of sites, motif-x first constructs a positional frequency matrix from the foreground set. This matrix contains the observed count $f_{ij}$ of each residue $j$ at position $i$. Similarly, for the background, a positional probability matrix is computed, containing the likelihood $p_{ij}$ of observing residue $i$ at a certain position $j$. Given the number of sites in foreground data $n$, the significance of residue/position pairs are then identified using a binomially distributed model:

$$P(n, f_{ij}, p_{ij}) = \sum_{k=f_{ij}}^{n} \binom{n}{k} {p_{ij}}^{k} (1 - p_{ij})^{n-k} \tag{2.1}$$

Significant residue/position pairs are then identified using a threshold of $P_{\text{binomial}} < 10^{-6}$. These pairs are used to construct a motif, which is reported. Sequences in the foreground and background data matching these motifs are then removed and the process repeats iteratively until one of the following criteria is met (1) no significant residue/pairs exist below the threshold or (2) fewer than 10 sites remain in the foreground or background data.

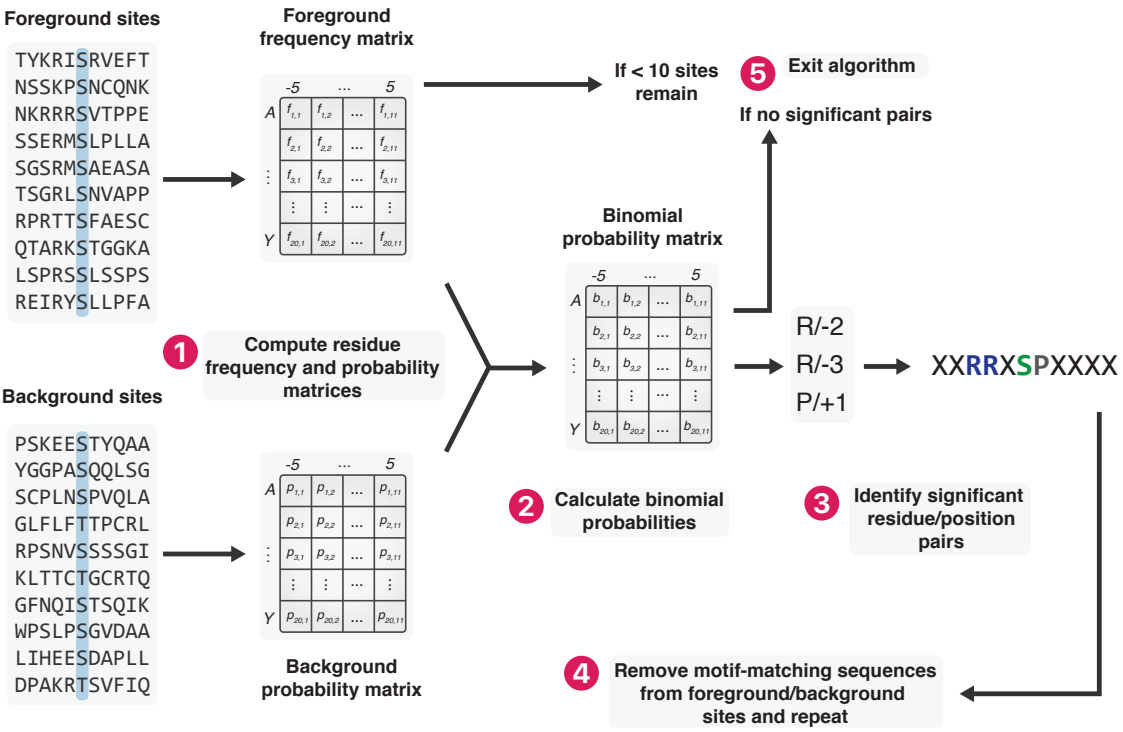



Fig. 2.16 Overview of the motif-x algorithm. (1) Foreground and background sequences are used to construct a frequency matrix and probability matrix, respectively which are used to (2) compute binomial probabilities. (3) Significant residue/position pairs are identified and used to construct the motif reported. (4) Sequences matching the reported motif are then removed from the foreground and background sets and this process repeats until (5) the algorithm converges (no significant pairs remaining) or too few input data remains.

A score $s$ reflecting the significance of the extracted motifs are calculated using the binomial probabilities of pairs used to generate the motifs:

$$s(motif) = \sum_{i}^{m} -log(b_i) \tag{2.2}$$

where $m$ is the total number of significant residue/position pairs used to generate the motif and $b_i$ is the binomial probability of the $i^{th}$ residue/position pair. This score is ultimately used to define top significant motifs.

Since the motif-x tool was only available via an online web server, the algorithm was re-implemented for the R programming language and made publicly available at https://github.com/omarwagih/rmotifx.

### 2.3.4 Kinase specificity models and performance assessment

Specificity models were constructed as PWMs, which are commonly used to model specificities of linear motifs [58]. PWMs can then be used to score peptides. A single PWM is constructed using a set of phosphosites. If $S$ is a set of $n$ phosphosites, each of length $l$, $s_1, \cdots, s_n$, where $s_k = s_{k1}, \cdots, s_{kl}$ and $s_{kj}$ represents one of the 20 amino acids. A PWM $M_{20\times l}$ with weights $p_{ij}$ as the relative frequency of each amino acid $i$ at a particular position $j$ is constructed as follows:

$$p_{ij} = \frac{1}{n}\sum_{k=1}^{n} f_i(s_{kj}) + \varepsilon \qquad f_i(q) = \begin{cases} 1, & \text{if } i = q. \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

Where $\varepsilon$ is a pseudo-count added to each frequency value to avoid infinite values upon log transformations.

An adapted version of the matrix similarity score (MSS), originally developed in the MATCH algorithm [279], as described in Wagih et al. [146] is then used to score a phosphosite $q$ also of length $l$, $q_1, \cdots, q_l$. The MSS uses the positional information content to assign position-specific weights of importance. Additionally, scores are normalized against the highest and lowest relative frequencies per position in the PWM. This results in a score $p_{ij}$ reflecting the likelihood of binding ranging from 0 to 1, where 0 represents no binding and 1 represents a perfect match. The MSS defined as:

$$
\begin{aligned}
MSS &= \frac{Current - Min}{Max - Min} \\
Current &= \sum_{j=1}^{l} I(j) p_{q_j,j} \\
Min &= \sum_{j=1}^{l} I(j) p_j^{min} \\
Max &= \sum_{j=1}^{l} I(j) p_j^{max} \\
I(j) &= -\sum_i p_{i,j} log(\frac{p_{i,j}}{p_b})
\end{aligned}
\tag{2.4}
$$

Here, $q_j$ denotes the residue at position $j$ of the query sequence, $p_j^{min}$ and $p_j^{max}$ denote the minimum and maximum relative frequency at position $j$ of the PWM, respectively, and $p_b$ is the background frequency of a particular amino acid in the proteome.

The performance of a given PWM was evaluated using the AUC, which is the curve representing the relationship between the false positive rate and true positive rate as the MSS score cutoff is varied:

$$
FPR = \frac{FP}{FP + TN} \qquad\qquad TPR = \frac{TP}{TP + FN} \tag{2.5}
$$

Here, FP, TP, TN, FN represent the number of false positives, true positives, true negatives, and false negatives, respectively. The PWM is used to score positive and negative sequences in order to generate these values. For a kinase of interest, the positive sequences are defined as the set of phosphosites annotated to the kinase, whereas the negative sequences as phosphosites annotated to any kinase not belonging to the same kinase family, where the kinase classification is defined by Manning et al. [122].

In the case where the performance of experimental models was evaluated (i.e. using the gold standard sequences), 10-fold cross-validation was carried out. The kinase sequences are split into 10 random bins and each bin is iteratively used as the test set, while the remaining nine are used to construct the PWM. This results in 10 AUCs, which are then averaged to provide an unbiased proxy of the PWM's prediction power.

### 2.3.5 Profiling *in vitro* kinase substrates

Identification of *in vitro* kinase substrates was carried out by collaborators Naoyuki Sugiyama and Yasushi Ishihama at the University of Kyoto using the previously described approach by Imamura et al. [270]. Briefly, lysate proteins were extracted from HeLa S3 cells at about 80% confluence in 15 cm dishes, and the total protein amount was measured by a BCA protein assay kit. Dephosphorylation was then carried out with TSAP (Promega, Madison, MI, USA) at 37°C for 1 h, and TSAP was inactivated by heating to 75°C for 30 min. For *in vitro* kinase reactions, each 100 µg µl$^{-1}$ of dephosphorylated proteins (1 µg/µl) was reacted with 1 µl of each recombinant kinase (0.5 µg/µl) or distilled water as a control at 37°C in kinase reaction buffer (40 mmol Tris-HCl at pH 7.5, 20 mmol $MgCl_2$, 1 mM ATP) for 3 h. AKT2, catalytic domain [120–481(end), accession NP_001617. 1], full-length EIF2AK4 [1–1649(end) accession Q9P2K8.2], full-length HIPK2 [1–1198(end) accession Q9H2X6] and full-length SRPK2 [1–688(end) accession NP_872633.1] were obtained from Carna Biosciences Inc. (Kobe, Japan). The kinases were expressed as N-terminal GST-fusion protein using the baculovirus expression system with SF9 cells and were purified using glutathione Sepharose chromatography. The reaction was stopped by heating to 95°C for 5 min. After protein reduction/alkylation, Lys-C/trypsin digestion (1/100 w/w) was performed and phosphopeptides were enriched by $TiO_2$-based hydroxyl-acid-modified metal oxide chromatography [280].

Phosphopeptides were desalted by StageTips and analyzed by nanoLC-MS/MS using a self-pulled analytical column (150 mm length × 100 µm) inner diameter) packed with ReproSil-Pur C18-AQ materials (3 µm), Dr. Maisch, Ammerbuch, Germany). An Ultimate 3000 pump (Thermo Fisher Scientific, Germering, Germany) and an HTC-PAL autosampler (CTC Analytics, Zwingen, Switzerland) were used coupled to an LTQ-Orbitrap XL (Thermo Fisher Scientific). A spray voltage of 2,400 V was applied. The MS scan range was m/z 300–1,500. The top 10 precursor ions were selected in MS scan by the Orbitrap with r = 60,000 for MS/MS scans and the ion trap in the automated gain control (AGC) mode, where automated gain control values of 5.00 × 105 and 1.00 × 104 were set for full MS and MS/MS, respectively. To minimize repetitive MS/MS scanning, a dynamic exclusion time was set at 20 s with a repeat count of 1 and an exclusion list size of 500. The normalized CID was set to be 35.0. Mass Navigator v1.2 (Mitsui Knowledge Industry, Tokyo, Japan) was used to create peak lists on the basis of the recorded fragmentation spectra with the default parameters for the LTQ-Orbitrap XL. Peptides and proteins were identified by automated database searching using Mascot v2.3 (Matrix Science, London, UK) against SwissProt release 2010_11 (02/11/2010, 522,019 entries). A precursor mass tolerance of 3 ppm, a fragment ion mass tolerance of 0.8 Da, and strict trypsin specificity allowing for up to two

missed cleavages. Carbamidomethylation of cysteine was set as a fixed modification and oxidation of methionines; phosphorylation of serine, threonine, and tyrosine was allowed as variable modifications. Peptides were considered identified if the Mascot score was over the 95% confidence limit based on the "identity" score of each peptide and if at least three successive y- or b-ions with a further two or more y-, b-, and/or precursor-origin neutral loss ions were observed, based on the error-tolerant peptide sequence tag concept. After identification, phosphopeptides identified from the control samples were rejected. A randomized decoy database created by a Mascot Perl program gave a 1% false-discovery rate for identified peptides with these criteria. Phosphosite localization was evaluated using a site-determining ion combination method based on the presence of site-determining y- or b-ions in the peak lists of the fragment ions, which supported the phosphosites unambiguously.

## 2.4 Discussion

The advances in MS have expanded tremendously our knowledge of exact protein modifications sites for a number of different PTM types. However, there is almost no information regarding the regulatory interactions connecting regulators to target proteins. Determining the recognition preferences for PTM enzymes and binding domains in large scale is still an open problem and remains a limiting factor in achieving this goal. In this chapter, we used phospho-regulation as a model system and showed that it is possible to combine PTM information with interaction network data to derive accurate models of enzymes and binding domains. A resource that contains all of the information used for the specificity predictions of each kinase can be accessed from http://evocellnet.github.io/kpred. The code required to apply this approach can be found in the help page along with a tutorial.

It should be noted that even though some models do not perform better than that of randomly sampled sites, this does not necessarily reflect the reliability of the predicted model. Some kinase specificities are well-modelled by the most common motifs that are recovered from a random sample. For these cases, the added information from the network data does not result in a model that is more accurate than random. The power of the proposed approach is, therefore, more obvious for regulators that have specificities that are less common such as the DNA damage kinase ATR. For this kinase, the recovered model is accurate (AUC = 0.94) while performing much better than models produced by random sites.

In the current implementation of this approach, kinases that are not CMGC are assumed to not be proline-directed and P+1 phosphosites are removed prior to enrichment. This may result in mispredicting cases where a non-CMGC kinase is proline-directed and where CMGC kinases are not proline-directed. An alternative approach that does not require the removal of

P+1 phosphosites was tested but, overall, yielded a lower performance. However, in cases where P+1 phosphosites are retained, motifs can be recovered that are proline-directed. For example, the CSNK2A1 kinase is a casein kinase belonging to the CMGC group. It is one of the few non-proline-directed kinases belonging to the CMGC family. CSNK2A1 is known to have a strong preference for acidic residues, predominately C-terminal to the phosphosite [281]. Despite that P+1 phosphosites were not removed prior to enrichment for this kinase, a strong bias for an acidic residue at positions +1, +3 and +4 is recovered (Figure 2.14). Additionally, the class of the kinase (i.e. serine/threonine or tyrosine) is required a priori to filter only phosphosites matching the class of the kinase. This is primarily due to the fact that phosphotyrosine is in many regards a different PTM from that of phosphoserine and phosphothreonine. In particular, it occurs at a much lower frequency and thus, if one would not discriminate between these two types, the predicted specificities would be dominated by serine/threonine phosphosites.

It is important to take into account that most phosphosite information was retrieved from phosphoproteomics experiments that have used trypsin for protein digestion. Given that trypsin cleaves C-terminal to arginine and lysine residues, it is possible to expect a bias for arginine or lysine residues in the phosphopeptides. One would otherwise expect any bias to be equally possible at positions before or after the phospho-residue and also not specifically biased for arginine or lysine. Instead, arginine determinants are more frequent than lysine determinants and are not symmetrically distributed. Of the 202 Arg determined positions (defined as having >0.25 relative frequency at the position), 96% (194/202) are found N-terminal to the phosphosite, whereas 0.039% (8/202) are found C-terminal to the phosphosite. Overall, there are only 19 positions where lysine is the major determinant, and these tend to be more evenly distributed with 42.1% (8/19) occurring N-terminal to the phosphosite and 57.89% (11/19) occurring C-terminal to the phosphosite. Thus, this bias is unlikely to influence the recovered motifs.

Kinase families show different average performance in their predictions and the performance of gold standard specificity models is correlated with that of the predicted models. These observations highlight the inherent limitation of the approach proposed here. PTM-interacting proteins that recognize their target sites mostly by residues flanking the target phosphosite will be more amenable to this approach than those that use multiple recognition mechanisms. These include docking motifs, colocalization, coexpression, and scaffolding interactions [123]. In addition, this approach assumes that the recognition occurs in a linear epitope at the PTM position. It has recently been shown that kinase targeting can also occur in a three-dimensional epitope [282]. As such, this linear motif enrichment strategy would not be appropriate if a PTM enzyme or binding domain often recognizes the target site

through three-dimensional epitopes. These observations should be taken into account prior to future use of this method on other PTM recognition domains.

The proposed method was also successfully applied to different modes of site-directed motif-binding domains, such as 14-3-3 domains and bromodomains, suggesting that the method could thus be extended further to analyze specificities of other PTM recognition domains. Finally, this approach was applied to study the conservation of kinase specificity between human and mouse kinases. For the kinases analyzed, at least 34% appear to have conserved specificity. Thus, in combination with an analysis of potential mutations in specificity-determining residues, this approach could be used to identify PTM recognition domains with diverged specificities across species. Given that these regulators interact with many different target PTMs, it is expected that their specificity diverges slowly. This is in contrast to the fast changes in PTMs targeted by these proteins, that can diverge more quickly [283, 284]. Conserved regulator specificity with diverged target phosphosites is a scenario that is analogous to what is observed in transcriptional regulation [285]. However, there have been cases described for the divergence of transcription-factor specificity [286], suggesting that analogous cases of PTM divergence recognition are likely to exist. In addition to studying the evolution of specificity, applying this method to different organisms could lend further confidence to the true specificity of a PTM recognition domain, since models trained in different species could contribute complementary specificity determinants and ultimately be combined to provide better models.

In summary, we describe here a novel approach to predict PTM recognition motifs and believe it can be applicable to a wide range of recognition domains and contribute significantly to our understanding of these signalling systems.

# Chapter 3

# Assessing performance of methods for predicting impact of variants on transcription factor binding

*In this chapter, I employ over 146,000 allele-specific binding (ASB) ChIP-seq variants across 43 TFs as a gold standard to assess how TF-binding models across five different methods perform at predicting variant impact. I compare the performance of different methods and explore alternative mechanisms beyond sequence specificity that may be altered by variants. This work was carried out by myself in collaboration with the company Deep Genomics and under the supervision of Daniele Merico, Andrew Delong and Brendan Frey.*

## 3.1 Introduction

Gene expression is a tightly regulated process governed by a multitude of variables [42]. One of the primary mechanisms contributing to the regulation of gene expression is the binding of TFs to regulatory genomic elements. Differential gene expression can drive and contribute to almost every aspect of disease phenotypes. Understanding the intricate process of TF-DNA binding can, therefore, provide mechanistic hypotheses for variants and propel the discovery of novel therapies. Fortunately, through the aid of high throughput techniques such as ChIP-seq, SELEX and PBMs, the binding specificities of many TFs have been exhaustively catalogued over the past decade [287].

Genetic variation falling within specificity determinants of TFBSs can alter binding by introducing novel binding sites or diminishing existing binding sites, often resulting in a substantial impact on molecular phenotypes through changes in gene expression. Approaches

such as pooled ChIP-Seq have been used to experimentally map variants to molecular-level traits such as TF-binding [288, 289]. These approaches are, however, costly and cannot yet be routinely applied to the sizeable quantity of genetic variation data available. As such, much effort has gone into modelling TF-DNA binding *in silico*, which range from rudimentary approaches such as the PWM to state-of-the-art deep-learning methods. These methods have also been employed to predict variants likely to alter TFBSs and have thus become an essential component of many variant prioritization pipelines.

The performance by which TF binding models are able to distinguish their binding regions from random genomic regions has been well characterised [290, 291]. To assess how well these predictors perform at identifying the impact of variants, known regulatory variants are often employed, which include variants from the Human Genome Mutation Database (HGMD), GWAS-derived variants, and QTLs [292, 293, 103]. However, little has been done to explore the ability of these models to assess the impact of genetic variants on binding in a TF-specific manner. ASB ChIP-seq provides a valuable dataset to carry out such performance assessments. Here, ChIP-seq reads are mapped to either allele of heterozygous variants within an individual or cell line, allowing for the explicit identification of variants that alter TF occupancy. Several studies have utilised ASB data to explore TF-specific performance at assessing variant impact. For instance, Zeng et al. used a small number of ASB variants for six TFs to validate their GERV method at identifying TFBS-altering variants [294]. Shi et al. compiled a dataset of over 10,000 ASB variants across 45 ENCODE ChIP-Seq datasets and demonstrated that ASB variants lie within highly relevant PWM positions [105]. These studies are, however, often based on a small number of TFs or are focused on individual variant impact methods.

In this chapter, we aimed to carry out a systematic and unbiased analysis of the performance of TF-binding models at assessing variant impact. Using a compiled compendium of over 20,000 ASB variants across 100 TFs, we compare a total of five methods. We devise several scoring metrics for each model and assess how they affect the identification of impactful variants. We also explore the performance of TFs individually, identifying TFs that are able to accurately predict variant impact as well as those that, although have distinct binding specificities, are unable to do so. We explore mechanisms that may explain this poor performance and suggest improvements for modelling impact of variants on TFBSs. This study offers novel insight into non-coding variant impact prediction in TFBSs.

## 3.2 Results

### 3.2.1 A compendium of allele-specific binding data

To assess the performance of TF-binding impact predictions, we require a set of variants known to alter binding or have no effect on binding. This is conveniently provided by ASB ChIP-seq data. ASB variants were collected from five studies [106, 104, 295, 105, 296], with each study providing heterozygous variants, the sample or cell line from which it was obtained, and reference and alternate allele read counts and the TF affected (Figure 3.1a). If an ASB variant was reported across multiple studies, the one with the highest number of total mapped reads was retained. Variants with at least 10 reads mapping to the reference or alternate allele were retained, for a total of 146,947 TF-variant pairs (ASB events) reported across 94 TFs. The largest fraction of TFs was reported in a single study, with a total of 50 TFs and as few as three TFs were reported across all five studies (Figure 3.1b). Different studies also contained a disproportionate number of TFs and samples for which ASB data was available. The largest number of TFs was contained within the Santiago et al. dataset with a total of 80 TFs across from 14 samples [106] (Figure 3.1c-d).

The binomial test was used to define how the significance of the imbalance between the reference and alternate read counts (Methods, section 3.3.1), which is commonly used in ASB studies [104, 297]. ASB variants that exhibit significant differences between reference and alternate read counts were defined by a significance threshold $P_{\text{binomial}} < 0.01$, resulting in 21,183 ASB events, of which 57.5% (12,715) were loss events where the alternate read count is lower and 9,397 (42.5%) were gain events, where the reference read count was lower. A total of 54,826 balanced reads, or non-ASB events, were defined as those with $P_{\text{binomial}} > 0.5$ (Methods, section 3.3.1). In total, 46 TFs had at least 10 non-ASB and ASB variants, while 43 TFs had at least 20 non-ASB and ASB variants. To our knowledge, this is the largest available ASB dataset compiled.

ASB variants are implicated in altering TF-binding and should be less likely to exist with high frequency. We confirmed this by analysing the proportion of ASB variants which are rare at a MAF < 1% using data from the ExAC consortium [298], 1000 genomes project [299], and the ESP6500 project [300] (Methods, section 3.3.4). ASB variants consistently demonstrated a higher fraction of rare variants, compared to non-ASB variants ($p<9.1\times10^{-3}$, Figure 3.1f). We also assessed whether commonly used non-coding variant impact predictors, such as GWAVA [301], Eigen [302] and CADD [303], could accurately distinguish ASB variants from non-ASB using the area under the receiver operating characteristic curve (AUROC) measure (Methods, section 3.3.4). However, near-random performance was observed for all three methods (CADD AUROC = 0.51, Eigen AUROC = 0.46, GWAVA AUROC = 0.48,

Figure 3.1g). This suggests that current approaches, which do not incorporate TF specificity are unable to identify variants altering TF-binding.

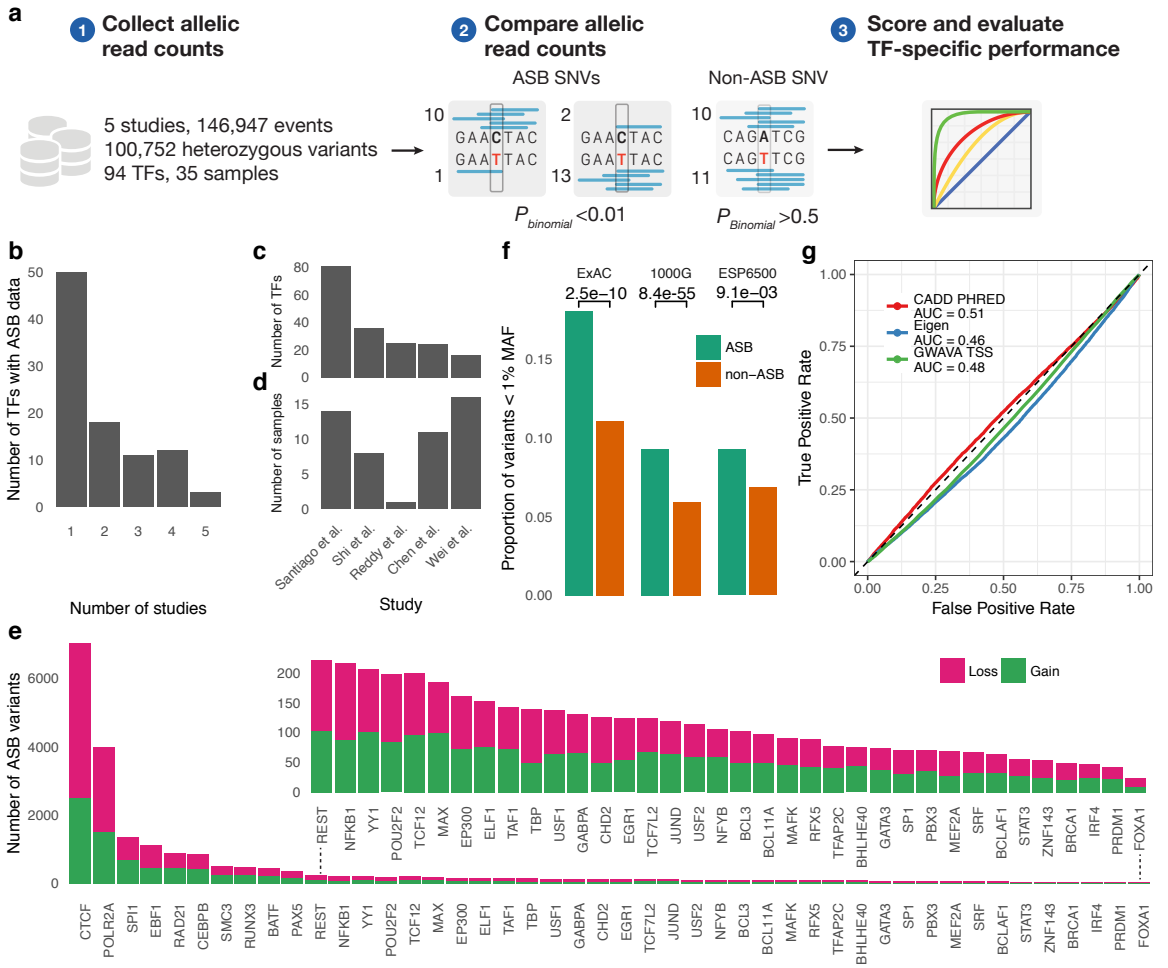



Fig. 3.1 Properties of ASB and non-ASB variants. (a) The use of ASB data for assessing the performance of TFBS variant impact. (b) The total number of TFs covered by a different number of studies. Only three TFs have ASB data in all five studies. (c-d) The number of TFs and samples per ASB study. (e) The number of ASB variants per TF at a $P_{\text{binomial}} < 0.01$ and at least 10 reads mapped to either allele. Only TFs with at least 20 ASB variants are shown. Loss and gain ASB variants are shown in magenta and green, respectively. (f) ASB variants (green) are relatively rare compared to that of non-ASB variants (orange). Significance $p$-values represent a one-sided Fisher's exact test. (g) Non-coding variant impact predictors are unable to distinguish ASB variants from non-ASB.

We utilised the collected ASB data to assess and compare the performance of several computational predictors of TF-binding variant impact (Figure 3.1a). The approaches included in the analysis were those based on PWMs [287, 68], $k$-mer-based approaches

GERV [294] and gkmSVM [103], and deep learning-based approaches DeepBind [68] and DeepSEA [69].

### 3.2.2 Scoring metrics for evaluation of transcription factor binding variant impact

The different methods available offer a variety of scoring metrics that describe the quantitative impact of a variant on TF-binding. These metrics are typically signed, where strong negative and positive values indicate loss and gain, respectively. DeepSEA produces a single probability of binding for both the wildtype and mutant sequences and uses two metrics to quantify the impact of a variant: the difference (*diff*) and log fold change (*log FC*) between the probabilities (Methods, section 3.3.3). gkmSVM provides a single score *deltaSVM* reflecting the change in the sum of *k*-mer weights for wildtype and variant sequences and GERV provides a single unsigned score (*GERV score*) that reflects the change in predicted ChIP-seq read counts (Methods, section 3.3.3).

In contrast, PWMs and DeepBind only provide a score reflecting the likelihood of binding and not the impact of a variant. For these approaches, we devise a number of metrics to assess the impact of a variant. Because the TF specificity models receive as input a fixed-length sequence, we score multiple overlapping sequences ("*k*-mers") along the region of interest with the reference and alternate allele. The defined metrics serve as a good starting point for assessing how different approaches perform at scoring the variant impact on TF-binding and are described in more detail.

#### Raw score metrics of variant impact

The difference in raw model scores is typically used to assess the impact of a variant [304–306] (Figure 3.2). We similarly define *delta raw* as the maximum difference between the raw wildtype and mutant scores across the *k*-mers. Because TF-binding can be made robust through homotypic clusters of redundant binding sites, they can often mitigate effects of impactful variants [307]. In line with this reasoning, we devised *delta track* as the difference between the maximum of all wildtype *k*-mer scores and the maximum of all mutant *k*-mer scores. Both metrics are signed, such that losses are indicated by negative scores and gains by positive (Table 3.1, Methods, section 3.3.3).

#### Probability-transformed metrics of variant impact

We sought to aid interpretability and strengthen baselines for variant effect prediction. To do this, we convert raw scores (which are not on any particular scale and not comparable

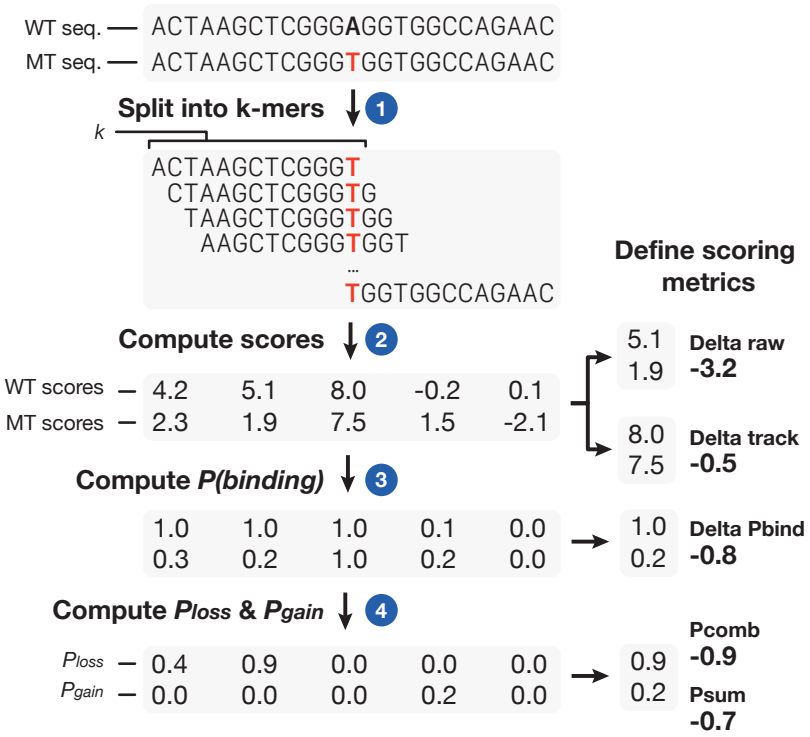



Fig. 3.2 Defining TFBS variant-impact scoring metrics. (1) Wildtype and mutant $k$-mers flanking the variant position are scored. (2) The generated raw scores are used to derive both the *delta raw* and *delta track* metrics. (3) $P_{bind}$ values are then computed for each wildtype and mutant $k$-mer. (4) $P_{loss}$ and $P_{gain}$ scores are defined using the generated $P_{bind}$.

across TFs) to likelihoods of binding (which are normalized to [-1, 1] and are comparable across TFs). We define positive and negative sequences as those used to train the DeepBind or PWM model and random genomic regions, respectively (Methods, section 3.3.3). We found that, in some cases, distributions of raw scores from the background followed a normal distribution and in some cases, distributions of foreground scores were bimodal with one component of scores exhibiting similar properties to that of the negative distribution. Because the ChIP-seq/SELEX data obtained was used to train deep learning models, this is likely due to lenient threshold used to call the ChIP-seq peaks, which was done to maximise the number of sequences available to train models. As such we use a Gaussian mixture model (GMM) to learn the two components comprising the foreground distribution. One component is fixed to the parameters of the negative distribution and the "true positive" component is learned. A linear model is then trained and used to compute the probability of binding ($P_{bind}$) from raw PWM or DeepBind scores (Methods, section 3.3.3).

Using the $P_{bind}$ score, we define the *delta* $P_{bind}$ score as the maximum difference between the mutant and wildtype $P_{bind}$ probabilities across all $k$-mers. This value ranges from -1 to 1,

where low negative values indicate a loss of binding and high positive values indicate a gain of binding.

We additionally define a probabilistic score $P_{loss}$ and $P_{gain}$ that range from 0-1 reflecting the likelihood of a binding site being lost or gained, respectively. For $P_{loss}$, this is computed by taking the joint probability of binding for the wildtype sequence and the probability of the mutant not binding and vice versa for $P_{gain}$ (Figure 3.2a, Methods, section 3.3.3). $P_{loss}$ and $P_{gain}$ are combined into a single score by first signing $P_{loss}$ negatively and computing the probability with the higher absolute value as $P_{comb}$. Since $P_{loss}$ and $P_{gain}$ can have low to moderate magnitudes we also compute $P_{sum}$ as the sum of the signed probabilities, resulting in a near-zero score for such cases (Table 3.1, Methods, section 3.3.3).

| Score metric | Method | Description |
|---|---|---|
| *diff* | DeepSEA | Difference in the probability of binding |
| *log FC* | DeepSEA | Log fold changes of the binding probability |
| *deltaSVM* | gkmSVM | Difference in the sum of SVM-based *k*-mer weights |
| *GERV score* | GERV | Difference in predicted ChIP-seq reads |
| *delta raw* | PWMs/DeepBind | Maximum difference across each scored window |
| *delta track* | PWMs/DeepBind | Difference between the maximum score for all windows |
| *delta* $P_{bind}$ | PWMs/DeepBind | Maximum difference between probability-transformed scores across each scored window |
| $P_{comb}$ | PWMs/DeepBind | Maximum signed $P_{loss}$ or $P_{gain}$ score |
| $P_{sum}$ | PWMs/DeepBind | Sum of signed $P_{loss}$ and $P_{gain}$ scores |

Table 3.1 Summary of existing and devised scoring metrics used across different methods.

### 3.2.3 The use of allele-specific binding data for benchmarking variant impact prediction

Given the numerous available predictors and scoring metrics available for prioritising the impact of variants on TFBSs, we investigated how well each method and scoring metric performed at distinguishing TFBS-altering variants using the ASB data as a gold standard.

We collected and trained models for TFs with ASB data from the described methods. DeepBind models for 94 TFs were utilised, which we had previously trained on ENCODE ChIP-seq data and SELEX data [68]. For DeepSEA, pre-trained models for 67 TFs were used that were trained on similar datasets to those used for DeepBind, matched by the cell

line from which the training data was obtained. The same data used to train DeepBind models was used to train 94 corresponding gkmSVM models and pre-trained GERV models for 62 TFs were collected, based on ChIP-seq data (Methods, section 3.3.2). We further utilised PWMs for 56 TFs from the JASPAR database along with 92 sets of PWMs based on over-represented motifs derived using MEME-ChIP [308] from the data used to train DeepBind models. For each TF, sequences matching a set of the top five over-enriched motifs were used to construct at most five PWMs. Using the set of PWMs, predictions were generated for the (1) "signif" most significant PWM and (2) "best" the PWM that resulted in highest magnitude variant-impact score (Methods, section 3.3.2).

Each method, model, and scoring metric variants was used to score both ASB and non-ASB data. The resulting scores were used to assess the performance of the predictor at discriminating variants implicated in loss or gain ASB from that of non-ASB variants using the receiver operator curve (ROC) and precision-recall (PR) curve. The AUROC and area under PR curve (AUPRC) are used to provide a quantitative measure of performance, where both metrics provide a different view on performance.

The performance was measured using the definition of an ASB and non-ASB variants as those with a $P_{\text{binomial}} < 0.01$ and $P_{\text{binomial}} > 0.5$, respectively and exhibited at least 10 reads mapped to the reference or alternate allele. We further only considered TFs with at least 20 ASB or non-ASB variants to improve robustness.

### A comparison of variant-impact scoring metrics within methods

We first explore performance of PWM-based scoring metrics. We compared the performance of five scoring metrics used for PWMs in JASPAR and MEME-based PWMs. The performance of scoring metrics in most cases was equivalent to one another in each of the PWM sets, with the exception of *delta raw* which consistently demonstrated poor performance ($p<8.5\times10^{-4}$, Figure 3.3). For instance, in JASPAR PWMs, the average difference between AUROCs between *delta raw* and *delta track* was, on average, 0.06 and 0.07 for gain and loss, respectively and 68% (19/28) of TFs show a 10% increase in *delta track* performance for either loss or gain ASB events (Figure 3.4a). The source of poor performance for *delta raw* can be attributed to the inflation of scores caused by maximising differences over all *k*-mers. High *delta raw* scores do not necessarily indicate a loss or gain of due to the positional independence of PWMs. For instance, a low-scoring wildtype sequence harbouring a variant in a position of importance for the PWM will result in a high *delta raw* score. This effect coupled with taking the maximum over sliding *k*-mer windows results in an inflation of scores, which affects the identification of true negatives (non-ASBs) and true positives (ASBs). These

effects are partially mitigated by metrics such as the *delta track* and probabilistic metrics (Figure 3.3).

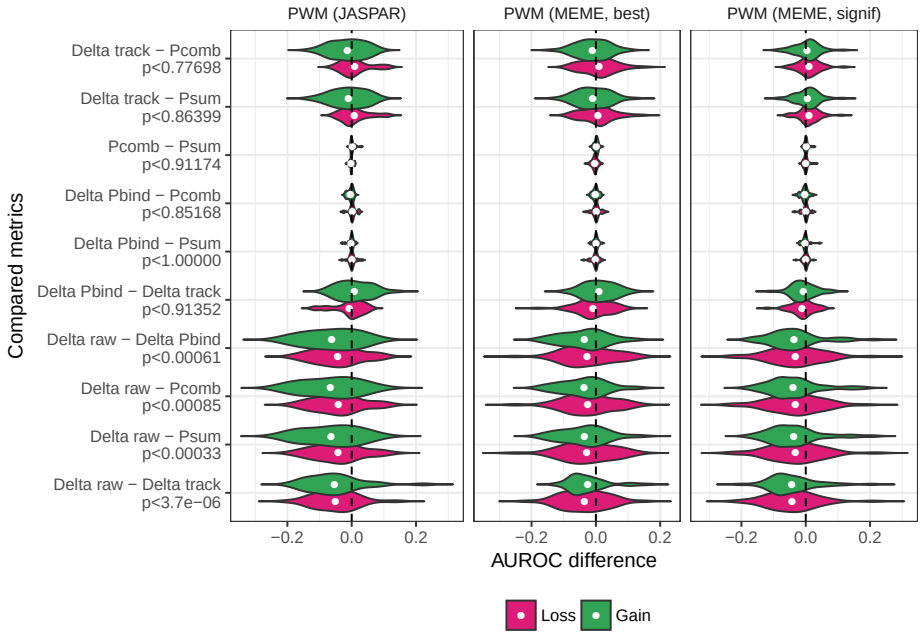



Fig. 3.3 Distribution of differences in AUROCs between PWM scoring metrics. The *p*-value on the y-axis represents a two-sided Wilcoxon test between AUROCs of the compared metrics.

The early B-Cell Factor 1 (EBF1) is one of the TFs with lower performance for *delta raw* compared to the other metrics. For the MEME signif PWM, the *delta track* showed an AUROC of 0.75 and 0.70 for loss and gain, respectively whereas *delta raw* showed AUROCs of 0.66 and 0.61, respectively. Figure 3.4b shows the distributions of scores for loss and gain ASB and non-ASB variants, highlighting the inflation of scores for non-ASB variants.

Detailed examples showing the calculation of *delta raw* and *delta track* for an EBF1 ASB and non-ASB variants are shown in Figure 3.5a-c. Here, the scores for the wildtype and mutant track are shown, along with the difference for each *k*-mer and the final computed scores. The first example highlights a non-ASB variant, where a near-zero predicted score is desired, yet despite no predicted binding occurring on either the wildtype or mutant tracks, the *delta raw* metric still results in an inflated score through differences computed in non-binding regions (Figure 3.5a). The second example highlights a loss ASB variant and the loss event is correctly identified by both metrics. However, the maximum difference for *delta raw* here is obtained not from the *k*-mer exhibiting the loss (*k*-mer window 19), but rather at another *k*-mer (*k*-mer window 14) (Figure 3.5b). The third and final example highlights a gain ASB event that shows how the drawbacks of the *delta raw* metric lead to an incorrect

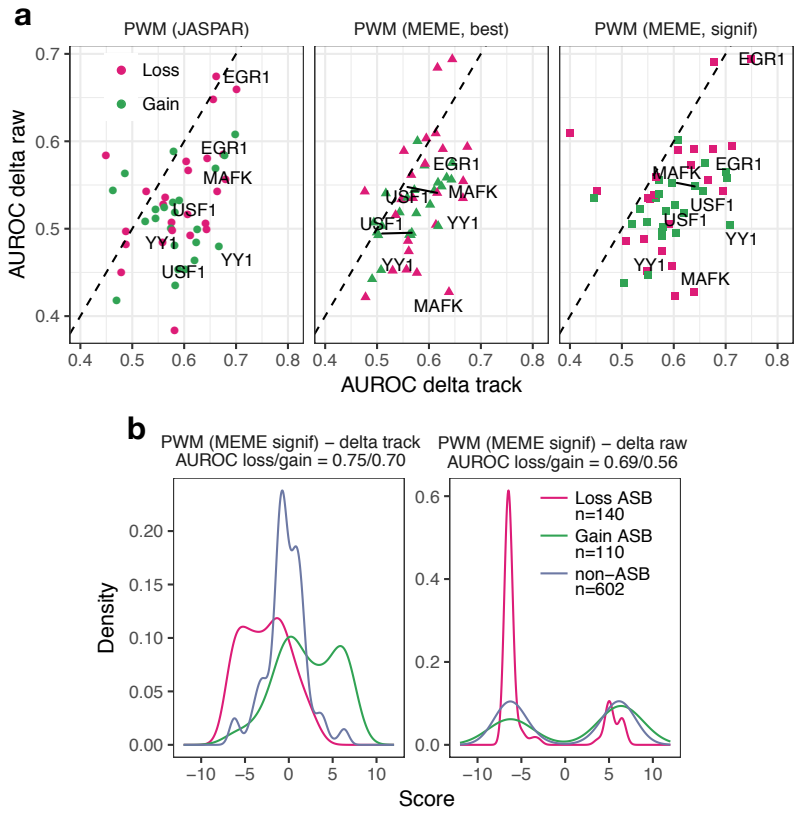



Fig. 3.4 Comparison of *delta raw* and *delta track* metrics for the PWM. (a) AUROCs for *delta track* compared to that of *delta raw* for the three PWM sets. (b) Density plots showing the inflation of scores in the *delta raw* metrics for the EGR1 TF.

prediction of the variant as loss, whereas *delta track* correctly predicts the directionality (Figure 3.5c). The *delta track* and similar metrics offer numerous advantages over identifying the largest possible difference. All metrics will, however, be bottlenecked by the high degree of false positives produced by PWMs.

The performance of scoring metrics used in both DeepBind and DeepSEA were also assessed. We compared the two DeepSEA metrics and found that, overall, neither metric significantly outperformed the other ($p < 0.42$). The *log FC* metric did, however, show an average increase of 0.032 and 0.019 in AUROC for loss and gain, respectively (Figure 3.6a). For DeepBind, no significant difference was observed in performance between the five used metrics. However, *delta raw* did show a modest increase in AUROC over other approaches, with an overall average AUROC difference of as much as 0.021, compared to $P_{comb}$ (Figure 3.6b). Unlike the PWM, the *delta raw* did not overall show difference to that of *delta track* (AUROC difference = 0.0089).

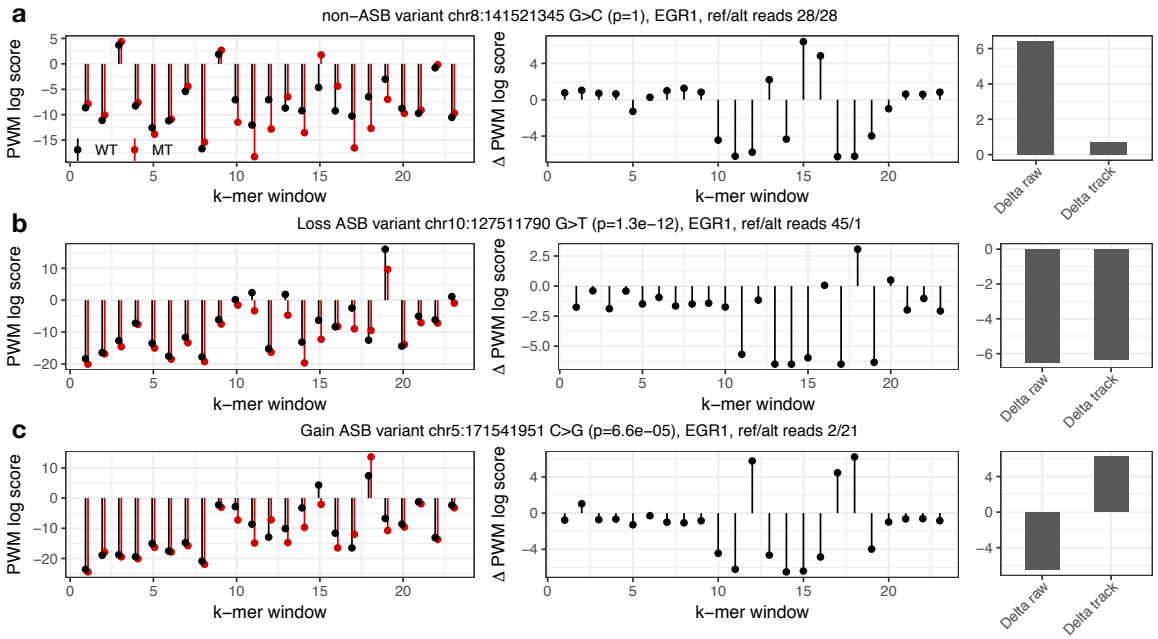



Fig. 3.5 EGR1 examples of scores for individual *k*-mers highlighting the differences between *delta track* and *delta raw*. The left plot shows the wildtype (black) and mutant (red) scores for each *k*-mer, the middle plot shows the score difference, and the right plot shows the final *delta raw* and *delta track* scores. This is shown for (a) a non-ASB that are misclassified by *delta raw*, and correctly classified by *delta track* (b) loss ASB correctly identified by both metrics, and (c) gain ASB misclassified by *delta raw* and correctly identified by *delta track*.

The choice of the scoring metric used in variant impact can often be critical to both interpretability and performance. For PWMs, the *delta raw* metric in PWMs has been long used in studies to quantify effect of a variant of a TFBSs [304, 306, 305, 76]. The results demonstrated here indicate that the choice of score metric when using PWMs, particularly *delta raw*, can drastically impact the reliability of predictions made on regulatory variants. Alternative metrics such as *delta track* and the probabilistic metrics $P_{comb}$ and $P_{sum}$ offer good approaches to mitigating effects by *delta raw* but still are bottlenecked by the inherent limitations of PWMs. For deep learning approaches, little overall difference was observed between metrics and the choice of metric in this case remains purely for interpretation purposes.

### Performance of binding models vary depending on the definition of ASB variants

We then asked whether performance varied if thresholds used to define ASB and non-ASB variants were changed. We measured the AUROC for a combination of thresholds for both the $P_{\text{binomial}}$ ($p < 0.1$, 0.01, $10^{-3}$, $10^{-4}$ and $10^{-5}$) and the minimum number of reference or

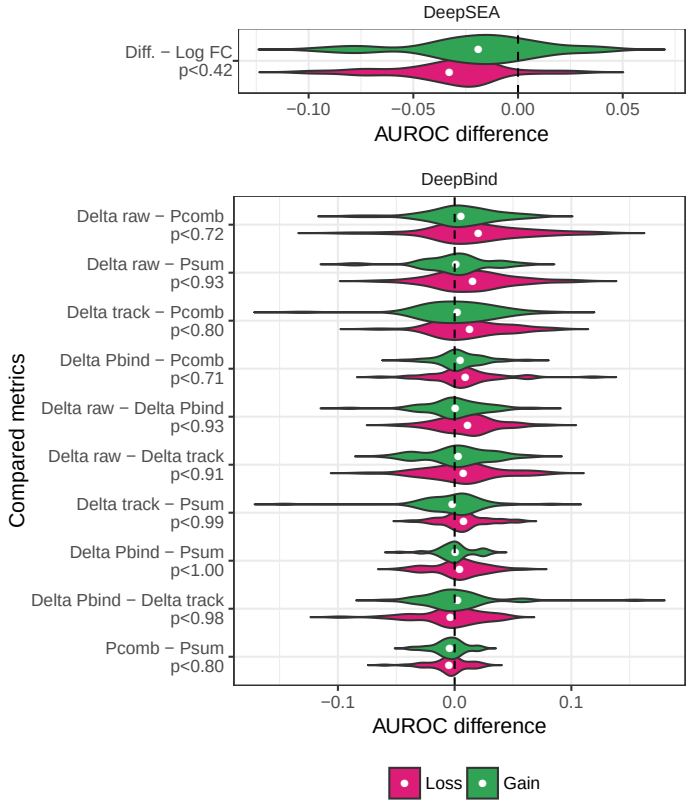



Fig. 3.6 Distribution of differences in AUROCs between (a) DeepSEA and (b) DeepBind scoring metrics. The *p*-value on the y-axis represents a two-sided Wilcoxon test between AUROCs of the compared metrics.

alternate reads ($\geq$10, $\geq$20 and $\geq$30 reads). Performance was measured for seven TFs (BATF, CEBPB, CTCF, EBF1, RUNX3, SMC3, TBP) which had $\geq$20 ASB variants at $10^{-5}$ and $\geq$30 reads.

Utilising performance measures for seven TFs with sufficient data at $\geq$10 reads and $P_{\text{binomial}} < 10^{-5}$ we found that, on average, more stringent definitions of $P_{\text{binomial}}$ thresholds exhibited higher AUROCs, which was consistent across both loss and gain ASBs (Figure 3.7a). For instance, for gain ASBs in DeepBind, at a $\geq$10 reads the average AUROC at $P_{\text{binomial}} < 10^{-5}$ and $P_{\text{binomial}} < 0.10$ is 0.70 and 0.59, respectively. Conversely, increasing the minimum number of reads did not show any substantial shift in performance (Figure 3.7a). These results suggest that models are better able to distinguish variants with a higher imbalance in the number of reads and that higher read imbalance is more likely driven by changes in sequence specificity.

To determine if more stringent thresholds used to define non-ASB variants affected the AUROC, we fixed the ASB $P_{\text{binomial}}$ to 0.01 with $\geq$10 reads and assessed performance at

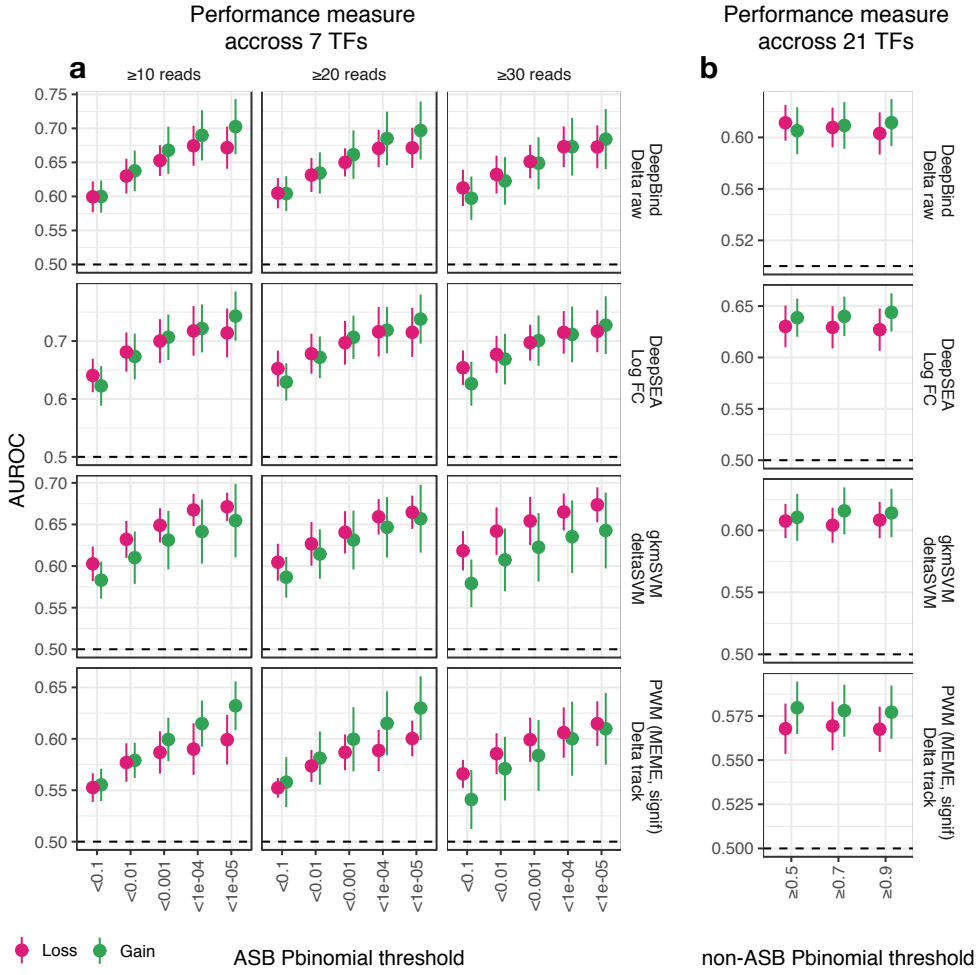



Fig. 3.7 Performance based on different definitions of ASB and non-ASB variants. Performance as measured by the median AUROC for (a) ASB variants across seven TFs where both $P_{\text{binomial}}$ and minimum reads thresholds were varied. (b) Similar performance was measured for 21 TFs at different $P_{\text{binomial}}$ thresholds for non-ASB variants. Magenta and green represent loss and gain, respectively and error bars represent the standard error.

different $P_{\text{binomial}}$ thresholds of 0.5, 0.7, and 0.9 for 21 TFs which had at least $\geq$20 non-ASB variants at $p > 0.9$. However, higher thresholds of $P_{\text{binomial}}$ did not show any significant variation in performance (Figure 3.7b).

At stringent thresholds of $P_{\text{binomial}}$, the number of TFs for which ASB data is available is minute. Thus, to assess performance with a sufficient number of TFs, we retain the thresholds of $P_{\text{binomial}} < 0.01$. Furthermore, since no significant increase was observed at higher reads we retain the $\geq$10 reads for further analyses. Lastly, it is not expected that the performance of gain and loss differ significantly and any changes observed are likely due to the small number of TFs being assessed.

### Machine learning-based methods outperform PWMs at predicting the impact of variants on transcription factor binding

PWMs have been the *de facto* approach to modelling TF specificity and assessing impact of regulatory variants on TF-binding. Machine learning, and in particular deep learning approaches are able to capture more complex relationships and reduce false positive predictions. We, therefore, next asked how the performance of *k*-mer-based and machine-learning approaches compared to that of PWMs at predicting variant impact. For methods with more than one scoring metric, we selected the top performing metric, which included *delta track* for PWMs, *log FC* for DeepSEA and *delta raw* for DeepBind, the *deltaSVM* score from gkmSVM and the *GERV score* from GERV. We compared performance based on the AUROC and AUPRC (Figure 3.8).

Because there exists a different number of trained models with ASB data for each method, we compared performances for 11 TFs with models across all five methods. GERV showed near random performance across both loss and gain, performing poorer than PWMs. The other machine-learning approaches including gkmSVM, DeepBind and DeepSEA significantly outperformed PWMs with respect to AUROCs (Figure 3.8a, $p$=0.034 DeepBind, $p$=$5.91\times10^{-04}$ DeepSEA, $p$=0.038 gkmSVM) and AUPRCs ($p$=$4.57\times10^{-3}$ DeepBind, $p$=$1.24\times10^{-3}$ DeepSEA, $p$=0.024 gkmSVM). We further limited the predictors being compared in order to retain a larger number of common models between the methods. We compared MEME-based PWMs, with gkmSVM and DeepBind for a total of 39 common TFs, where both DeepBind and gkmSVM similarly outperformed the PWM-based models with respect to AUROCs (Figure 3.8b, $p$=$4.07\times10^{-3}$ DeepBind, $p$=$4.85\times10^{-3}$ gkmSVM) and AUPRCs ($p$=$2.32\times10^{-3}$ DeepBind, $p$=$7.31\times10^{-3}$ gkmSVM).

Comparing the AUROC of deep-learning-based methods to that of PWM *delta track*, we identify the TFs SRF, CHD2, IRF4, BATF and CEBPB amongst those where deep learning models perform better at predicting variant impact (Figure 3.8c-d). A more in-depth

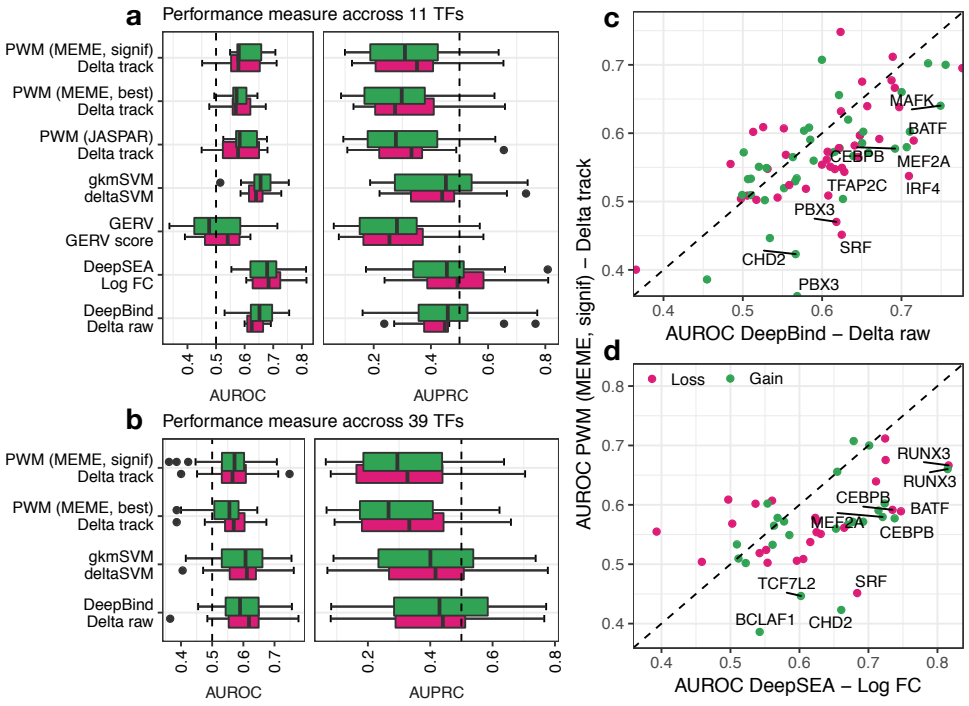



Fig. 3.8 Comparing performance of machine learning approaches to PWMs. (a-b) Comparison of AUROCs (left) and AUPRCs (right). Performance is shown for (a) nine TF models shared amongst five methods and for (b) 31 TF models shared amongst four methods. (c-d) Scatter plots showing the AUROCs for individual TFs for deep learning models (c) DeepBind *delta raw* and (d) DeepSEA *log FC* against PWM *delta track*.

examination of DeepBind and PWM scores reveals that even the best performing PWM metric often results in high numbers of false positives and false negatives. These results further illuminate the importance of machine learning models in variant impact.

### Alternative binding mechanisms explain differences in variant impact prediction performance

Having established that machine learning approaches outperform PWMs, we sought to focus on DeepBind and DeepSEA models and investigate the performance of individual TFs.

Amongst TFs that performed well are RUNX3, BATF, MAFK, and CEBPB, which had an AUROCs of $> 0.7$ in either DeepBind or DeepSEA models. Conversely, TFs like SP1, BRCA1, TBP, and TAF1 consistently showed near-random performance (Figure 3.10a-b). We asked whether poor ASB performance is dictated by the model's performance at identifying binding sites. A model that is unable to correctly identify binding sites should not perform well at identifying the impact of ASB variants. Indeed, we found that models with an AUC $< 0.80$ in DeepBind also demonstrated poor ASB performance (Figure 3.10c-d).

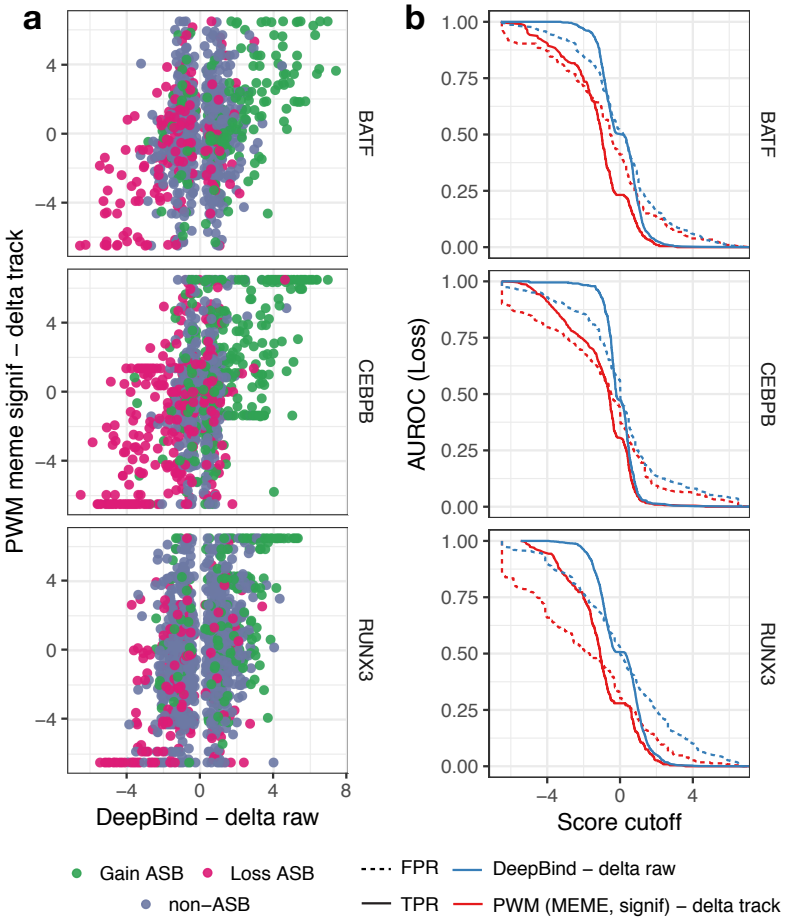



Fig. 3.9 Examples of three TFs showing score differences between machine learning approaches and PWMs. (a) The relationship between DeepBind *delta raw* scores and PWM *delta track* scores for ASB and non-ASB variants highlight the high degree of false positives. (b) True positive rates (blue) and false positive rates (red) computed at different DeepBind *delta raw* and PWM *delta track* thresholds.

However, high-performing models showed a high degree of variation with respect to ASB performance. For instance, SP1 and BRCA1 both have model AUROCs of 0.99. Despite having explicit sequence specificities, such models are unable to detect *in vivo* occupancy differences introduced by variants. This suggests alternate mechanisms beyond simple binding site specificities affecting binding. Since a number of mechanisms have been shown to contribute to TF-binding specificities such as methylation, DNA shape, TF cofactors and regulatory PTMs on the TF, we explored whether such mechanisms are able to explain the poor performance observed.

TFs that are involved in binding complexes can obtain their specificity by the binding of cofactors [74]. We collected known physical TF-TF interactions from the transcription

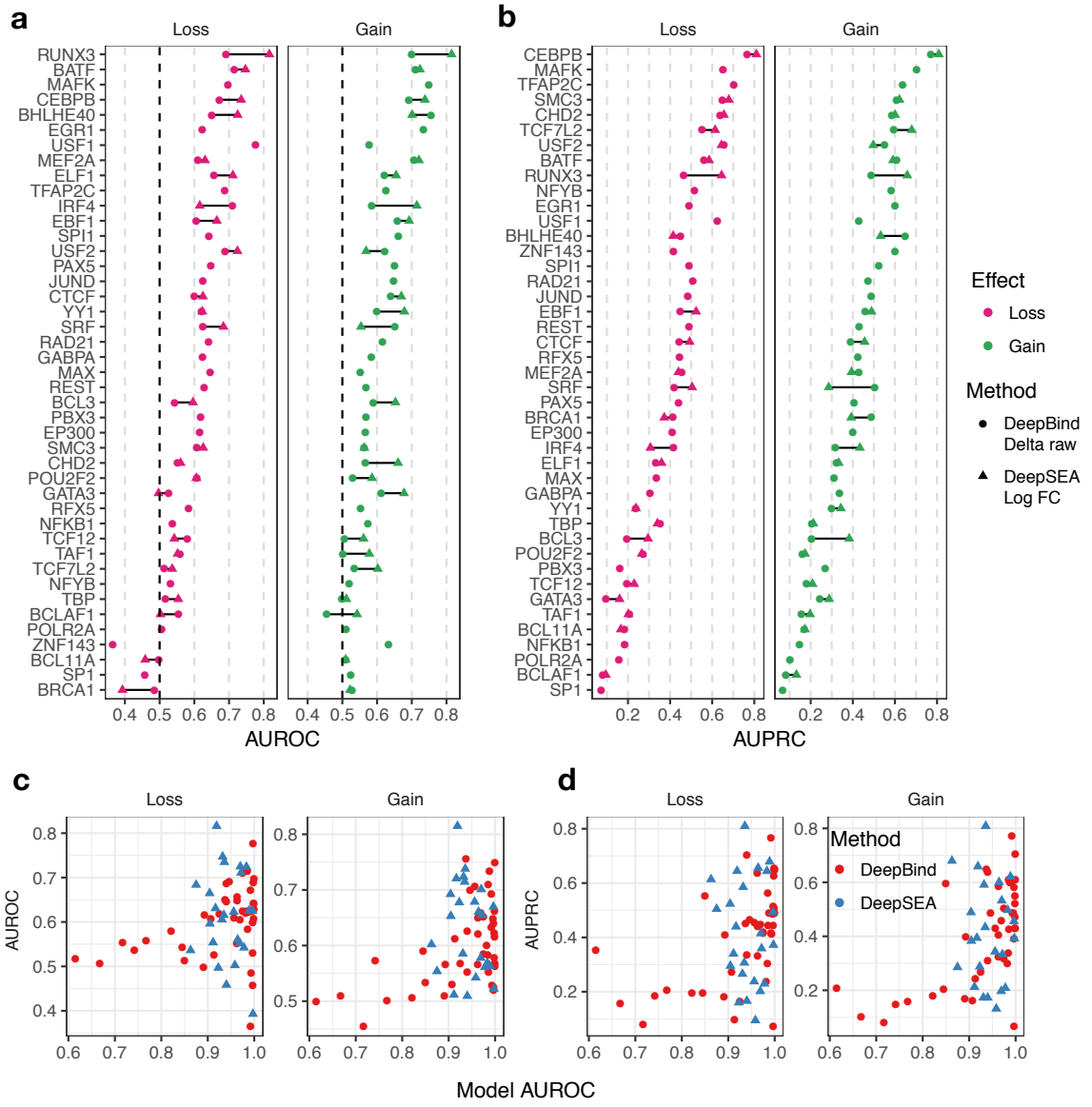



Fig. 3.10 Exploring performance of individual TFs for deep learning methods. (a) AUROCs and (b) AUPRCs for loss (magenta) and gain (green) ASBs. TFs are ordered by the maximum performance metric across methods and effects (c-d) AUROCs of binding performance is compared against performance of models to identify impact of variants, as defined by (c) AUROCs and (d) AUPRCs for DeepBind *delta raw* (red) and DeepSEA *log FC* (blue) models.

cofactors (TcoFs) database [309] for 35 TFs with performance measures and asked whether the degree of interactors predicted ASB performance. We found that TFs such as the TATA-binding protein (TBP) and the specificity protein 1 (SP1) which showed upwards of 50 interactions with other TFs also correlated with poor performance (Figure 3.11a).

Cytosine methylation is another major factor shown to dictate binding for many TFs [72]. A recent study by Yin et al. used a methylation-sensitive derivative of SELEX to identify TFs influenced by methylation for over 500 TFs [72]. Here, TFs were characterised into three groups MethylPlus, where the TF preferred to bind methylated sequences, MethylMinus, where little to no TF binding was found for methylated binding sites and LittleEffect, where methylation had little to no effect on binding. We collected classes for 14 TFs with performance measures across the different methods and compared performance for each class. Interestingly, we found that TFs classified as MethylPlus consistently showed significantly lower performance compared to that of MethylMinus in DeepBind (Figure 3.11b, $p$=$9.9\times10^{-3}$), gkmSVM ($p$=$6.1\times10^{-3}$) and PWMs ($p$=$2.8\times10^{-4}$) (Figure 3.11b). MethylPlus TFs included included SP1, RFX5, POU2F2, and GATA3, which demonstrated low average AUROCs across methods RFX3 (0.53), SP1 (0.49), POU2F2 (0.55), GATA3 (0.59). Indeed, TFs such as RFX3 and SP1 methylation has been shown to positively regulate binding [310, 311]. This suggests that TFs relying on methylation for binding are likely to perform worse when only sequence information is used for model training.

TFs are known to be able to detect three-dimensional shape of DNA [312]. We utilised data from Mathelier et al., where models were trained that incorporated DNA shape features to show performance of binding can be improved [71]. We identify TFs that rely on DNA shape for binding by computing the percent increase in AUROC ($\Delta\%$) for models. The percent increase is binned values into three bins, $< 5$, $6-10$, and $> 10$, which represent minimal improvement, medium improvement and strong improvement. We found that TFs that showed strong improvement had significantly lower performance when compared to those that showed minimal improvement for PWMs ($p$=$4.4\times10^{-3}$) and gkmSVM ($p$=0.039). Significance was borderline significant for DeepBind ($p$=0.061) and DeepSEA ($p$=0.035). This can be explained by the fact that deep learning approaches should at least in part be able to extract DNA shape features from the sequences they are trained on (Figure 3.11c). TFs in which DNA shape aided binding prediction (>10%) included SRF (mean AUROC = 0.62, $\Delta\% = 23$), BRCA1 (mean AUROC = 0.48, $\Delta\% = 15$), NFYB (mean AUROC = 0.50, $\Delta\% = 12.6$), MEF2A (mean AUROC = 0.62, $\Delta\% = 12$), MAFK (mean AUROC = 0.68, $\Delta\% = 11.7$), TBP (mean AUROC = 0.53, $\Delta\% = 11$), PAX5 (mean AUROC = 0.63, $\Delta\% = 10.1$), and SP1 (mean AUROC = 0.50, $\Delta\% = 10.1$). In contrast, TFs where DNA shape played a minimal role in binding prediction included ELF1 (mean AUROC = 0.65, $\Delta\% = 1.8$), TFAP2C (mean AUROC = 0.62, $\Delta\% = 2.6$), BHLHE40 (mean AUROC = 0.69, $\Delta\% = 3.5$), and USF2 (mean AUROC = 0.65, $\Delta\% = 3.8$). This suggests DNA shape as a valuable feature when assessing variant impact on TF-binding.

Finally, we explore the impact of PTMs on TFs binding. PTMs are key regulators of transcriptional activity and are known to govern binding specificity [313]. We asked if TFs that are more likely regulated by PTMs also performed poorly. We collected 1,645 PTM sites for seven modifications in 43 TFs from PhosphoSitePlus [117]. We binned the TFs by the number of PTM sites they harboured by percentiles. In many cases, we found that heavily modified TFs such as BCLAF1 (mean AUROC = 0.49, n = 192), POLR2A (mean AUROC = 0.52, n = 171), EP300 (mean AUROC = 0.59, n = 139) and SMC3 (mean AUROC = 0.57, n = 90) showed significantly lower performance levels, compared to TFs that harboured fewer than 10 PTM sites in DeepBind ($p$=$6.6\times10^{-5}$), DeepSEA ($p$=$5.6\times10^{-5}$), gkmSVM ($p$=$2.1\times10^{-4}$) and PWMs ($p$=$1.6\times10^{-3}$, Figure 3.11d). We similarly utilised 42 PTM sites across 17 TFs known to be regulatory and compared performance of TFs with a single regulatory PTM to those with more than one. We observed a similar trend, where TFs such as BRCA1 (mean AUROC = 0.48, n = 6), SP1 (mean AUROC = 0.50, n = 5), and NFKB1 (mean AUROC = 0.51, n = 4) with a high number of known regulatory PTMs displayed significantly lower performance across DeepBind ($p$=0.011) and gkmSVM ($p$=0.042) Figure 3.11d).

For certain TFs with distinct sequence specificities, elucidating impact of variants can be more challenging due to the sequence specificity depending on a multitude of factors that involve mechanisms beyond proximal sequence information alone. These results demonstrate the importance of such factors, *in silico*, when assessing variant impact in TFBSs.

## 3.3 Methods

### 3.3.1 Collection of allele-specific binding data

ASB data were collected from five studies [106, 104, 295, 105, 296]. In each study, ChIP-seq reads are mapped to both alleles of heterozygous variants in individuals or cell lines. A count for the number of reads mapping to the maternal and paternal allele of each locus is provided by the studies. Allelic read imbalance is computed across all studies using a binomial test:

$$P_{binomial}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.1}$$

where $n$ is the total number of reads mapped at a given loci, $p$ is the probability of success, which is fixed to 0.5. This assesses deviation from the expected 50/50 read count. Finally, SNP positions for hg18-mapped variants are converted to hg19 using `liftOver` [314] and loci that did not map were discarded.

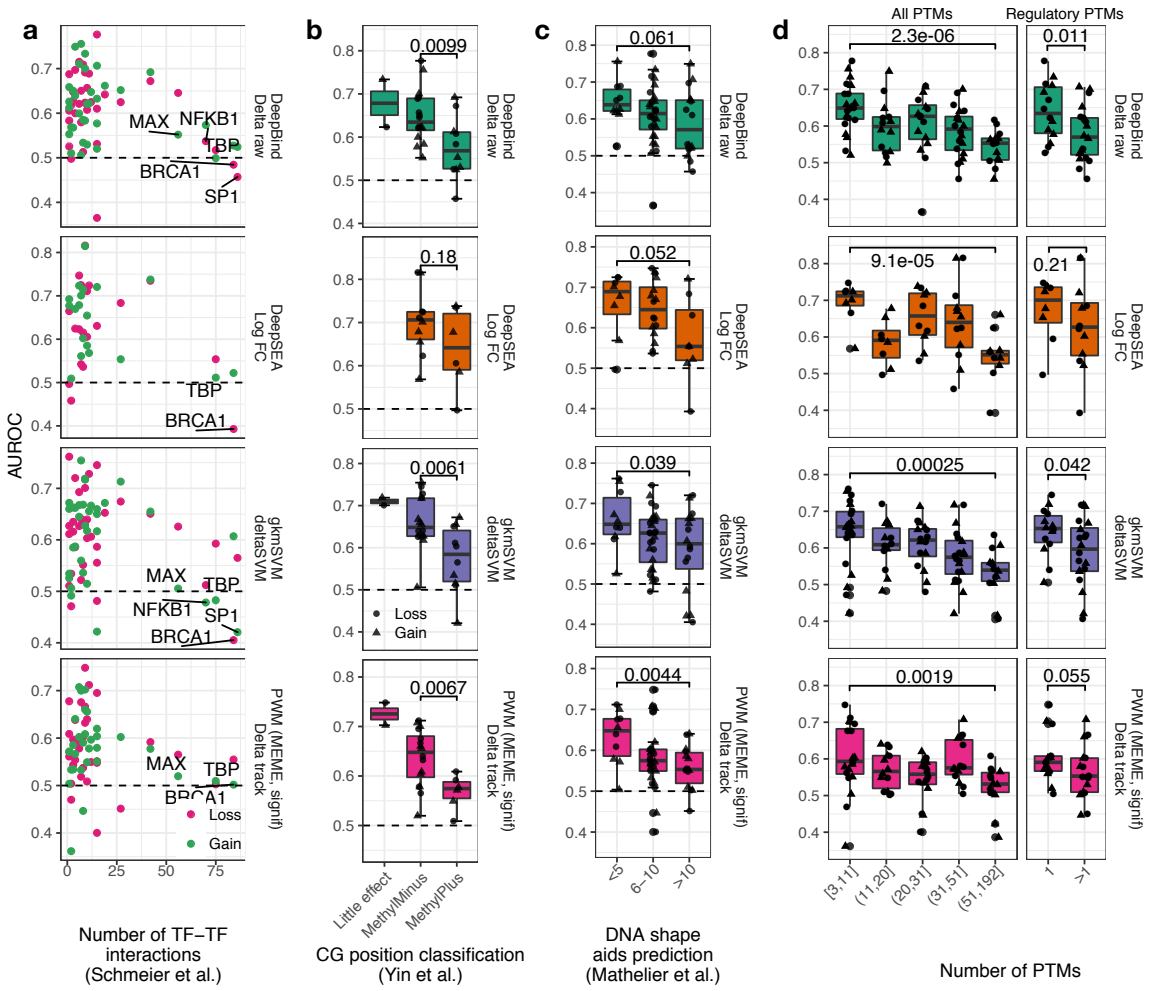



Fig. 3.11 Alternative mechanisms that contribute to poor variant-impact prediction. Performance, as measured by AUROCs across DeepBind, DeepSEA, gkmSVM and PWMs for (a) Number of TF-TF interactions, (b) MethylPlus versus MethylMinus TFs (c) TFs where binding is influenced by DNA shape and (d) PTMs. Significance *p*-values are based on a one-sided Wilcoxon test.

### 3.3.2 Transcription factor binding model training and scoring

DeepBind models for a total of 94 TFs based on SELEX and ChIP-seq datasets were obtained from Alipanahi et al. [68]. Performance of each model was evaluated by applying models to left out test sequences (sequences not used to train the model) and random genomic regions. In the cases where there were multiple DeepBind models per TF, the model with the highest performance was selected. Scoring was carried out using the `deepbind` executable v0.11 with default parameters. Scores for DeepSEA were obtained through the online web server http://deepsea.princeton.edu/. Models for a total of 62 TFs were used that matched DeepBind models by cell line.

A total of 54 PFMs for 54 TFs with a DeepBind model were collected from JASPAR [287]. If a TF has more than one model, the model with the latest accession version is used. Motif enrichment data carried out using MEME-ChIP [308] on ChIP-seq data used to train DeepBind models was used to construct a second set of PWMs. Each contained a set of enriched motifs along with matching ChIP-seq sequences and an *e*-value reflecting the enrichment significance. Motifs with an *e*-value > 0.05 or less than 10 associated sequences were discarded and the sequences associated with the top five enriched motifs were used to construct PWMs. The "signif" PWM set was defined as the PWM for each TF with the most significant *e*-value, whereas in the "best" PWM set the top five most significant PWM was used for scoring and the PWM that gave off the highest variant effect prediction was used. All PWMs were constructed using the `toPWM` function of the `TFBStools` package [315] and the `PWMscoreStartingAt` function of the `Biostrings` package was used to score sequences using the generated PWMs [316].

The gkmtrain command from the LS-GKM library (https://github.com/Dongwon-Lee/lsgkm) was used to train gkmSVM models [103] with default parameters, except for word length option "-l" set to 10. ChIP-seq and SELEX sequences were used as positive sequences, and random genomic sequences with the same length were used as negative sequences. The *deltaSVM* scores were generated from using the `gkmpredict` command along with the `deltasvm.pl` script (http://www.beerlab.org/deltasvm/). Finally, pre-trained GERV models for a total of 60 TFs were obtained from http://gerv.csail.mit.edu/ ChIP-seq experiments from ENCODE project and the `preprocess` and `score` options of the `run.r` script with default options was used to score the impact of variants.

### 3.3.3 Variant impact scoring metrics

#### DeepBind and PWMs

Given a PWM or DeepBind model, we define a window of size $k$ using the width of the PWM or the detector length of the DeepBind model (see [68]), respectively. Given a variant at position $q$, we score both the wildtype and mutant sequences starting $q-k$ to $q+k$ at increments of $k$ for a set of raw wildtype scores $w_1, w_2, \ldots, w_k$ and mutant scores $m_1, \ldots, m_k$. Given the set of indices $S = 1, \ldots, k$, the *delta raw* ($\Delta R$) and *delta track* ($\Delta T$) metrics are computed as follows:

$$\begin{aligned} i^* &= \underset{|m_i - w_i| \forall i \in S}{\operatorname{argmax}} \\ \Delta R &= m_{i^*} - w_{i^*} \end{aligned} \tag{3.2}$$

$$\Delta T = \max_{\forall i \in S} m_i - \max_{\forall i \in S} w_i \tag{3.3}$$

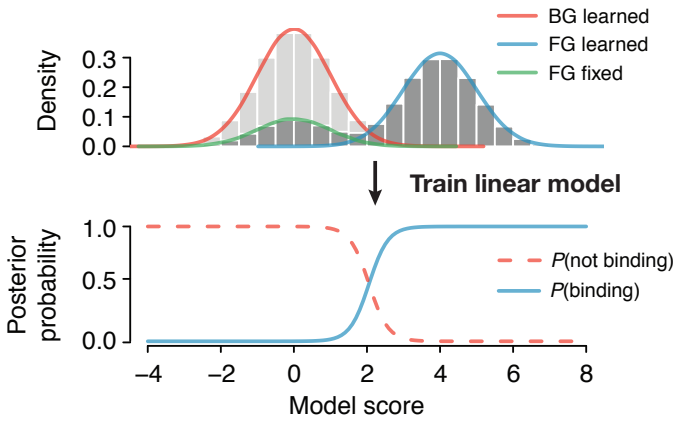



Fig. 3.12 The conversion of raw scores into probabilities of binding using generalised linear models.

To compute $P_{bind}$ scores we score a given an individual raw score, we compute a foreground and background distribution of raw scores using a set of positive and negative sequence respectively. The negative sequences are defined as 10,000 randomly sampled genomic sequences of size $k$. The positive sequences for JASPAR PWMs are defined as generated sequences from the PWM, whereas for MEME-ChIP PWMs this is defined as the corresponding matching ChIP-seq sequences used to construct the PWM. The positive sequences as the ChIP-seq or SELEX sequences used to train the DeepBind models. We assume the background distribution follows a Gaussian distribution $N \sim \mathcal{N}(\mu_n, \sigma_n)$ and learn the parameters of the true positive distributions by fitting a two-component Gaussian model mixture model. Here, one component is fixed to $\mu_n$, $\sigma_n$ and the true positive parameters are learned as $\mu_p$, $\sigma_p$. A total of 10,000 random samples are generated using the given the parameters of background and used to train a generalised linear model, which is used to compute a posterior probability (Figure 3.12) of binding ($P_{bind}$) and not binding ($P_{not\ binding}$). The $P_{bind}$ scores are computed for both the wildtype $Pb^w_1, \ldots, Pb^w_k$ and mutant $Pb^m_1, \ldots, Pb^m_k$ $k$-mers. The $P_{not\ binding}$ scores are computed similarly as $Pn^w_1, \ldots, Pn^w_k$ and $Pn^m_1, \ldots, Pn^m_k$. The delta $P_{bind}$ ($\Delta P$) score is then computed similarly to that of $\Delta R$:

$$\begin{aligned} i^* &= \operatorname*{argmax}_{|Pb^m_i - Pb^w_i| \forall i \in S} \\ \Delta P &= Pb^m_{i^*} - Pb^w_{i^*} \end{aligned} \tag{3.4}$$

The probabilistic scores of a loss ($P_{loss}$) and gain ($P_{gain}$) events are computed by multiplying the likelihood of the wildtype allele binding and the mutant allele not binding for loss, and vice versa for gain:

$$P_{loss} = \underset{Pb_i^w \cdot Pn_i^m \forall i \in S}{\operatorname{argmax}} Pb_i^w \cdot Pn_i^m \tag{3.5}$$

$$P_{gain} = \underset{Pb_i^m \cdot Pn_i^w \forall i \in S}{\operatorname{argmax}} Pb_i^m \cdot Pn_i^w \tag{3.6}$$

Both probabilities are then combined into individual scores $P_{sum}$ and $P_{comb}$ as follows:

$$P_{sum} = -P_{loss} + P_{gain} \tag{3.7}$$

$$P_{comb} = \begin{cases} P_{gain}, & \text{if } P_{gain} > P_{loss} \\ -P_{loss}, & \text{otherwise} \end{cases} \tag{3.8}$$

**DeepSEA**

Given a probability of binding in the wildtype and mutant alleles, as $\rho_w$ and $\rho_m$ respectively, DeepSEA utilises two scoring schemes: the difference ($DS_D$) and log fold change ($DS_L$) computed as follows:

$$DS_D = \rho_m - \rho_w \tag{3.9}$$

$$DS_L = log(\frac{\rho_m}{1-\rho_m}) - log(\frac{\rho_w}{1-\rho_w}) \tag{3.10}$$

**gkmSVM**

Given $k$-mer weights computed for wildtype and mutant sequences flanking the variant position as $\omega_1^w, \ldots, \omega_{10}^w$ and $\omega_1^m, \ldots, \omega_{10}^m$, the deltaSVM ($\Delta SVM$) score is computed as follows:

$$\Delta SVM = \sum_{i}^{n=10} \omega_i^m - \omega_i^w \tag{3.11}$$

### 3.3.4 Allele frequencies and non-coding variant impact predictions

Allele frequencies for the 1000 genomes project, ExAC variants and ESP6500 project, along with non-coding variant impact predictions for CADD, Eigen and GWAVA were obtained from the ANNOVAR tool [317], using the `table_annovar.pl` script.

### 3.3.5 Performance measures

ROC and PR curves were generated by assessing the TPR (or recall), FPR and precision. The ROC curves compare the FPR against the TPR, whereas PR curves compare the TPR against the PPV (or precision). Given the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), these are computed as follows:

$$\begin{aligned} TPR &= \frac{TP}{TP+FN} \\ FPR &= \frac{FP}{FP+TN} \\ PPV &= \frac{TP}{TP+FP} \end{aligned} \tag{3.12}$$

All ROC and PR curves, along with area under the curve measures were computed using the PRROC R package [318].

## 3.4 Discussion

Non-coding variation has the ability to greatly alter gene expression and influence disease phenotypes. Understanding the impact of non-coding variation is an ongoing challenge in genetics. One of the primary modes for this is through impacting TFBSs. Yet, despite the wealth of TF specificity available through high throughput technologies, accurate *in silico* prediction of TFBS-altering variants remains a non-trivial task.

This chapter describes efforts to compare TF-based variant impact predictors using ASB variants as a gold standard. Since both alleles exist in the same cellular environment, ASB variants serve as a valuable source to assess performance of TF-binding models at assessing impact of variants. We have shown that the ability for machine learning models, in particular, deep learning methods, to significantly reduce the number of false positives allows for more accurate variant impact predictions. Deep learning approaches are able to utilise the full extent of ChIP-seq and SELEX data to learn far more complex positional dependencies in binding sites. Deep learning approaches are also not confined to the exact motif location and therefore can model sequence context of the binding site, which has been shown to contribute to binding [64, 319].

We finally show that TFs with poor performance at assessing variant impact often rely on additional mechanisms such as binding partners, methylation, DNA shape and PTMs (Figure 3.11).

Assessing TF-DNA binding and how it is influenced by genetic variation *in silico* is a much more complex process than once thought. Current methods available for interpreting effects of TFBS variants rely primarily on binding specificity. Although this provides a useful framework for prioritizing non-coding variants, as demonstrated by results, even the most sophisticated methods are often unable to capture the full extent of altered binding in the genome. This can be attributed to several reasons. First, there are several other mechanisms that have been known to significantly contribute to binding specificity such as epigenetic modifications, cooperative binding, geometric shapes of DNA, PTM modifications and more. Epigenetics, in particular, methylation, can play a major role in enhancing or inhibiting TF-binding [320, 72]. The recent study by Yin et al. carried out methylation-sensitive SELEX in 542 TFs and identified many methylation-dependent TFs [72]. Epigenetics can also greatly affect regions TFs can occupy. Nucleosome occupancy, for instance, results in closed chromatin which is inaccessible to TFs. Indeed, DNAase hypersensitivity sequencing (DHS-seq) has revealed regions of open chromatin in many cell lines, which have been shown to improve binding prediction [321]. PTMs is another major regulator of TF activity through altering its structural conformation, stability or sub-cellular localization thereby affecting binding [313]. For instance, phosphorylation of p53 on S378 allows it be recognized by 14-3-3 proteins, which associate with p53 and significantly enhances DNA-binding [322]. Second, binding preferences of a TFs have been shown to be heterogeneous across different cell lines. For instance, Arvey et al. comprehensively analysed ChIP-seq data for 67 TFs across multiple different and found that many cell-type-specific sequence models were able to capture binding variability, which was primarily due to differences in heteromeric complex formations [323]. Since the samples and cell lines from which ASB variants were obtained do not always match that of the experiments used to generate TF-specificity models, this is a potential confounding factor of poor performing models.

Another factor greatly limiting the prediction of TFBS-altering variants is the availability of TF motifs. It is estimated that the human genome contains approximately 1,400 TFs containing DBDs [324]. Although the current catalogue of TF-binding specificity has significantly expanded in the past decade with the aid of high throughput approaches such as ChIP-seq, SELEX and PBMs, almost half of identified TFs are yet to have their specificity determined [325]. This is perhaps due to technical limitations, such as transient binding or expression of the TF. The lack of such data further hampers ability for us to systematically understand variant impact in TFBSs.

Significant advances in interpreting non-coding variation have been greatly aided by the emergence of deep learning methods to the field of genetics over the past few years. However, accurate assessment of variant impact on TFBSs will require models to systematically

integrate additional epigenetic, proteomic and genetic data in order to account for mechanisms beyond sequence specificity in a cell-type-specific manner.

# Chapter 4

# Functional consequences of single nucleotide variants across different molecular features

*In this chapter, I describe a comprehensive effort to compile and benchmark commonly-used sequence and structure-based predictors of mutational consequences, which are used to precompute the effects of coding and non-coding variants in the reference genomes of H. sapiens, S. cerevisiae, and E. coli. I utilise this data to analyse known pathogenic disease variants and provide mechanistic hypotheses for causal variants of unknown function. Lab members Danish Memon and Marco Galardini have contributed multiple sequence alignments used for SIFT and FoldX-generated data to the project for H. sapiens and E. coli. All work was otherwise carried out by myself under the supervision of Pedro Beltao. Parts of this work were published in the following article:*

Marco Galardini, Alexandra Koumoutsi, Lucia Herrera-Dominguez, Juan Antonio Cordero Varela, Anja Telzerow, Omar Wagih, Morgane Wartel, Olivier Clermont, Erick Denamur, Athanasios Typas and Pedro Beltrao (2017). Phenotype prediction in an Escherichia coli strain panel. *Biorxiv*, page 141879.

## 4.1 Introduction

GWASs have come a long way at identifying causal genetic variants. Over the past decade, thousands of associations have been made between genetic variation and phenotypic traits including disease risk [326]. However, GWASs are inherently limited in their ability to

explain the underlying mechanism that is likely influenced by the variant in question. This missing mechanistic layer poses several roadblocks in the comprehensive understanding of how variants influence phenotypic variability.

In previous chapters, I have discussed the importance of modelling sequence specificity mediating both kinase-substrate phosphorylation and TF-binding and their role in uncovering mechanistic consequences of genetic variation. In addition to such mechanisms, variants occurring in coding and noncoding regions can influence a diversity of molecular functions. For instance, non-coding variants can affect chromatin accessibility [327], splice sites [328], and epigenetic modifications [329]. Coding variants can affect PTM sites [121, 146], protein folding and stability [330], protein interaction interfaces [226], sub-cellular localization [331], as well as introduce premature stop codons. Understanding the disrupted biological mechanisms underlying genetic variation is key to many applications in genetics such as genetically engineering organisms, assessing drug efficacy and drug discovery [332–334].

The ability to predict the degree to which genetic variation would alter such mechanisms offers a time and cost-effective alternative over classical experimental validations and can greatly facilitate the understanding of mechanisms underlying causal variants. A multitude of *in silico* predictors aimed at predicting such effects have been proposed [146, 189, 305], yet, for the average user, they are often cumbersome to set up and use and/or require significant computational power and time. For instance, structure-based protein stability predictors can take on the order of minutes to hours to assess the impact of a single variant [189]. Furthermore, the currently available tools do not comprehensively combine effects across different molecular mechanisms [335] or are limited to analysing coding or noncoding variation [305].

In this chapter, I first investigated natural and disease-variation in both *H. sapiens* and *S. cerevisiae* in the context of functional elements in the genome to show that such regions display higher evolutionary constraint. Accordingly, I have compiled and benchmarked commonly-used sequence and structure-based predictors of mutational consequences and predicted the effect of all possible amino acid and nucleotide variants in the reference genomes of *H. sapiens*, *S. cerevisiae*, and *E. coli*. The impact of variants was measured in the context of conserved protein regions, protein stability, PPI interfaces, PTMs, kinase-substrate interactions, SLiMs, start and stop codons, and TFBSs. This data is deposited in the mutfunc platform, which allows for prioritization of variants while providing insight into the altered mechanisms.

Because all data is precomputed, variants can be rapidly annotated and prioritised. The data available in mutfunc was validated by analysing both natural and disease genetic variation data in *S. cerevisiae* and *H. sapiens* data. For instance, we have shown that genes deemed

essential for survival are less likely to harbour impactful variants and that common genetic variants are less likely to be impactful. Variants curated to be relevant for function or deemed pathogenic were also more likely to be predicted as impactful. We further demonstrate the utility of mutfunc by analysing clinical variants that identified in disease patients but have an uncertain significance of pathogenicity. Mutfunc is a valuable resource that will facilitate the understanding of the mechanistic impacts of genetic variation.

## 4.2 Results

### 4.2.1 Functional genomic regions display evolutionary constraint across *S. cerevisiae* and *H. sapiens*

We first aimed to investigate how both natural and disease variants manifest themselves within functional regions of the genome. We asked if certain functional regions relevant to protein structures, PTMs and TFBSs were under evolutionary constraint or negative selection. If these regions are indeed critical for function, arising deleterious variants would be purged over time in order to retain function. The evolutionary constraint, defined as $c$, can be measured by taking the ratio between observed mutation counts in a region of interest and random regions, where, values below 1 confer negative selection, and values above 1 confer positive selection. To do this, publicly available genetic variation data for both *S. cerevisiae* and *H. sapiens* were used. For *S. cerevisiae*, 896,772 natural variants for over 400 *S. cerevisiae* strains were collated from four studies [336–339], of which 478,857 were coding variants. For *H. sapiens*, over 3.2M coding variants from over 65,000 individuals were obtained from the ExAC consortium [298] (Methods, section 4.3.1).

#### Buried protein regions and interaction interfaces exhibit negative selection

It is well established that buried residues contribute more to protein stability relative to surface residues [340] and have been shown on a smaller scale to harbour fewer mutations [341]. To assess this on a larger scale, variants from *S. cerevisiae* and *H. sapiens* were mapped to a total of 9,837 resolved protein structures and homology models (n=6,737 *H. sapiens*, n=3,100 *S. cerevisiae*) and residues were grouped based on the computed relative surface accessibilities (RSAs) of residues (Methods, section 4.3.4). Residues were binned into four RSA groups (0-25%, 26-50%, 51-75% and >76%). The number of mutations falling within positions of each bin is then counted and compared to counts observed by the same number of random positions in the protein, permuted 1,000 times. Buried residues were

found to harbour fewer variants in both *S. cerevisiae* and *H. sapiens*, relative to those of higher RSA (Figure 4.1a, $p<1.28\times10^{-34}$).

This was similarly measured in structures of PPI interfaces. A total of 9,883 structures (n=7,693 *H. sapiens*, n=2,190 *S. cerevisiae*) for binary PPIs were collected from Interactome3D [205]. The difference in RSA between, $\Delta RSA$, between monomeric proteins and proteins in the interaction complex, was used as a measure of how buried or exposed a residue was within an interface (Methods, section 4.3.4). Residues were grouped into four $\Delta RSA$ bins, similar to those defined for stability. Residues buried within interface residues were found to harbour far fewer variants compared to those that are exposed, with a low $\Delta RSA$ (Figure 4.1b, $p<2.28\times10^{-33}$).

The strong negative selection observed in buried residues with respect to both monomeric proteins and interaction interfaces suggest that they are of higher functional relevance, which is in agreement with what is reported regarding the importance of buried residues in the stability of proteins [340] and interaction interfaces [342].

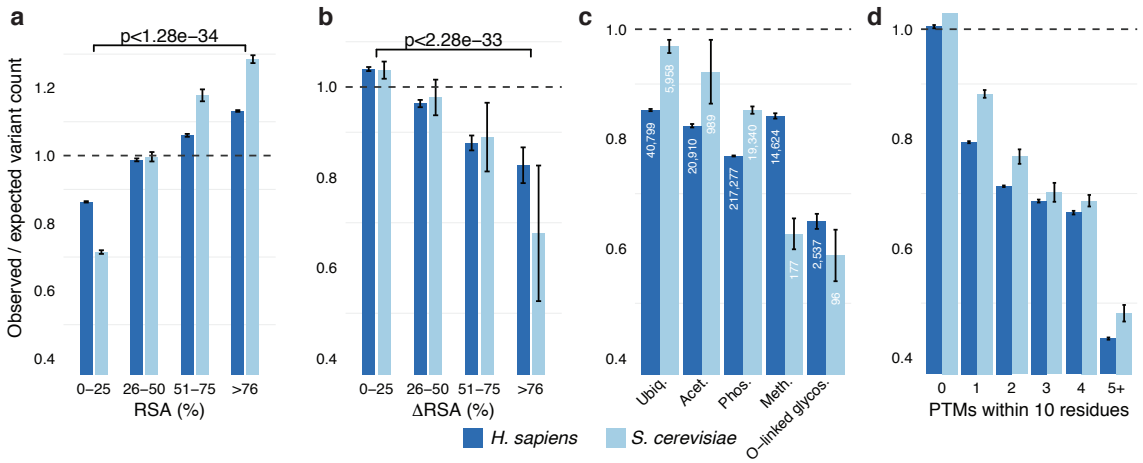



Fig. 4.1 The evolutionary constraint in monomeric protein structures, interaction interfaces and PTMs. (a) Regions buried within a protein structure with a low RSA typically exhibit higher evolutionary constraint. Similarly, (b) regions buried within interaction interfaces exhibit a high $\Delta RSA$ and demonstrate stronger evolutionary constraint. *P*-values represent a one-sided Wilcoxon test. (c) Evolutionary constraint on PTMs, where numbers reflect the number of PTM sites for each modification. (d) PTMs with a higher number of neighbouring PTMs are much under stronger constraint, compared to those that exist individually.

### Post-translational modification regions exhibit negative selection

To explore PTM-associated variation, a total of 296,147 and 26,560 *H. sapiens* and *S. cerevisiae* PTM sites were gathered from publicly available databases and the common PTMs

included ubiquitination, acetylation, phosphorylation, methylation and *O*-linked glycosilation (Methods, section 4.3.5). The frequency of variant frequency within $\pm 5$ flanking residues of modified sites was compared to that of randomly sampled residues. The random set is defined as non-PTM matching residues sampled from the same proteins harbouring the modifications. Different PTMs exhibited variable levels of constraint with modifications like *O*-linked glycosilation displaying the strongest (Figure 4.1c, c=0.64 *H. sapiens*, c=0.58 *S. cerevisiae*), followed by methylation (c=0.84 *H. sapiens*, c=0.62 *S. cerevisiae*). In contrast, modifications such as ubiquitination demonstrated lower constraint (c=0.85 *H. sapiens*, c=0.97 *S. cerevisiae*). This could be partly explained by the fact that ubiquitination is far more robust to genetic variation. There have been numerous documented cases in the literature suggesting that the disruption of one ubiquitination site has little impact on the targeting and degradation of the protein since the ubiquitination of a proximal lysine will often achieve a similar function [343, 344].

Clusters of PTM sites, where a high number of PTM modifications are observed, have been shown to confer functionally relevant regions of the protein [283]. An example of this is the multi-phosphorylation cascade in beta-catenin, where the phosphorylation of several neighbouring phosphosites must be achieved in order for beta-catenin to be targeted for degradation [345]. To test whether clusters of PTM sites confer stronger evolutionary constraint, PTM sites were binned depending on the number of neighbouring PTMs within a $\pm 10$ window and the variants were analysed within each bin. A strong positive relationship was observed between the number of neighbouring PTMs and negative selection. Sites with 5 or more other neighbouring PTMs demonstrated much stronger negative selection when compared to those that occurred individually (Figure 4.1d, $p<5.41\times10^{-6}$). This suggests that such clusters could indeed represent signalling hotspots, relevant for carrying out critical biological functions.

### Transcription factor binding sites exhibit negative selection

We next sought to measure the constraint of non-coding variation in TFBSs in the *S. cerevisiae* genome. To do this, we predicted genes likely regulated by a TF by identifying differentially expressed genes in TF-knock-out strains (Methods, section 4.3.7). This resulted in a network of 1,711 TF-gene relationships across 93 TFs. Using a total of 176 PWMs from JASPAR [346] the corresponding ChIP-seq or ChIP-chip regions in promoters of associated genes were scanned. A total of 4,523 potential binding sites were identified across 93 TFs. The constraint was measured by computing the ratio of variant frequency within the predicted binding sites to that of random genomic sites of the same length and within the same gene promoter and ChIP regions. Interestingly, TFs that were relevant for regulating genes in

response to lack of nutrients (PHD1, SFP1) and oxidative/osmotic stresses (SKN7, MOT3) were not found to be under negative selection. This is in contrast to TFs that regulate more basic cellular function such as those involved in the expression of respiratory genes (HAP4), termination of RNA polymerase I transcription (REB1), and general transcription activation, repression and chromatin silencing (RAP1), which exhibited a much higher degree of negative selection (Figure 4.2a).

Clusters of multiple TFs (heterotypic clusters) or single TFs (homotypic clusters) are have been shown to have increased importance in the regulation of gene expression, likely through promoting cooperative TF binding [347, 348, 348]. To assess evolutionary constraint within heterotypic and homotypic TFBS clusters, predicted binding sites were binned based on the number of neighbouring TFBSs within a 50 bp window and constraint was measured for each bin, similar to PTMs. Clusters of TFBSs displayed high negative selection (Figure 4.2b), where binding sites with 6 or more adjacent neighbours showing significantly stronger negative selection (c=0.37) compared to those that occurred individually (Figure 4.2b, c=0.77, $p$=$1.26\times10^{-34}$), suggesting that regions harbouring a higher number of TFBSs likely represent functional regulatory hotspots.

Given that TFBSs were overall constrained, we next asked if positions relevant for binding were under stronger negative selection. Using the PWMs, the position-specific information content (IC) was computed and used as a proxy for binding relevance. For TFs with greater than 20 putative binding sites, the IC of positions was binned based on whether it was low (<0.5), medium (0.5-1.5) and high (>1.5). Positions with high IC were found to display significantly stronger negative selection compared to those with low IC ((Figure 4.2d), $p$=0.017). Examining position-specific constraint for individual TFs further demonstrates the relevance of high IC positions (Figure 4.2e).

### 4.2.2 mutfunc: a one-stop resource for mechanistic effects of single nucleotide variants

Given that functionally-relevant regions of the genome are under negative selection, we sought to better understand the mechanistic impact of point mutations affecting these functional elements. To do this, a set of commonly-used predictors were used to assess the impact of every possible single amino acid or nucleotide substitution across *H. sapiens*, *S. cerevisiae*, and *E. coli*, where applicable. We precomputed data of variants that impact conserved protein regions, protein stability, protein interaction interfaces, kinase-substrate phosphorylation and other PTMs, linear motifs, TFBSs and start and stop codons (Methods, section 4.3.8). These

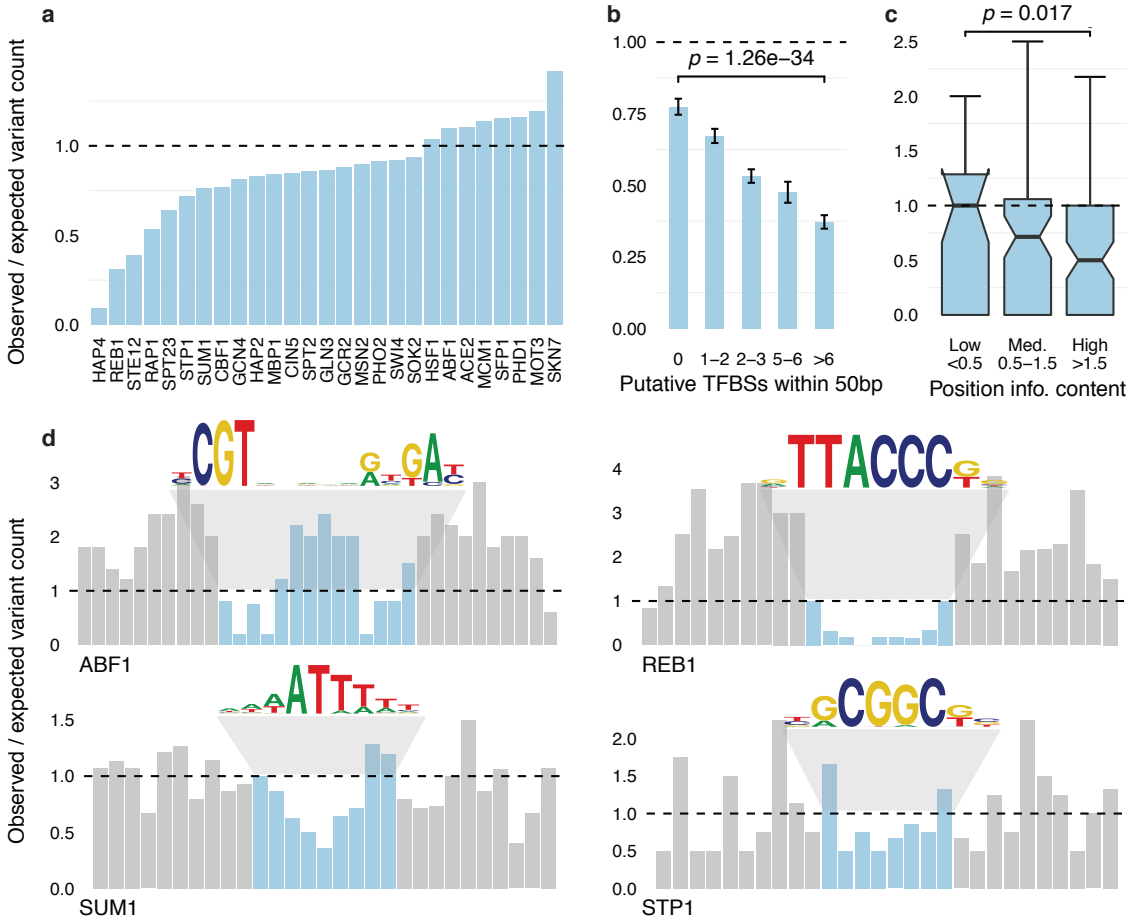



Fig. 4.2 Evolutionary constraint of TFBSs in *S. cerevisiae*. (a) Variability in constraint amongst bindings sites for TFs with at least 40 sites. (b) TFBSs that co-exist with other binding sites are under stronger constraint. *P*-value shown is computed using a one-sided Wilcoxon test (c) Position-specific constraint shows that positions of higher relevance for binding in TFs with at least 20 sites are under stronger constraint. *P*-value shown is computed using a one-sided Kolmogorov-Smirnov test. The clear correlation between the positions relevant for binding and constraint is visually represented through (d) four examples where the bar plots reflect the position-specific constraint in (blue) and around (grey) the binding site, along with sequence logos for the binding specificities.

results were deposited in a web tool, mutfunc, which offers a quick and interactive way by which users can gain mechanistic insight into variants of interest.

### A compiled resource of precomputed mechanistic variant impact

To measure the impact on conserved regions, we constructed 29,027 multiple sequence alignments for proteins of the three organisms (n=19,497 *H. sapiens*, n=5,498 *S. cerevisiae*,

n=4,032 *E. coli*), and used the SIFT algorithm [244] to assess the impact of all possible 291.7M variants (n=212.2M *H. sapiens*, n=53.4M yeast, n=26.1M *E. coli*). To measure the impact on protein stability, the FoldX algorithm [189] was applied to 17,893 structures (including homology models) across the three organisms, and precomputed effects of 66.3 million all substitutions (n=42.7M *H. sapiens*, n=10.3M *S. cerevisiae*, n=13.4M *E. coli*, Methods, section 4.3.4). We identified interface residues in 10,675 structures of binary PPIs from Interactome3D across the three organisms and similarly applied FoldX to compute the effects of all 11.2M possible mutations on binding stability (n=7.2M *H. sapiens*, n=2.3M *S. cerevisiae*, n=1.6M *E. coli*). To identify variants that could impact kinase-substrate sites, we used MIMP [146] to predict the impact of all possible 541,161 variants (n=485,736 *H. sapiens*, n=55,425 *S. cerevisiae*) falling within $\pm$5 residues of a known kinase-substrate phosphorylation site (phosphosite) on a kinase's specificity. Specificities for 56 kinases in *H. sapiens* and 46 kinases in *S. cerevisiae* were considered. For all other PTMs such as methylation, ubiquitination, and acetylation that do not exhibit explicit flanking sequence specificities, a variant was considered damaging if it directly altered the modified site. This resulted in a total of 6.3M possible variants that could alter such PTM sites across the three organisms (n=5.8M *H. sapiens*, n=537,434 *S. cerevisiae*, n=9,177 *E.coli*). For linear motifs, we gathered 1,668 experimentally identified linear motifs (n=1,525 *H. sapiens*, n=143 *S. cerevisiae*), along with their derived regular expression pattern from the ELM database [161] and computed the impact of all possible variants 226,920 (n=205,120 *H. sapiens*, n=21,800 *S. cerevisiae*) on binding patterns. Finally, for TFBSs, for organisms without well-defined functional TFBSs (*H. sapiens* and *S. cerevisiae*), we defined putative TF-gene regulatory network using TF-knockdown expression data and/or ChIP-seq/ChIP-chip, as previously described. We then used PWMs to identify putative binding sites, and predict the impact of all possible 3.6M variant substitutions (n=3.3M *H. sapiens*, n=236,382 yeast, n=46,768 *E. coli*) on specificities of 217 TFs (n=72 *H. sapiens*, n=104 *S. cerevisiae*, n=41 *E. coli*). It is noteworthy to mention that although in Chapter 3 (section 3.2.3) we show that machine learning-based approaches outperform PWMs (used here) at predicting TFBS variant impact, the work presented in this chapter was carried out prior to that of Chapter 3.

The precomputed mechanistic variant effects data are stored in a normalised MySQL database in a mechanism-specific manner (Figure 4.3). All variants are stored within a primary mutation table, which then relates to the individual mechanism-specific tables in a one-to-many manner. The mutation table also references a position table containing positions of all variants, which in turn references a gene/chromosome table in a many-to-one manner. All these tables are indexed to allow for rapid lookup amongst millions of entries.

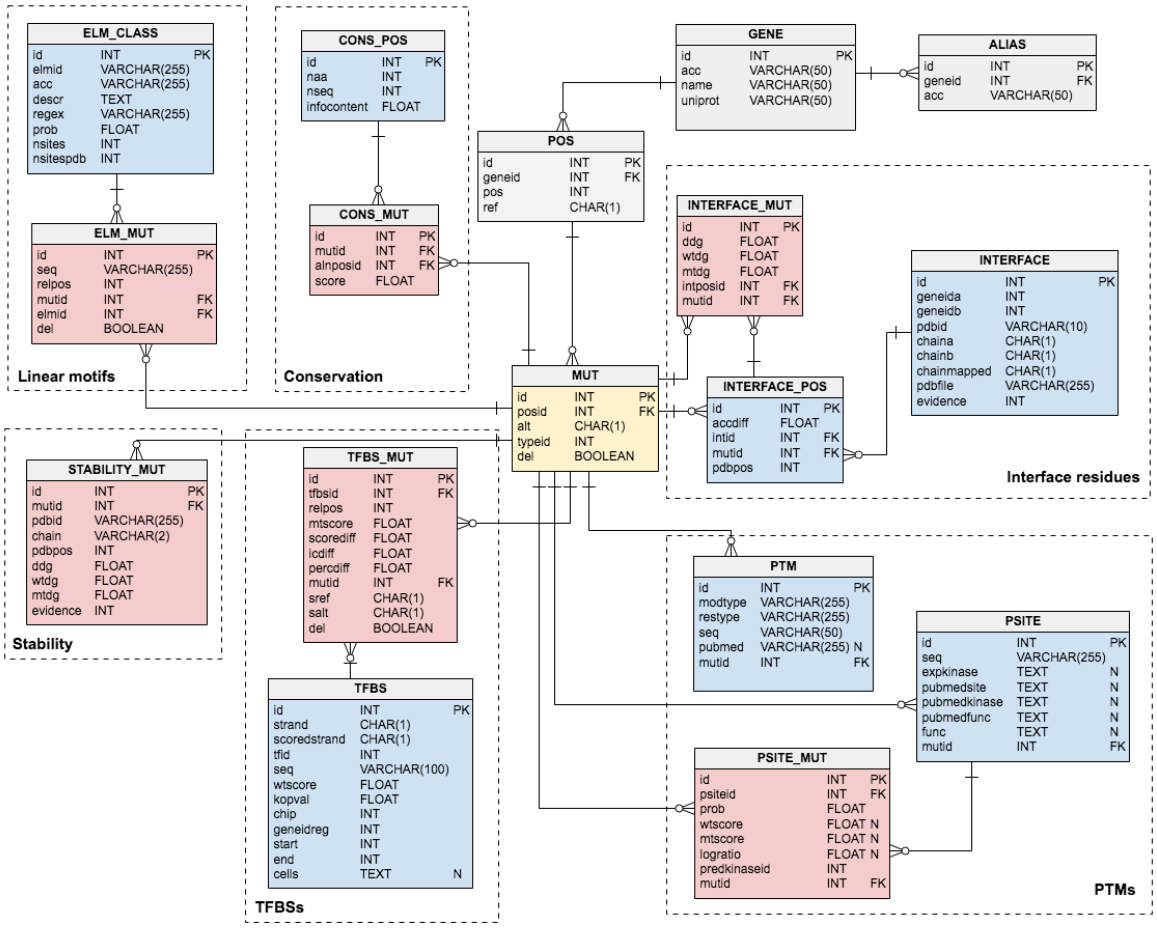



Fig. 4.3 The MySQL mutfunc database schema showing the structure of the database, tables, and relationships. The primary table `MUT` (yellow) stores all possible DNA and amino acid variants, which relate to mechanism-specific tables (red). These then relate to additional tables (blue) containing information on the affected mechanisms.

### The mutfunc web server user interface

The mutfunc user interface provides an intuitive, user-friendly and interactive way by which users can query the database using their own variants. Both DNA or protein substitutions can be provided to mutfunc in one of two formats, plain text format or the variant call format (VCF). The plain text format is a simplified format for variants, where variants are line separated. Variants should be formatted as follows `NAME_X123Y` or `NAME_123_X_Y`, where `NAME` is name of the gene (UniProt accessions, gene names or IDs are acceptable) or chromosome (number or NCBI IDs), `123` is the position of the variant, and `X` and `Y` are the wildtype and mutant amino acid or base, respectively (Listing 4.1). Alternatively, the VCF format can be used, which is a format commonly used by variant calling pipelines. In its simplest form, a VCF file is a five-column tab-delimited file containing the chromosome,

position, id, wildtype and mutant alleles of a DNA variant (Listing 4.2). Both formats can be provided either via the text box or by uploading a file.

Listing 4.1 Plain text format example

```
chr1_61177_G/A
YPL228W_N274C
MYO4_K1366N
YJL158C K63A
```

Listing 4.2 VCF format example

```
chr17   10987590   ID1   G   T
chr10   23508363   ID2   A   G
chr16   52599188   ID3   C   T
chr16   20932709   ID4   T   C
```

Variants are processed then queried against the database. If DNA variants fall within a coding region and encode a nonsynonymous substitution the corresponding protein variant is also queried against the database. Since all predictions are precomputed, mutfunc is able to analyse an extremely large number of variants typically within seconds.

An interactive report of variant effects is returned to the user, which contains a table of variants matching the database (Figure 4.4a). The predicted effect of each variant are categorised into six classes: (1) PTMs and linear motifs (2) Stability (3) Interfaces (4) Conservation (5) TFBS (6) Start-stop codons. Each row in the table contains a series of coloured and labelled badges, where each badge is coloured and labelled distinctively based on the mechanism class (Figure 4.4b). Expanding the row allows for mechanistic effects to be further explored, providing additional information on the prediction made such as the score as provided by the predictor, visual representations of the variant, and external links to references. For instance, details of a variant affecting protein stability or interfaces will show FoldX-predicted $\Delta\Delta G$ values, alongside an interactive visualization of the variant in the context of the three-dimensional structure. Variants impacting conserved regions will show SIFT scores and an MSA of the affected position in context. Variants impacting SLiMs, PTMs or TFBSs will show the local sequence context before and after the mutation as well as sequence logos if a motif is involved (Figure 4.4b).

Variants of a single protein can be visually inspected using the interactive protein viewer (Figure 4.4c). Using an adapted version of the neXtProt [349] feature viewer, variants within a protein are visualised in the context of different protein feature tracks including protein domains from PFAM [350], regions of disorder from MobiDB [351], secondary structure from UniProt [149] and PTMs. A separate track is separately displayed for each class of variants. The viewer allows interactive zooming and interfaces with the results table i.e. selecting a variant in the viewer will highlight the corresponding row.

Results can be filtered by gene or chromosome keywords and by different classes of variants. All results can also be exported for analysis offline, in which an archive of tab-delimited flat (one for each variant class) files are made available to download. All jobs are

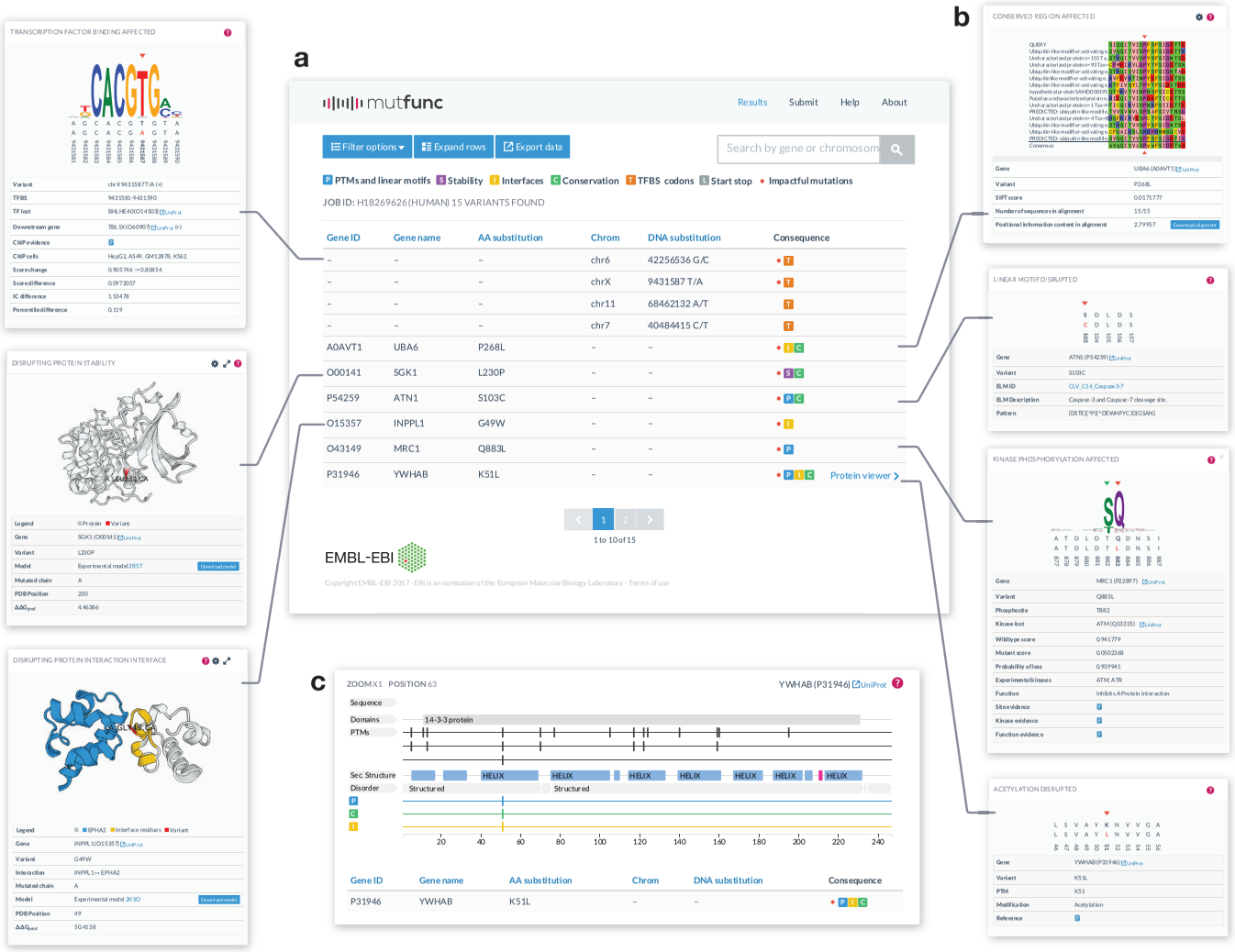



Fig. 4.4 An overview of the mutfunc web server. (a) After variants are processed, a results page shows a table of both DNA and/or amino acid variants submitted, along with the predicted mechanism affected. (b) Each row in the result table can be expanded to further explore the altered mechanism. (c) The protein viewer allows variants within a protein to be interactively explored in the context of other feature tracks, such as protein domains, secondary structure, and intrinsic disorder.

stored for 48 hours, after which all submitted data and generated results are removed from the server.

The mutfunc resource serves as a useful platform for small and large-scale studies, allowing variants to be mechanistically explained and prioritised.

### 4.2.3 Validation of predictions

To demonstrate the ability of predictions provided in mutfunc, we aimed to explore the properties of a large number of predictions generated for yeast and *H. sapiens*. We explored the deleteriousness of variants in the context of essential genes as well as allele frequency. We further leveraged a number of data sets that have manually curated variants as either being deleterious to function or having no effect in order to validate predictions.

#### Essential genes harbour fewer deleterious variants

Essential genes are those required for survival and are often identified by disrupting the gene and assessing the viability of the organism or cell. In *S. cerevisiae*, roughly 20% (1,114) genes have been identified as essential well over a decade ago by the *Saccharomyces* Genome Deletion Project (SGDP) consortium. More recently, this has been made possible in *H. sapiens* through CRISPR and gene trapping technology [352, 353]. For instance, Blomen et al. identified roughly 8% (1,734) of *H. sapiens* genes as essential in two cell lines [352].

Utilising knowledge of essential genes in both *S. cerevisiae* and *H. sapiens*, we explored how often essential genes exhibit variants impacting conserved regions (sift score $< 0.05$), protein stability and interface residues ($\Delta\Delta G > 2$), altering start codons or stop codons (nonsense and nonstop variants). We counted deleterious variants in essential and non-essential genes, normalised by the length of the protein. We found that essential genes consistently demonstrated significantly lower frequencies of variants predicted to affect conservation, stability, interfaces and alter start and stop codons in across *H. sapiens* and *S. cerevisiae* (Figure 4.5). Specifically, conservation showed the most significant separation between essential and non-essential genes ($p$=$1.04\times10^{-46}$ *H. sapiens*, $p$=$1.52\times10^{-22}$ *S. cerevisiae*), followed by protein stability ($p$=$1.82\times10^{-12}$ human, $p$=$7.07\times10^{-10}$ *S. cerevisiae*). Although in *S. cerevisiae* fewer deleterious interface variants impact essential genes ($p$=$5.43\times10^{-4}$), in *H. sapiens* there was surprisingly no observed difference ($p$=0.70, Figure 4.5c). Finally, of the variants affecting the start and stop codons, nonsense variants showed the highest significance ($p$=$8.49\times10^{-23}$ *H. sapiens*, $p$=$1.08\times10^{-5}$ *S. cerevisiae*), followed by start loss and nonstop variants (Figure 4.5d-f). However, similar to interfaces, start loss variants did

not show a significant difference in *H. sapiens* ($p$=0.51), whereas *S. cerevisiae* exhibited mild significance (Figure 4.5f, $p$=0.012).

These results demonstrate that essential genes, in most cases, harbour fewer variants that would affect its function, confirming both the essentiality of the genes as well as the reliability of predictions made.

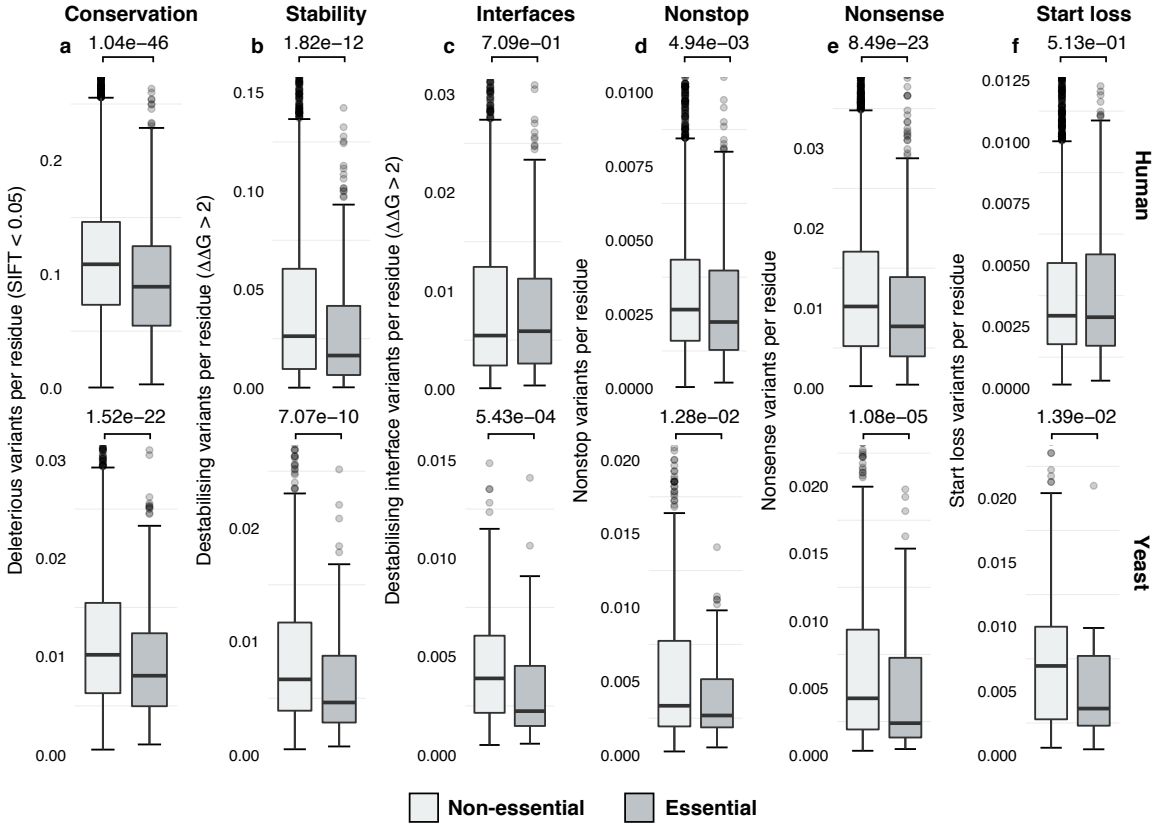



Fig. 4.5 Essential genes harbour fewer variants impacting mechanisms across *H. sapiens* (top) and *S. cerevisiae* (bottom). (a-f) Box plots show the count of variants, normalised by protein length for essential and non-essential genes in each mechanism. *P*-values are calculated based on a one-sided Wilcoxon test.

### Common variants are more tolerated

Variants that occur commonly in the population should by definition less likely to have strong effects on molecular phenotypes. To confirm this, we analysed the predicted impact of variants in the context of allele frequencies. Variant effect predictions for conservation, stability, and interfaces were binned by their MAF into 5 groups (0-1%, 1-5%, 5-20%, 20-50% and >50%) and the distribution of variant impact scores in each bin was assessed.

Both conservation and stability showed a clear linear relationship between the MAF and predictor score across *H. sapiens* and *S. cerevisiae*. Variants with a MAF < 1% showed an average SIFT score of 0.27 and 0.4 in *H. sapiens* and *S. cerevisiae*, respectively, compared to variants with MAF > 50%, which showed significantly higher SIFT scores of 0.60 and 0.60 ($p<2.2\times10^{-16}$). Stability effects for variants followed similar trends with < 1% MAF variants showing an average $\Delta\Delta G$ of 1.5 and 0.75 in *H. sapiens* and *S. cerevisiae*, respectively, compared to 0.2 and 0.15 in high MAF variants (> 50%). Interfaces did not follow such trends, likely due to the highly imbalanced number of variants across MAF bins.

Grouping MAFs by bins often relies on having a significant number of MAF values per bin. To more directly compare distributions of MAFs, we compared the quantiles of MAF distributions of impactful variants to that of all nonsynonymous variants. For conservation, stability and interface predictions, deleterious variants were defined by cutoffs of $s < 0.05$ for conservation, and $\Delta\Delta G > 2$ for stability and interface predictions. Results showed that impactful variants typically exhibited significantly lower MAF values with the exception of both nonstop and start lost variants (Figure 4.6b). Variants affecting PTMs in *S. cerevisiae* also did not show a significant difference (Figure 4.6b).

The AF is a useful metric to consider when interpreting variant effects. These results demonstrate that it is often the case predicted deleterious variants demonstrate lower AF.

### Predicted deleterious variants are enriched in functionally important variants

We next tested the ability of predictors included in mutfunc to identify functionally significant variants in *S. cerevisiae* and *H. sapiens*. For *H. sapiens* we used a total of 34,600 variants annotated to be pathogenic (n=17,167) or benign (n=17,433) in ClinVar database[354]. For *S. cerevisiae* we utilised 8,083 variants consolidate by Jelier et al. [355] as either tolerated (n=5,271) or affecting function (n=2,812) (Methods, section 4.3.1).

Using the quantitative score of conservation, stability, and interface predictions, we assessed the performance at which predicted scores are able to discriminate functional variants. Predictors consistently demonstrated satisfactory performance across both *H. sapiens* and *S. cerevisiae*. Predictions based on SIFT performed the best at discriminating pathogenic variants from benign (AUC *H. sapiens* = 0.87, *S. cerevisiae* = 0.92), followed by FoldX interfaces (AUC *H. sapiens* = 0.64, *S. cerevisiae* = 0.72) and FoldX stability (AUC *H. sapiens* = 0.70, *S. cerevisiae* = 0.62, Figure 4.7a). A possible explanation for stability and interface predictors having lower performance is that they explain only individual mechanisms and, therefore, misclassified variants are likely involved in alternate mechanism driving the pathogenicity.

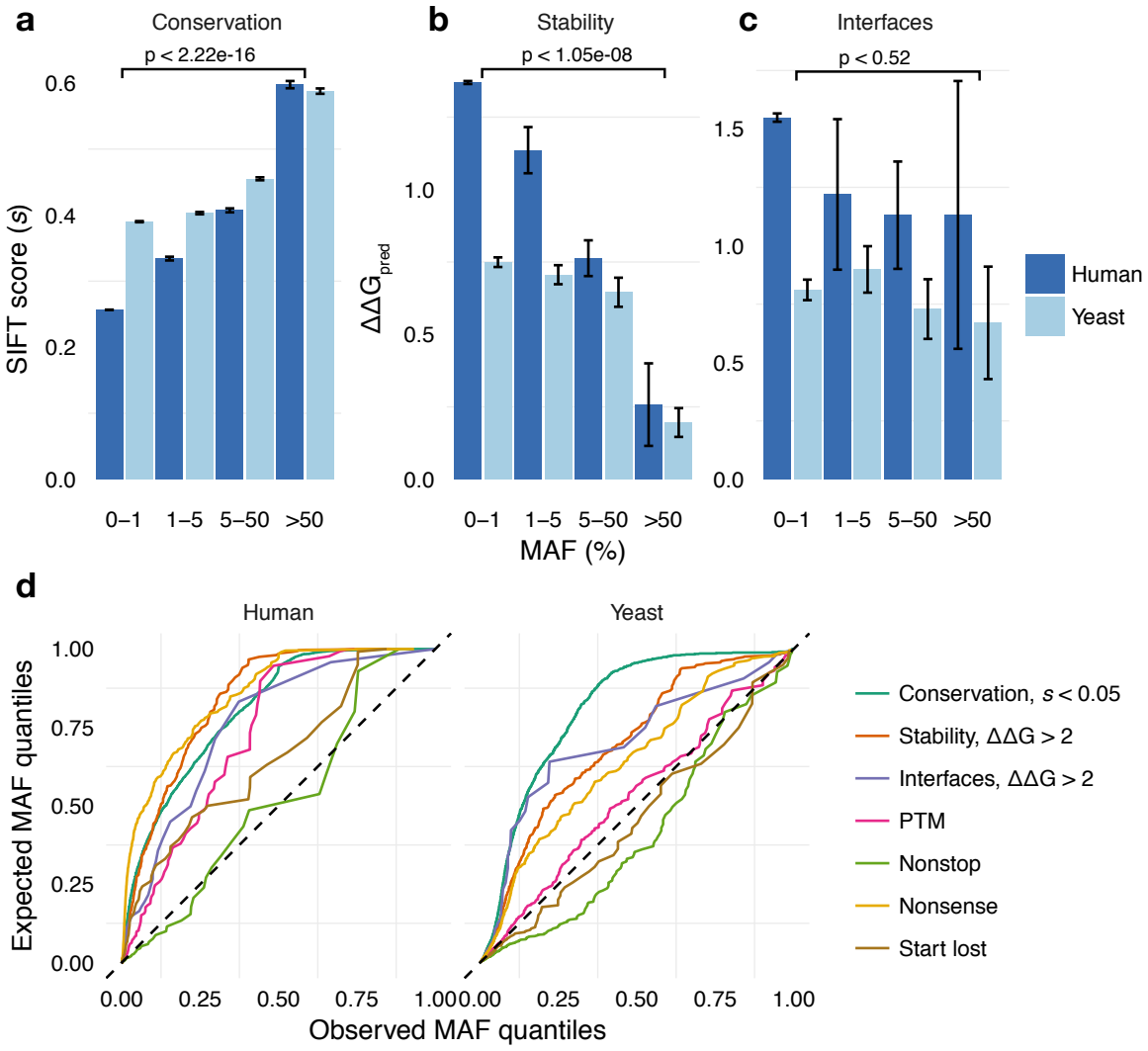



Fig. 4.6 Common variants are less impactful across *H. sapiens* and *S. cerevisiae*. (a-c) Bar plot of mean SIFT scores and predicted $\Delta\Delta G$ values for variants within different MAF bins. Error bars represent the standard error and *p*-values are calculated based on a one-sided Wilcoxon test. (d) Quantile-quantile plots of MAFs between observed MAFs of variants impacting a mechanism and expected MAFs of all coding variants.

Other heuristic-based predictors such as SLiMs and PTMs that provide a binary classification of deleteriousness were also checked if variants predicted to be affecting a cellular mechanism were enriched in pathogenic/deleterious variants. For each class of predictions, the proportion of functional variants is computed and compared against a background set of variants using a one-sided Fisher's exact test to obtain a *p*-value describing the significance of observing the proportion by random chance. This was carried out on variants that lie within a SLiM and disrupt or retain the regular expression, variants that impact PTM and

non-PTM residues, and variants altering start and stop codons. We found that, despite low numbers, variants disrupting regular expression patterns of linear motifs were enriched ($p$=5.23×10$^{-3}$ *H. sapiens*, $p$=0.12 *S. cerevisiae*). The insignificant $p$-value observed in *S. cerevisiae* is likely due to a scarcity of observed SLiM-disrupting variants. Similarly, in *S. cerevisiae*, variants altering PTM sites were shown to be enriched in deleterious variants ($p$=8.44×10$^{-7}$). This observation did not extend to *H. sapiens*, likely due to the much higher quantity of PTMs, many of which are non-functional [283]. Finally, variants that disrupt start codons as well as nonsense and nonstop variants all displayed high proportions of deleterious/pathogenic variants (>80%), which was significantly higher than the proportion of all variants ($p$<2.14×10$^{-4}$ *H. sapiens*, $p$=0.052 *S. cerevisiae*).

The results here demonstrate that predictors utilised in mutfunc are capable of identifying variants of functional significance, further demonstrating the practicality of mutfunc.

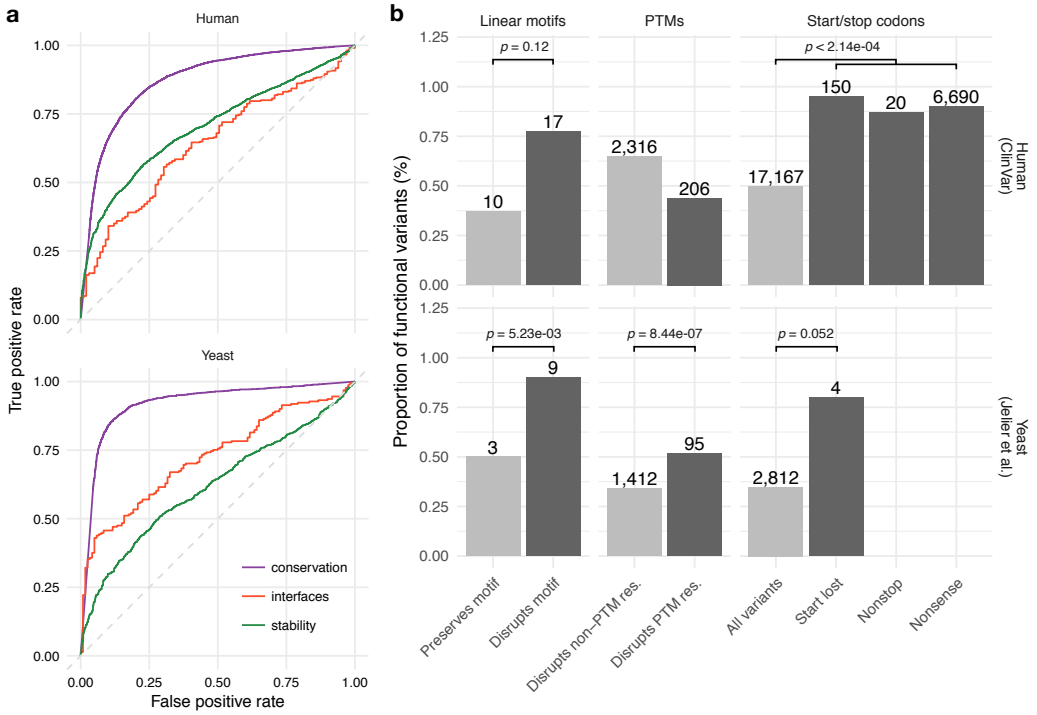



Fig. 4.7 Validation of predictors in mutfunc using functional and pathogenic variants in *S. cerevisiae* and *H. sapiens*. (a) The predicted impact of variants on conservation, stability and interaction interfaces can accurately distinguish functional variants. (b) Functional variants are enriched in those predicted to alter SLiMs, PTMs and start and stop codons. Bar plots show the proportion of variants in each bin deemed functional. Numbers above each bar plot denote the number of variants. All $p$-values are calculated using a one-tailed Fisher's exact test.

### 4.2.4 Predicting mechanistic insight into variants of uncertain significance

Variants that have been identified through genetic testing but are yet to be deemed benign or pathogenic are termed variants of uncertain significance (VUS). The interpretation of such variants is a common challenge in genetics, one that is often aided by computational predictors. We sought to employ the mutfunc database to predict protein-coding VUSs. A total of 64,692 variants labelled with "uncertain significance" were collected from ClinVar [354], along with the disease phenotype in which they were tested for. VUSs were annotated using mutfunc and 21,584 variants were predicted impactful by at least one of the mechanistic predictors, not including SIFT (n=7,547 stability, n=751 interfaces, n=139 linear motifs, 2,372 PTMs, 57 kinase-binding). We focus on variants predicted to impact the structural integrity of proteins (stability and interaction interfaces) since they hold the highest coverage.

Over 38% (751/1981) of VUSs are predicted to interfere with interface binding stability. VUSs were retained if (1) the protein also harbours a pathogenic variant predicted by the same predictor as impactful and (2) both the pathogenic variant and VUSs are identified in patients carrying the same disease. This allows us to focus on higher confidence VUS predictions for which we know a pathogenic variant alters the same mechanism. We demonstrate a few examples of VUSs that are predicted to alter binding to highlight the utility of predictions in mutfunc. For instance, primary hyperoxaluria is a disease caused primarily by mutations in the *GRHPR* gene, a glyoxylate and hydroxypyruvate reductase [356, 357]. Enzymatic activity of *GRHPR* requires homodimerization [358]. The VUS R171H is predicted to impact a conserved region as well as the homodimerization stability ($\Delta\Delta G = 2.19$, $s < 0.018$), thereby impinging on the function of GRHPR. Interestingly there have been two other pathogenic variants R302H and E113K that are implicated in primary hyperoxaluria and are also predicted to impact conserved regions and binding stability ($\Delta\Delta G > 2.15$, $s < 0.012$), further supporting the evidence of an altered mechanism for the VUS. Similar to this example, mutations in fumarate hydratase (*FH*) have been shown to play crucial roles in fumarase deficiency and cancer [359] and regular function for *FH* is attained through the formation of a homotetramer [359]. A number of VUSs throughout the FH homotetramer interfaces identified in fumarase deficient patients have been predicted as disrupting binding stability, such as the S334R mutation, which shows a $\Delta\Delta G$ of 6.31, which would result in loss-of-function of FH. The pathogenic variants R233C and D341N have been predicted as impacting binding stability ($\Delta\Delta G > 2.81$) and are implicated in cancer and fumarase deficiency, respectively.

Similar to interface variants, we analysed variants that destabilise the protein structure. We identified 1,182 VUSs predicted to alter stability in proteins containing pathogenic variants

also predicted destabilizing. For example, in the ubiquitin ligase PARK2, implicated in Parkinson's disease, two rare VUSs (R42H, V148E) identified in Parkinson's disease patients are predicted to destabilise the protein ($\Delta\Delta G > 4.7$, Figure 4.8d). PARK2 also contains other pathogenic variants implicated in Parkinson's disease and predicted to be destabilizing. In the tumour suppressor serine/threonine-protein kinase STK11, pathogenic and VUS identified in Peutz-Jeghers syndrome patients show $\Delta\Delta G$ scores predicting destabilisation (Figure 4.8e). In particular, the VUS G242W shows an exceptionally high destabilizing score ($\Delta\Delta G >$ 38.96).

The analysis here demonstrates how mutfunc could be applied to systematically describe altered mechanisms through candidate disease variants.

## 4.3 Methods

### 4.3.1 Genetic variant data collection

A total of 896,772 genetic variants occurring in for 405 haploid and diploid *S. cerevisiae* strains were collected from four studies [336–339]. All but one study by Strope et al. [336] provided processed variant calls in VCF format. Variants were called for the Strope et al. study using the following pipeline. Raw reads were obtained from the ENA resource [360]. Adapter sequences were removed using cutadapt v1.8.1 [361] and reads were mapped to the *S. cerevisiae* genome version 64 using BWA-MEM v0.7.8 [362]. Duplicate reads were discarded using picard v1.96 (https://github.com/broadinstitute/picard) and reads were realigned using the GATK indel realigner v3.3 [363]. Base alignment qualities were computed using samtools v1.2 [364] and variants were called using freebayes v0.9.21-15-g8a06a0b [365] and the following parameters `-no-complex`, `-genotype-qualities`, `-ploidy 1` and `-theta 0.006`. The VCF was filtered for calls with `QUAL > 30`, `GQ > 30` and `DP > 4`. VCF for individual *S. cerevisiae* strains were combined and coding variants were called using the `predictCoding` function of the VariantAnnotation R package [366].

A total of 3,198,692 coding variants in *H. sapiens* for over 65,000 individuals was collected from the ExAC consortium [298] in the ANNOVAR [317] output format along with corresponding adjusted allele frequencies. Ensembl transcript positions were mapped to UniProt by performing Needleman-Wunsch global alignment of translated Ensembl transcript sequences against the UniProt sequence using the `pairwiseAlignment` function in the Biostrings R package. The mapping between Ensembl transcript IDs (v81) and UniProt accessions was obtained from the biomaRt R package [367]. In the case that multiple alleles

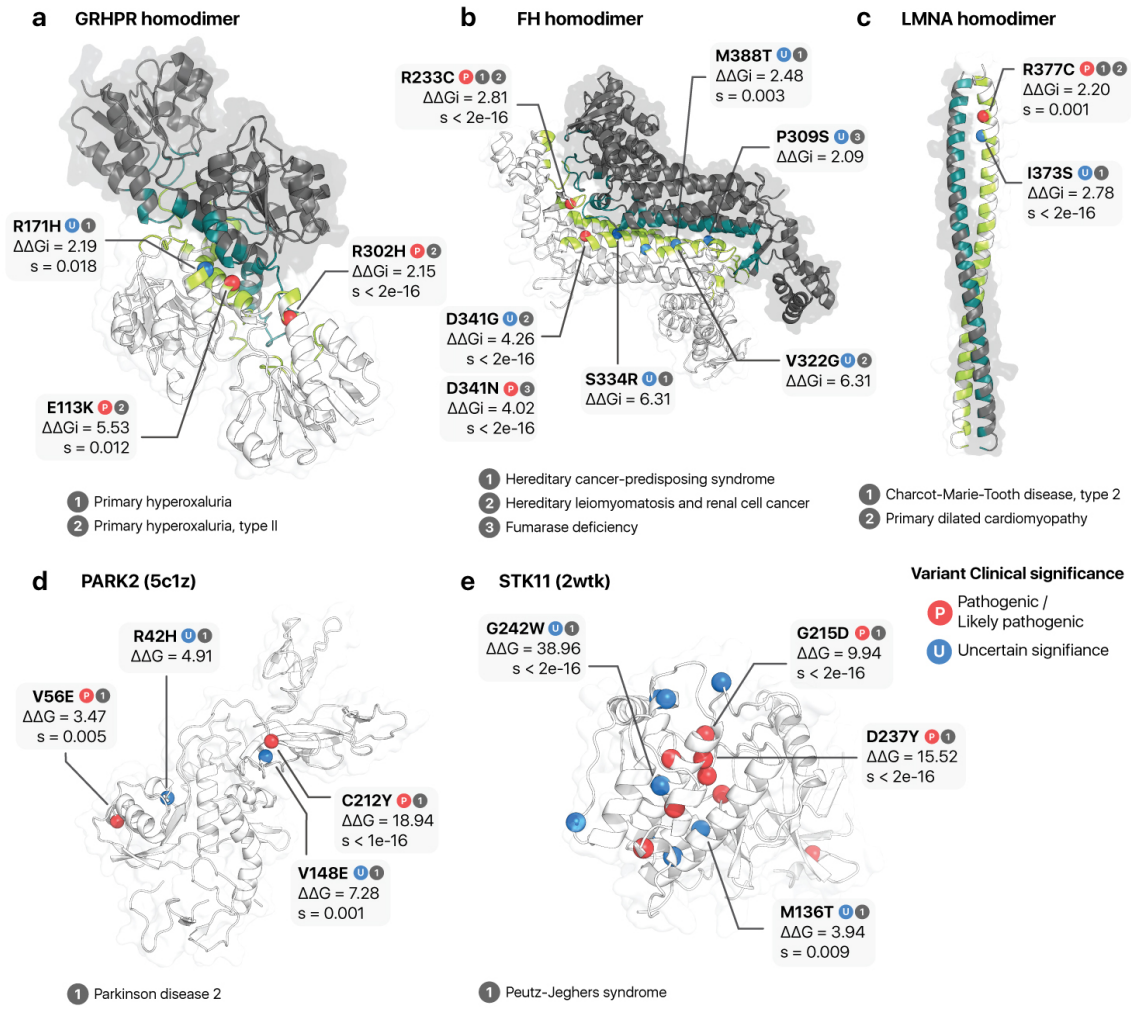



Fig. 4.8 *In silico* validation of VUSs using mutfunc predictions. (a-c) Three examples of interaction interfaces containing variants predicted to impact binding stability. Subunits of the interaction complex are coloured in dark grey and white, and respective interface residues in dark green and green. (b) Two examples of variants predicted to impact protein stability. Pathogenic variants are labelled "P" in red, and VUSs "U" in blue.

mapped to the sample single amino acid substitution, the one with the highest adjusted allele frequency was retained.

A total of 139,167 variants were obtained from ClinVar [354]. Only variants that did not match one of the following clinical significance terms were removed: 'Benign', 'Benign/Likely benign','Likely benign', 'Likely pathogenic', 'Pathogenic/Likely pathogenic', and 'Pathogenic'. Variants with a review status of 'no assertion criteria provided' were also removed, as those reflect variants that have been assigned clinical significance without any particular criteria. The final filtered set contained 39,597 variants. Of these variants, 44% were classified as pathogenic or likely pathogenic. For *S. cerevisiae*, a total of 8,083

manually curated variants were obtained from [355], 34.5% (2,812) of which were labelled as deleterious. Variants were collected from a combination of the UniProt database [149], Protein Mutant Database [368], Saccharomyces genome database [369] and mutations that are identified in essential genes [370].

### 4.3.2 Evolutionary constraint

The evolutionary constraint, defined as $c$, is computed by taking the ratio between observed mutation counts in a region of interest and that of random regions.

To assess the difference between $c$ in buried vs. exposed protein structures and interfaces, variants are counted in each of the four bins of RSA. An equal number of variants are then sampled 100 times and the number of variants in each bin are counted. The observed and expected counts for each bin are divided for 100 values of $c$.

For PTMs, the observed number of mutations in and around the modified site is counted. The expected number of mutations is calculated by sampling an equal number of un-modified sites in the same genes harbouring the observed PTMs. This is repeated 100 times, for each of which $c$ is calculated.

For TFBSs, the observed number of variants is computed by counting the number of mutations in predicted TFBSs which also overlap with ChIP-seq or ChIP-chip regions. This is compared to mutation counts in random regions of the same length in the ChIP-seq or ChIP-chip regions. This is similarly repeated 100 times.

There are a number of limitations related to this approach of computing evolutionary constraint. Namely, it does not consider variable mutation rates between nucleotides. Mutation rates have been shown to vary significantly in regions with base composition biases [371], local recombination rates [372], chromatin structure [373] and many other factors. Including these would allow for more accurate constraint measurements.

### 4.3.3 Essential genes

A total of 2,501 essential genes identified using gene trapping technology in two haploid *H. sapiens* cell lines KBM7 and HAP1 were obtained from [352]. These were further filtered for genes that were essential in both cell lines, for a total of 1,734 genes. A total of 1,156 essential genes in *S. cerevisiae* were obtained from the *Saccharomyces* Genome Deletion Project [374].

### 4.3.4 Predicting impact on protein stability and protein interaction interfaces

Experimentally determined structures were obtained from the protein data bank (PDB). Large structures that did not have a corresponding PDB file were downloaded in mmCIF format and converted to PDBs using the PyMOL Python library v1.2r3pre [375]. Mapping of coordinates from PDB to UniProt residues was derived from the SIFTS database [376]. Structures with a resolution above 3 were discarded and a single representative structure maximising the coverage of the protein was retained. Homology modelling was carried out for proteins with no experimentally determined structures using ModPipe version 2.2.0 [377] and the following parameters: `-hits_mode 1110` and `-score_by_tsvmod OFF`. For each protein, the model with the highest normalised DOPE score was retained. Experimental and homology modelled structures for protein interactions were obtained from the Interactome3D database [205]. Relative solvent accessibility (RSA) for all residue atoms was computed using NACCESS [378] for proteins individually, and in the interaction complex. Interface residues were defined as those with any change in RSA. All other cases of RSA were computed using freeSASA v1.1 [379].

The impact of variant on stability was computed using FoldX v4.0 [189]. All structures were first split by chain into individual PDB files and repaired using the `RepairPDB` command, with default parameters. The `Pssm` command is then used to predict $\Delta G$ with `numberOfRuns=5`. This performs the mutation multiple times with variable rotamer configurations, to ensure the algorithm has achieved convergence. The average $\Delta G$ of all runs is computed and the $\Delta\Delta G$ is computed as the difference between the wildtype and mutant. The impact of variants on interaction interfaces is measured similarly, with the exception of structures being provided in binary interaction, rather than individual chains.

### 4.3.5 Predicting the impact of variants on PTMs and linear motifs

For *S. cerevisiae*, a total of 20,056 phosphosites and 2,219 kinase-substrate associations were obtained from the PhosphoGRID database [380]. A total of 1,070 of other PTM sites was obtained from the dbPTM database [116]. For *H. sapiens*, all PTM data, including that of phosphorylation and kinase-substrate associations were obtained from PhosphoSitePlus [252], for a total of 296,147 sites. For *E. coli*, a total of 483 PTM sites were obtained from dbPTM [116]. Linear motif data for *S. cerevisiae* and *H. sapiens*, including annotated linear motif binding sites and regular expression patterns, were obtained from the ELM database [381].

Impact of variants on phosphosites and flanking regions was measured using the MIMP algorithm [146], with default parameters. For other PTMs, a variant was predicted to be impactful if it affected the modified residue. For linear motifs, a variant was predicted to be impactful if it causes a loss of match for associated regular expression pattern.

### 4.3.6 Predicting the functional impact of variants using conservation

All protein alignments were built against UniRef50 [382], using the `seqs_chosen_via_median_info.csh` script in SIFT 5.1.1 [244]. The siftr R package (https://github.com/omarwagih/siftr), an implementation of the SIFT algorithm that was developed by myself, was used to generate SIFT scores with parameters `ic_thresh=3.25` and `residue_thresh=2`.

### 4.3.7 Transcription factor binding sites

A total of 177 *S. cerevisiae* TFs binding models were collected in form of a position frequency matrices (PFMs) from JASPAR [287] and converted to position weight matrices (PWMs) using the TFBSTools R package [315]. PWMs were trimmed to eliminate consecutive stretches of low information content (<0.2) on either terminus. To identify genes likely regulated by a particular TF, a combination of TF-knockout expression and ChIP-chip experiments were used, as similarly described in [383]. Genome-wide gene expression profiles for 837 gene-knockout strains were obtained from three studies [384–386], 148 of which were a known TF with a defined PWM. Studies provided either a $Z$-score or $p$-value for each gene as a measure of over or under-expression, relative to the distribution of values for all genes. Two-tailed p-values were computed from $Z$-scores when a $p$-value was not provided [384]. In cases where TF knockout was repeated between studies, the lowest p-value for each gene was used. ChIP-chip tracks for 355 TFs were collected from four studies [387–390], via the Saccharomyces genome database [369]. Of the 355 of the TFs, 144 (56%) had a defined PWM. Potential binding sites were then only searched for in TF-gene pairs with a $p$-value below 0.01 and the corresponding ChIP-chip region upstream of the regulated gene. A normalised log score of 0.80 was used as the cutoff for defining putative binding sites. Similarly, for *H. sapiens*, 454 TF PWMs were generated from JASPAR PFMs. ENCODE clustered ChIP-seq data were obtained for 161 TFs, of which 72 had a PWM. Only those regions were scored against the corresponding PWM. For *E.coli*, a total of 1,905 TF-matching sequences across 84 TFs were obtained from RegulonDB [391] and used to construct PWMs. A total of 2,416 experimentally identified TFBS were obtained for 79/84

TFs from RegulonDB. These sites were used as putative binding sites for downstream variant predictions.

Potential target sequences could then scored against the PWM using the log-scoring scheme defined in [60] and normalised to the best and worst matching sequence to the PWM. The resulting score lies between 0 and 1, where 1 signifies strong predicted binding by the factor, whereas 0 signifies predicted lack of binding. Potential binding sites are scored in the presence ($s_{wt}$) and absence ($s_{mt}$) of a variant. Three separate metrics are used to quantify the change in binding between the reference and alternate allele. The first one is simply the difference in the normalised log score, $S_{wt} - S_{mt}$, where a large positive value indicates loss of binding. The second is the difference in binding percentile. Here, random oligonucleotides are used to generate a negative distribution of log normalised scores for each TF. The percentile of each wildtype $p_{wt}$ and mutant scores $p_{mt}$ is computed from this distribution, and the difference, $p_{wt} - p_{mt}$, is used to quantify the magnitude of impact. The last is the difference in the relative information content. This can be thought of as the difference of letter height in a sequence logo. Given that the wildtype and mutant bases have relative frequencies of $f_{wt}$ and $f_{mt}$, respectively and a position has an IC value of $\gamma$, then this is computed as $(f_{wt} \cdot \gamma) - (f_{mt} \cdot \gamma)$. This value ranges from 0 to 2, where 0 indicates little to no impact on a critical base, and 2 indicates a strong one.

### 4.3.8 Implementation of mutfunc

Described predictors were used to precompute effects for all amino acid and nucleotide substitutions. The resulting data serves as the basis for mutfunc, allowing users to rapidly query thousands of variants on predictions that would otherwise take on the order of days or weeks to compute, particularly those involving 3D structures.

The mutfunc web server at http://mutfunc.com is free and open to all users and requires no login. The web application uses the Java and Scala-based Play Framework v1.3.7 backend (http://www.playframework.org) along with a MySQL database. The front-end utilises a modified version of the the Twitter Bootstrap UI library (http://twitter.github.com/bootstrap). Visualization tools used include a modified version of the neXtProt feature viewer v0.1.52 (https://github.com/calipho-sib/feature-viewer) for interactive visualisation of protein sequence features, WebGL protein viewer v1.1 for interactive visualisation of protein structures v1.8.1 (https://github.com/biasmv/pv), and a modified version of the JSAV v.1.10 library (https://github.com/AndrewCRMartin/JSAV) for visualization of multiple sequence alignments.

## 4.4 Discussion

A complete understanding of how genetic variation drives phenotypic variability relies majorly on understanding the mechanisms they impinge on and how this propagates through to phenotype. Computational predictors that utilise both sequence and structure features have been developed to aid this process. In this chapter, I have described the mutfunc resource, in which numerous predictors were used to precompute millions of variant effects across *H. sapiens*, *S. cerevisiae*, and *E. coli*. I have further explored predictions made in the context of natural and disease variation in both *H. sapiens* and *S. cerevisiae* genomes and validated the performance of predictors. Such predictors have shown a promising capacity at not only identifying causality of a variant but also which mechanisms they likely impact.

The increasing availability of individual genomes is allowing for the detection of many common and rare variants in both coding and non-coding regions. Such variants can easily be queried in mutfunc, allowing for rapid hypothesis-driven annotation and prioritisation. Currently, conservation effects hold the highest coverage, (*H. sapiens* 98.6%, *S. cerevisiae* 87.9%, and 96.1% *E. coli*) followed by stability (*H. sapiens* 18.9%, *S. cerevisiae* 16.9%, and 49.2% *E. coli*) and interfaces (*H. sapiens* 2.20%, *S. cerevisiae* 2.84%, and 4.45% *E. coli*). Other mechanisms like kinase-substrate phosphorylation, SLiMs, and TFBSs have much lower coverage and depend on the availability of external, often manually curated, data. As additional data become available, mutfunc will be updated to improve coverage.

Hypotheses derived from *in silico* predictions, such as those provided in mutfunc, should be exploited with caution. Despite the accuracy of many predictors, the inherent effect of a genetic variation *in vivo* can be far more complex and depend on both genetic and environmental factors [392]. Several studies have shown that many variants annotated as disease-causing or predicted as deleterious have been identified in healthy humans [393], emphasizing the discretion required when deeming a variant deleterious. One of the major factors confounding variant effect predictions is epistatic effects, where the impact of a variant can be mitigated or aggravated by the occurrence of alternative genetic variants [394]. The genetic background in which a variant exists is therefore critical to understanding genomic regions that have undergone co-evolution to suppress deleterious effects. Ultimately, *in silico* predictions should not be used as actionable clinical evidence, but rather to guide follow-up validation experiments, which could then either confirm or deny the role of the variant in the underlying mechanism [395].

The utility of mutfunc lies within the precomputed effects of individual point mutations allowing a large number of variants to be rapidly queried without the need for on-the-fly computations, which can often be time-consuming. Within such a framework, epistatic effects cannot be precomputed due to a large number of possible combinations. Similarly,

many other types of genetic variation that largely contribute to phenotypes such as CNVs and indels [396, 397] cannot be consistently precomputed in mutfunc due to their atypical structure. Other aspects mutfunc that could be improved in future versions. For instance, despite there being several well-studied mechanisms available in mutfunc, there are many mechanisms that are yet to be integrated, such as splicing, protein localization, and epigenetic modifications. Some of these mechanisms are not currently included in mutfunc since they remain difficult to predict with existing poor accuracy options. The development of accurate predictors of molecular phenotypes relies upon both the biological understanding of molecular determinants underpinning the mechanism as well as the availability of experimentally-verified training data. Lastly, many organisms in which genetic variation is commonly studied are not included in mutfunc. These include as *M. musculus*, *D. melanogaster* and *A. thaliana*, which contain an abundance of data on PTMs, SLiMs, structural data and TFBSs. Predictors employed here could thus be applied to provide mechanistic variant impact predictions for these organisms, expanding the utility of mutfunc.

Understanding how disrupted cellular mechanisms propagate to changes in phenotypes is critical for variant interpretation. For instance, predicting the disruption of a phosphorylation event alone is less constructive if the functional role of the phosphorylation event is not known. Much effort has gone into identifying molecular phenotypes associated with a particular cellular event. For instance, prioritizing functional PTMs and understanding their function [283], investigating the role of particular PPIs in disease [398], and identifying TFBSs that are likely to influence expression [399, 400]. Such studies are critical to aiding the interpretation of mechanisms predicted to be disrupted by genetic variation.

All in all, mutfunc is a unique resource that will greatly facilitate the identification of the molecular mechanisms altered by point mutations that lead to phenotypic differences and can be broadly applied to different model organisms.

# Chapter 5

# Gene-level aggregation of mechanistic variant impact for gene-phenotype associations

*In this chapter, I describe the use of mechanistic variant impact predictors to construct gene burden scores in a panel of 93 S. cerevisiae strains. Phenotypic screening of S. cerevisiae strains under 43 different conditions was carried out, further allowing for the testing of gene-phenotype associations. All analysis was carried out by my self, under the supervision of Pedro Beltrao. I was not involved in the generation of experimental data. Phenotypic screens were carried out by lab member Bede Busby and the processing and scoring of phenotyped data was carried out by lab member Marco Galardini.*

## 5.1 Introduction

Rare genetic variants extensively contribute to disease biology [401]. Yet, traditional GWASs is often unable to implicate rare variants in phenotypic differences primarily due to their low prevalence. Such associations would often require the genotyping of sufficiently large cohorts, which is in many cases is impractical or infeasible. For cases where sufficient data is available, the high number of statistical tests combined with stringent multiple testing correction often result in the dissipation of any signal. In addition, the design of GWASs revolves around genotyping chips for a set of tag variants, which rely on linkage-disequilibrium for imputation of rare variants from a reference panel like the HapMap project [402]. However, the marker variant must be observed in the reference panel in order for successful imputation, which may not always be the case. Much effort has gone into designing rare variant association

studies that involve whole genome and exome sequencing as well as rare variant genotyping, yet, these approaches focus on achieving a higher number of samples, which will not always guarantee sufficient improvement statistical power [403].

One way to tackle the lack of statistical power for association studies is to combine the effects of rare variants by predicting their impact on gene function or "gene burden". This, in turn, can be used to carry out gene-phenotype associations. Studies utilising gene burden for associations have employed different approaches to quantify the effect of rare variants. For instance, DeBoever et al. used 18,228 protein truncating variants (PTVs), including nonstop, nonsense and frameshift variants to generate gene burden scores which were associated with 135 phenotypes from the UK Biobank and uncovered 27 high confidence associations [404]. Iorio et al. utilised prior information on known cancer driver variants and copy number variants to define whether a gene is affected across 1,001 cancer cell lines and computed associations across drug response profiles for 265 drugs uncovering many known and novel associations [405]. Olde Loohuis et al. quantified gene burden by assessing deleteriousness of rare coding variants by assessing PTVs, splice-site variants and deleteriousness predictions by predictors like SIFT and PolyPhen. Through the analysis of rare variants in 1,042 schizophrenia patients against 961 controls, they were able to uncover many schizophrenia-associated genes under high burden [406]. While such approaches are effective at improving our understanding of rare variant impact, they commonly focus a single predictor or rely on previously identified pathogenic variants. Little has been done to comprehensively assess effects across multiple predictors. In addition, predictors of deleteriousness such as SIFT are unable to explain the altered biological mechanism.

The use of mechanistic variant impact predictors can significantly contribute to improving statistical power while shedding light on the altered mechanisms caused by rare variants beyond single variant-based testing. In this chapter, we aimed to test this using *S. cerevisiae* as a case study. We utilised coding variants from whole genome sequences for 93 *S. cerevisiae* strains and collect pre-computed mechanistic variant effects for protein stability (FoldX), conservation (SIFT) and PTVs to define gene burden scores (see section 4.2.2). We phenotyped corresponding growth profiles of 166 strains across 43 conditions including common drugs, nutrient stressors and environmental stressors. The resulting data was used to test gene-phenotype associations and uncover several known and novel associations. We further show that gene burden can be expanded to compute complex-level burden, which provides additional power to statistical tests.

## 5.2 Results

### 5.2.1 Phenotypic variation across *S. cerevisiae* strains

We first phenotyped growth for a panel of 166 *S. cerevisiae* strain in 43 conditions including common drugs (e.g. ketoconazole, benomyl, caspofungin, cisplatin, rapamycin), nutrient stressors (e.g. glucose, glycerol, sorbitol, amino acid deprivation, nitrogen starvation), and other environmental stressors (e.g. high heat, anaerobic conditions, UV light, sodium chloride). Colony sizes for strains were quantified using the IRIS software [407] then normalised and scored relative to all strains in a condition to produce a phenotypic measure defined as the S-score, where a positive value indicates higher growth and negative values indicate poorer growth (Methods, section 5.3.1). S-scores for biological replicates demonstrated a high degree of concordance ($r$ = 0.91, $p<2.22\times10^{-16}$, Figure 5.1a), demonstrating a high degree of confidence in phenotypic measurements.

Hierarchical clustering of growth phenotypes revealed known clusters of related stressors (Figure 5.1a). Clusters of similar phenotypic profiles included, (1) UV light, cisplatin and MMS, which are all DNA damaging agents (mean pairwise $r$ = 0.51) (2) nystatin and caspofungin, which act by interfering with the fungal cell membrane ($r$ = 0.49) (3) rotenone, DMSO and pH levels of 7.5-8.5 all inflict oxidative stress (mean pairwise $r$ = 0.46) (4) caffeine and rapamycin, both involved in multiple signalling pathways, namely that of TOR ($r$ = 0.41), and (5) 5-fluorouracil and 6-azauracil, which both act by altering nucleotide pool levels ultimately influencing transcriptional elongation ($r$ = 0.42). Furthermore, strains belonging to the same population structure [336] or environmental origin often showed similar phenotypic profiles (Figure 5.1b).

Since strains belonging to the same population structure typically arise from a common ancestor, we asked if strains with similar SNP profiles also show similar phenotypic trends. We collected variants for 56% (93/166) of phenotyped strains from Strope et al. [336] (Methods, section 5.3.2) and computed SNP profile distances and phenotypic distances for all 4,278 pairs of strains using the Euclidean distance measure. We found that strains extremely similar in their SNP profiles were also phenotypically similar (Figure 5.1c). However, at increased genotype distance, pairs of strains exhibit a much higher variability in the phenotypic distance. This suggests that predicting phenotypic distance amongst strains displaying heterogeneous genotypes is a non-trivial task.

The screened phenotypes presented here, along with collected variants for 93 strains serve as a useful starting point for performing subsequent gene-phenotype associations.

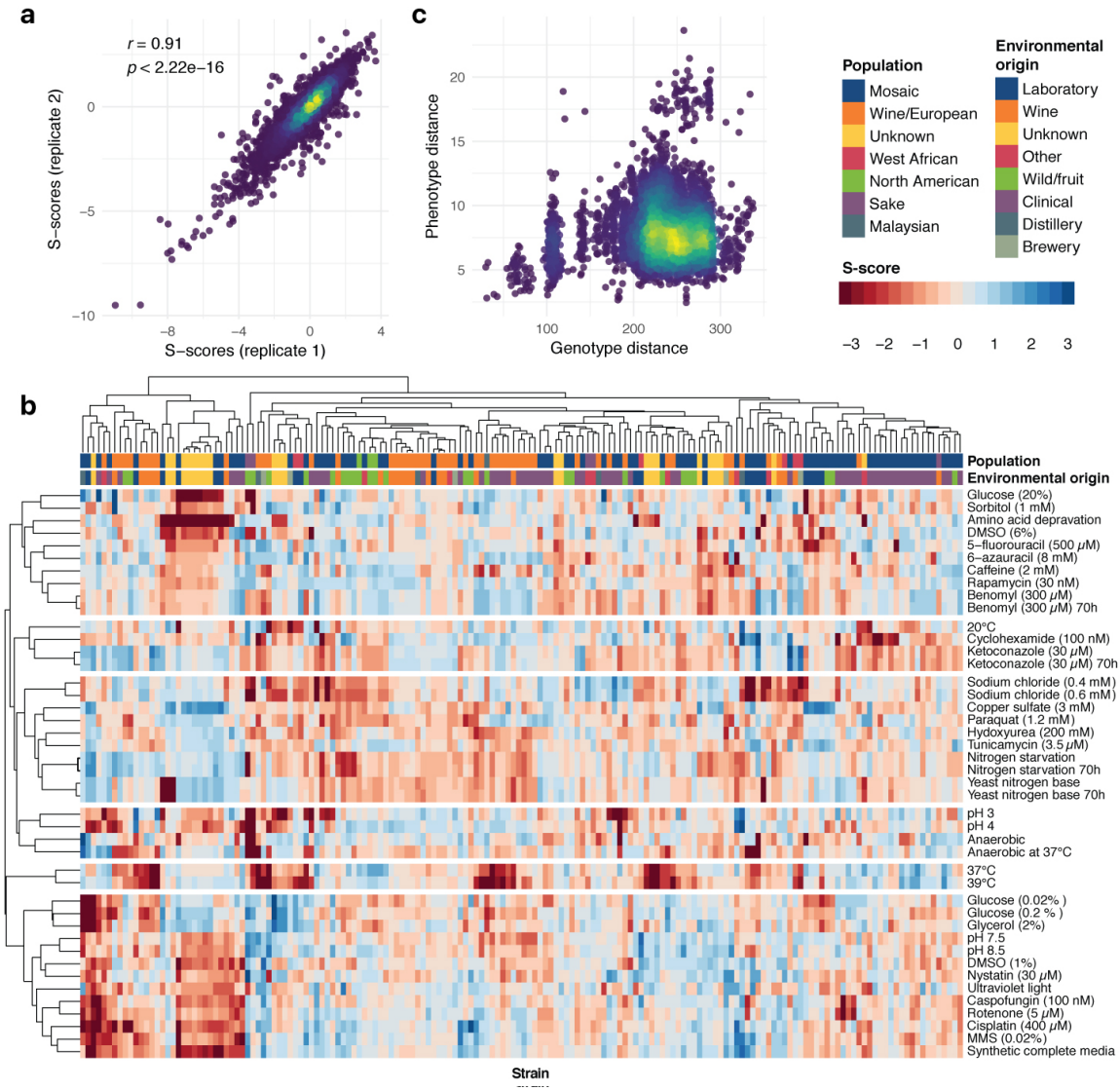



Fig. 5.1 Phenotypic screening of 166 *S. cerevisiae* strains. (a) Concordance between S-score measurements between two biological replicates. (b) Heatmap of S-scores showing hierarchical clustering of both strains and conditions reveals clusters of phenotypically similar strains and conditions. (c) Comparison of pairwise genotype and phenotype distances between strains shows little observable correlation.

### 5.2.2 Mechanistic gene burden scores identify novel gene-phenotype associations

We next sought to define gene-level burden scores to aid the interpretation of phenotypic variability amongst strains. To compute gene-level burden scores for a given protein, we utilised mechanistic predictions for conservation (SIFT), protein stability (FoldX) and protein truncating variants (PTVs, including start loss, nonstop and nonsense variants) (Figure 5.2a).

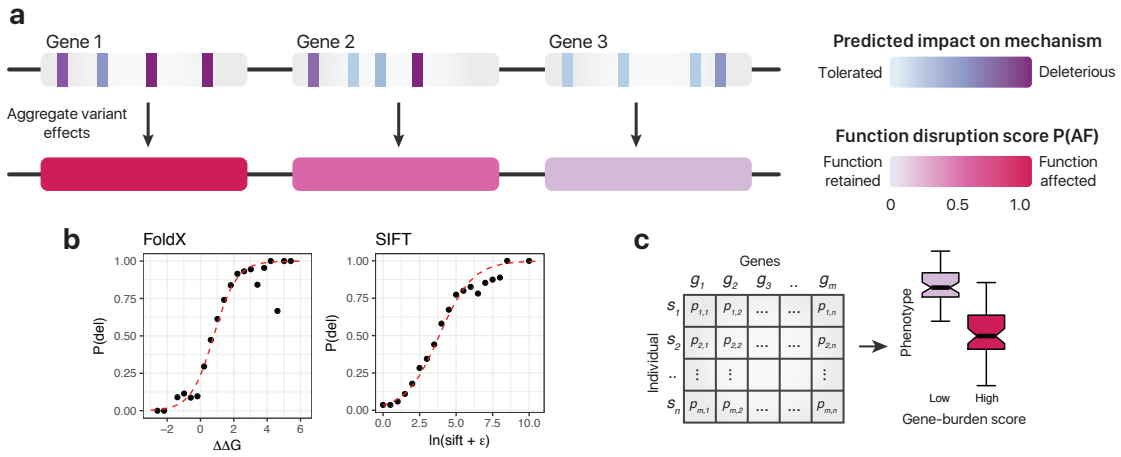



Fig. 5.2 Gene-level aggregation of variant effects. (a) Diagram demonstrating the aggregation of variant impact. Each variant is first assigned a probability of deleteriousness, which is aggregated at the gene level using the maximum impact. (b) The probability of deleteriousness for FoldX and SIFT is computed by assessing the proportion of deleterious variants in gold-standard data for FoldX and SIFT. A logistic regression model (red line) is fit to compute subsequent probabilities. (c) Once gene burden is computed for each gene and strain, gene-phenotype associations can be carried out by comparing growth of strains containing high and low $P_{AF}$ scores for a particular gene.

We first standardise the variant impact predictions for each of the mechanistic predictors. Scores produced by predictors are recalibrated to reflect the likelihood they are deleterious ($P_{del}$). For SIFT, a curated gold standard set of 8,083 variants in 1,346 *S. cerevisiae* genes with known tolerated or deleterious effects were obtained from Jelier et al. [355]. The negative natural logarithm of the SIFT score was binned by 0.5 and for each bin, the proportion of deleterious variants was computed. A binomial logistic regression was fit to the proportion values and used to compute subsequent $P_{del}$ values for subsequent SIFT scores. For FoldX, 964 gold-standard mutations across 34 experimentally identified proteins structures with both experimentally quantified $\Delta\Delta G$ values and FoldX-predicted $\Delta\Delta G$ values were obtained from Guerois et al. [408]. A variant was labelled destabilising if $\Delta\Delta G$ was greater than 1. Mutations were binned by predicted $\Delta\Delta G$ at intervals of 0.4 and for each bin, the proportion of destabilising variants was computed. A binomial logistic regression model was similarly fit to the data and used to compute subsequent $P_{del}$ for FoldX-predicted $\Delta\Delta G$ values (Figure 5.2b). For PTVs, we resorted to using heuristics to define $P_{del}$ values. Variants disrupting start or stop codons were assigned a value of 1. Since nonsense variants occurring closer to the C-terminal of a protein are less likely to impact function, we only assign $P_{del}$ value of 1 for nonsense variants occurring in the first 50% of the protein, otherwise a value

of 0 is assigned. Gene burden scores are then computed as the variant with the maximum $P_{del}$ score and describes the predicted likelihood that a protein has an affected function ($P_{AF}$). This allows for effects of rare variants to be combined across different protein positions and predictors, which can then be used to identify gene-level phenotype associations by comparing phenotypic readout for strains with high gene burden compared to those with low gene burden (Figure 5.2c).

Using natural variation data for the 93 strains, we computed $P_{AF}$ scores for all genes. If a gene did not contain any coding variants, a $P_{AF}$ score of 0 is assigned. Scores were binned based on high ($P_{AF} > 0.90$) or low ($P_{AF} < 0.90$) gene burden. Using a linear model, associations were carried out for 1,446 genes (with at least three strains containing a $P_{AF} > 0.90$) against growth phenotypes across 43 conditions (Methods, section 5.3.1). All reported $p$-values were corrected using the false discovery rate (FDR) method. In addition to statistical significance, to ensure sufficient magnitude in growth phenotypes, we compute the effect size using the Glass' $\Delta$ approach. Here, change in mean values relative to the standard deviation of one group is measured. We compute two Glass $\Delta$ values, relative to both groups and report the minimal absolute $\Delta$, signed by the direction of the association (Methods, section 5.3.4).

We identified a total of 626 statistically significant gene-phenotype associations at ($p < 1 \times 10^{-3}$ and FDR < 10%, Figure 5.3a). A total of 83% (520/626) of are negative associations i.e. decreased growth, and 17% (106/626) were positive. To validate associations we utilised chemical genetic data where genes are knocked out in the reference strain and allowed to grow under various conditions. The growth defect for a knock out can then be used to associate a gene with a particular condition. Although these experiments are carried out in the reference strain and conditions rarely match in concentration, this data provides a useful starting point to systematically validate associations. We collected genes associated with 35/43 of the assayed conditions using data from high-throughput chemical genetic screens [409] as well as literature-curated cases from the *Saccharomyces* Genome Database (Methods, section 5.3.3). Of all significant negative associations, a total of 9% (48/520) are validated by the chemical genetic data. It is also important to note that not all gene-phenotype association have been tested in the chemical genetic data, and thus this number could, in theory, be higher. Increasing the effect size threshold shows an increasing trend in the proportion of validated associations (Figure 5.3b). At $\Delta > 1$, 13% (38/282) are validated and at $\Delta > 1.7$, 24% (15/64) are validated, suggesting that associations of higher effect sizes are more reliable. To assess whether these values are obtained by chance, we sampled the same number of genes from the pool of 1,446 tested genes at each of the effect thresholds and similarly measured the proportion of these gene-condition pairs observed in the chemical genetic data. This was repeated 1,000 times for each threshold. We found that

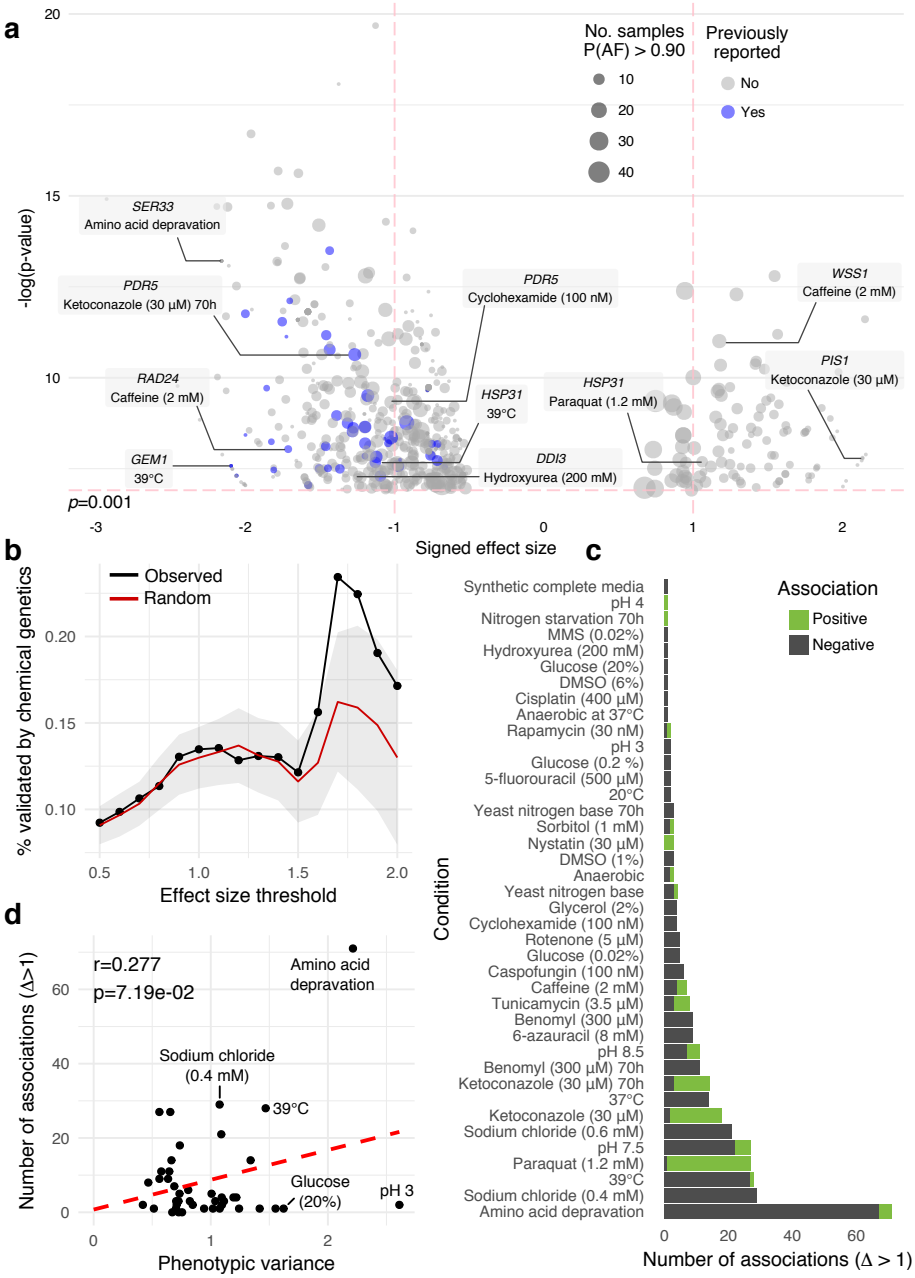



Fig. 5.3 Significant identified associations. (a) A volcano plot showing significant associations. The size of each point is proportional to the number of strains containing a high $P_{AF}$ score. Associations validated by chemical genetics are coloured in purple. (b) The proportion of associations validated by chemical genetics at different effect size thresholds for observed (black) and randomly permutated (red) associations. Grey ribbon represents the one standard deviation. (c) The number of associations across different conditions where positive and negative associations are shown in green and dark grey, respectively. (d) Phenotypic variance across strains compared against the number of significant associations.

at lower effect sizes expectation was near random. However, at larger effect sizes ($\Delta > 1.7$), observed proportions were significantly higher ($p$=0.03, Figure 5.3b).

We next checked if larger phenotypic variability for a condition results in a higher number of identified associations. However, no strong correlation was observed between growth variability and the number of identified associations at $\Delta > 1$ ($r$ = 0.27, $p$=$7.19\times10^{-2}$, Figure 5.3c-d). Specifically, conditions such as high heat, sodium chloride and amino acid deprivation showed high phenotypic variability and a high number of identified associations. However, many conditions that showed high phenotypic variability showed very few associations, such as low pH ($n$=2) and high glucose ($n$=1). Interestingly, several conditions including paraquat, ketoconazole and nystatin seemed to explain the majority of identified positive associations suggesting that coding variation are advantageous in such conditions. Why this is the case remains unclear.

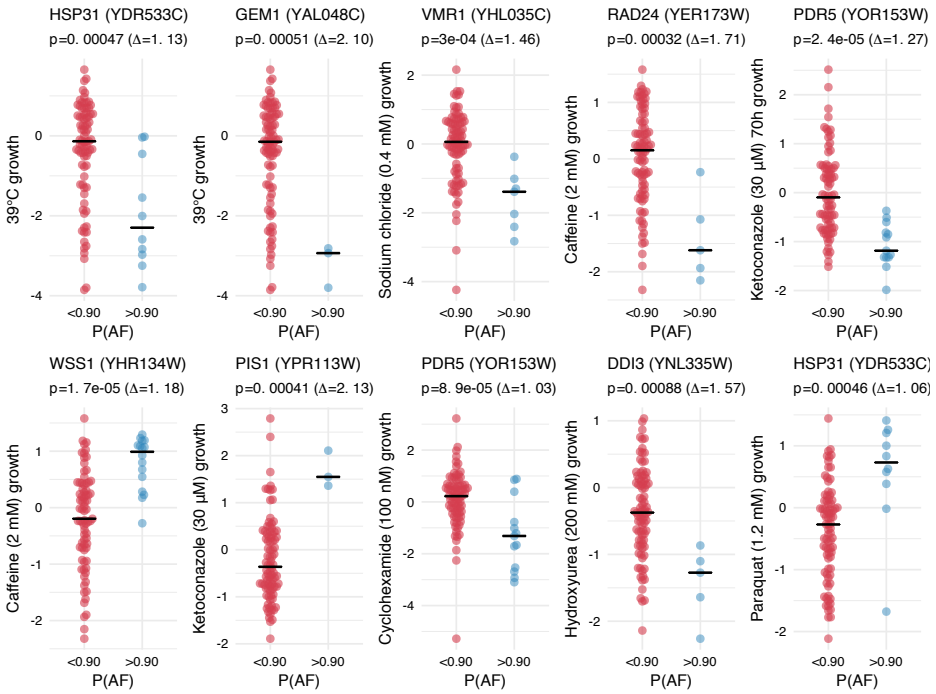



Fig. 5.4 Examples of high confidence associations. Each example shows the S-score of the negative (red) and positive (blue) group in a given condition. Horizontal black lines represent the median S-score.

To explore resulting associations we focus on a subset of 366 associations with a high effect size ($|\Delta| > 1$, Figure 5.4). For instance, high heat associated with mutations in the heat shock protein *HSP31* (Δ=-1.13, $p$=$4.65\times10^{-4}$), outer mitochondrial membrane GTPase *GEM1* (Δ=-2.09, $p$=$5.12\times10^{-4}$) and acetyl-CoA carboxylase *ACC1* (Δ=-1.51, $p$=$5.49\times10^{-4}$), all of which when knocked out in the reference strain result in growth

defects under high heat [410–412]. Nutrient stressors like sodium chloride showed significant associations to the ABC transporter VMR1 ($\Delta$=-1.46, $p$=$2.99\times10^{-4}$). Amino acid deprivation showed significant associations with numerous genes involved in the uptake, catabolism, and biosynthesis of amino acids including *BAP2* ($\Delta$=-1.58, $p$=$7.40\times10^{-6}$), *PDC6* ($\Delta$=-1.53, $p$=$6.02\times10^{-6}$), *ARO9* ($\Delta$=-1.49, $p$=$8.46\times10^{-6}$) and *SER33* ($\Delta$=-2.15, $p$=$1.83\times10^{-6}$). Common drugs like Caffeine, which is involved in cell cycle arrest and DNA damage [413, 414] showed associations to genes like the cell cycle checkpoint protein *RAD24* ($\Delta$=-1.71, $p$=$3.23\times10^{-4}$) and the SUMO-ligase *WSS1* ($\Delta$=1.17, $p$=$1.65\times10^{-5}$) which is involved in DNA repair. The anti-fungal drug ketoconazole interferes with ergosterol synthesis, thereby disrupting the cell membrane. The phosphatidylinositol synthase *PIS1* is key for biosynthesis of cell membrane polyphosphoinositides, and strains carrying impactful mutations show stronger growth in ketoconazole ($\Delta$=2.13, $p$=$4.08\times10^{-4}$). The ATP-binding multi-drug resistance transporter *PDR5* is another gene negatively associated with ketoconazole ($\Delta$=-1.26, $p$=$2.40\times10^{-5}$) as well as other drugs including 6-azauracil ($\Delta$=-1.08, $p$=$3.78\times10^{-4}$) and cycloheximide ($\Delta$=-1.02, $p$=$8.91\times10^{-5}$). Given its general role in drug resistance [415], it is appropriate that mutations disrupting gene function would result in such growth defects. Hydroxyurea is a drug that arrests DNA replication and is involved with DNA damage and is associated with the DNA damage inducible protein *DDI3* ($\Delta$=-1.57, $p$=$8.76\times10^{-4}$), which is over-expressed 100-fold by DNA damaging agents [416]. Paraquat is a drug that induces oxidative stress by interfering with the electron transport chain. We find the heat shock protein *HSP31* strongly positively associated with paraquat ($\Delta$=1.06 $p$=$4.57\times10^{-4}$). Heat shock proteins have been previously identified to suppress paraquat-induced effects in rat and *H. sapiens* [417, 418] suggesting the existence of a similar mechanism in *S. cerevisiae*. Other genes positively associated with paraquat include reductases *YJR096W* ($\Delta$=1.43, $p$=$1.32\times10^{-4}$) and *LYS2* ($\Delta$=1.75, $p$=$2.01\times10^{-4}$) as well as the hydrolase *YSA1* ($\Delta$=1.42, $p$=$9.26\times10^{-4}$, Figure 5.4).

Investigation of variants responsible for associations reveals the collective impact rare variants have on a gene. For instance, the *GLN4*-heat association is driven by destabilising rare mutations in glutaminyl-tRNA synthetase domain (Figure 5.5a) whereas the *ACC1*-heat association is driven by both destabilising and conservation-affecting variants in the biotin carboxylase domain (Figure 5.5b). The *VMR1*-sodium chloride association is driven by two nonsense variants in the ABC transmembrane domain and conservation-affecting variants in and around the ABC transporter domain (Figure 5.5c). The *RAD24*-caffeine association is driven by three conservation-affecting mutations within the Rad17-like domain (Figure 5.5d). The *DDI3*-hydroxyurea association driven by a start loss variant and two

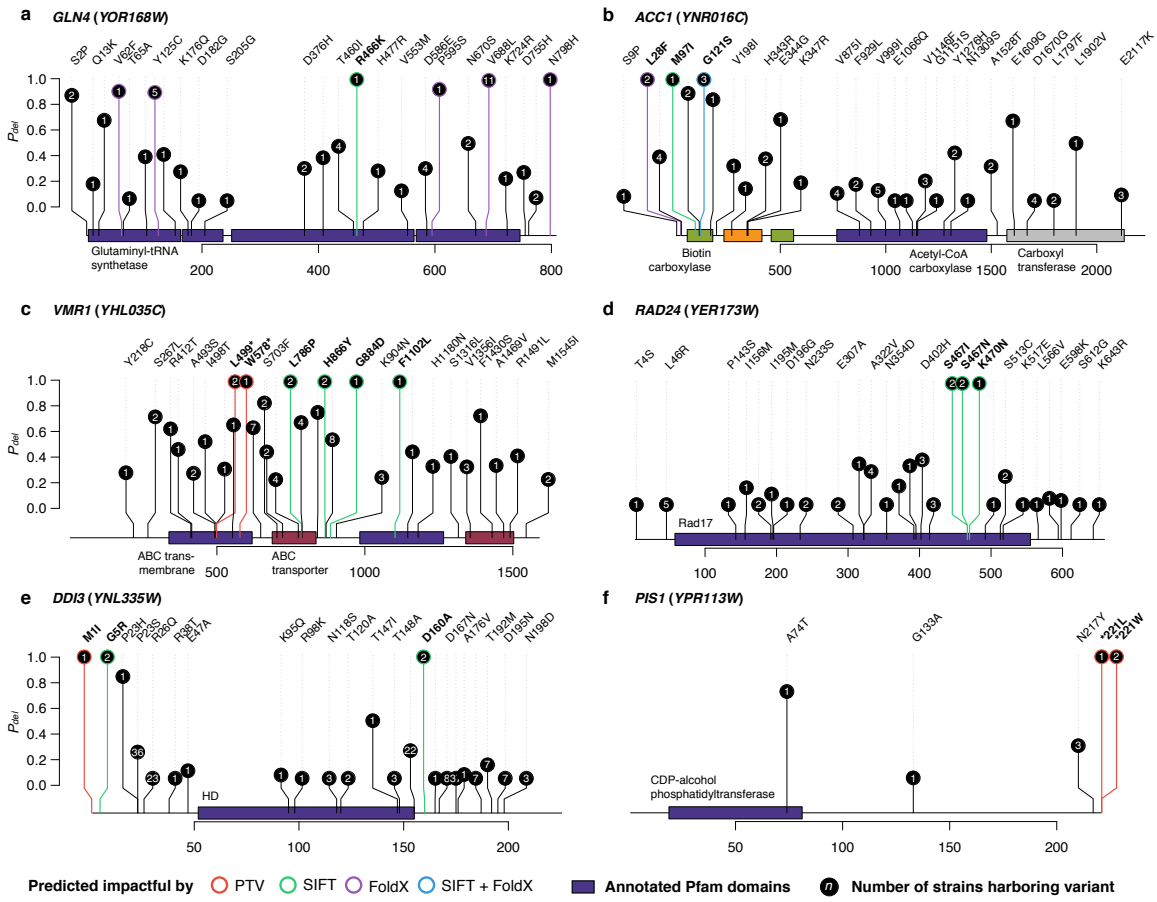



Fig. 5.5 (a-f) Variant impact plots for six genes from figure 5.4. Coloured boxes represent annotated Pfam [350] domains. The height of mutations reflect the $P_{del}$ score and the value indicated in white denotes the number of strains harbouring the variant. Variants predicted as deleterious ($P_{del} > 0.90$) show different coloured outlines depending on the predictor: PTV (red), SIFT (green), FoldX (purple) and both SIFT and FoldX (blue).

conservation-affecting variants. Finally, the *PIS1*-ketoconazole association is driven by two nonstop variants.

The suggested gene burden approach has proven valuable at revealing many known and novel gene-phenotype associations through combining effects of rare variants to increase statistical power.

### 5.2.3 Complex burden scores further improve association power

Genes often form complexes that carry out the majority of cellular processes. Altering any constituent parts of a complex could, therefore, result in a similar phenotypic outcome. Assessing burden on a complex level could thus provide additional power when conducting

associations. To test this, we collected gene sets for 408 *S. cerevisiae* complexes from the CYC2008 resource [419]. The complex-level burden was computed by taking the maximum $P_{AF}$ score across all genes in a complex (Methods, section 5.3.4). For 258 complexes that that contained at least two high $P_{AF}$ genes across strains, associations were carried out against the 43 conditions similarly to that of gene-level associations.

We identified 75 significant complex-phenotype associations ($p<1\times10^{-3}$, FDR < 10%), 25 (33%) of which had a high effect size ($\Delta > 1$). The 25 associations involved 15 conditions and were constituted of 17 (68%) negative associations and 8 (32%) positive associations. Complex-level burden scores were able to uncover associations not possible on the gene-level. For instance, high heat was shown to be associated with the FBP complex responsible for protein degradation ($\Delta$=1.01 $p$=$5.7\times10^{-4}$, Figure 5.6a). Members of the FBP complex with affected function ($P_{AF}>0.90$) *VID24*, *GID7*, *FYV10* and *RMD5* are not detected by gene-level burden associations, yet 3/4 show heat sensitivity within chemical genetic data [411]. The drug 5-fluorouracil suppresses DNA replication by blocking synthesis of the pyrimidine nucleotide thymidine and significantly associates with the guanyl-nucleotide exchange factor complex ($\Delta$=1.25 $p$=$7.4\times10^{-4}$, Figure 5.6b). Benomyl interferes with microtubule stability and associates with the AP-2 adapter complex, which is responsible for clathrin-mediated endocytosis (Figure 5.6c). Clatherin is also involved in stabilising microtubules that attach to kinetochores during meiosis [420]. Similarly, the AP-1 adaptor complex associates with ketoconazole ($\Delta$=1.61 $p$=$2.2\times10^{-4}$, Figure 5.6d). Since ketoconazole alters cell membrane by interfering with ergosterol biosynthesis and ergosterol play key roles in endocytosis, this potentially explains the link between mutations in members of the AP-1 adaptor complex and ketoconazole. The transcription factor TFIIH complex positively associates with ketoconazole ($\Delta$=2.03 $p$=$1.1\times10^{-4}$, Figure 5.6e), which can possibly be explained by the transcriptional repression and activating functions of ketoconazole. Lastly, the DNA replication factor C positively associates with ultraviolet light ($\Delta$=1.26 $p$=$8.2\times10^{-4}$, Figure 5.6f), which could be explained by the DNA-damaging properties of UV light.

Complex-level burden provides additional power when testing gene-phenotype associations, particularly when the sample size is limited. Associations performed on the pathway or domain level may also provide additional insight into novel associations.

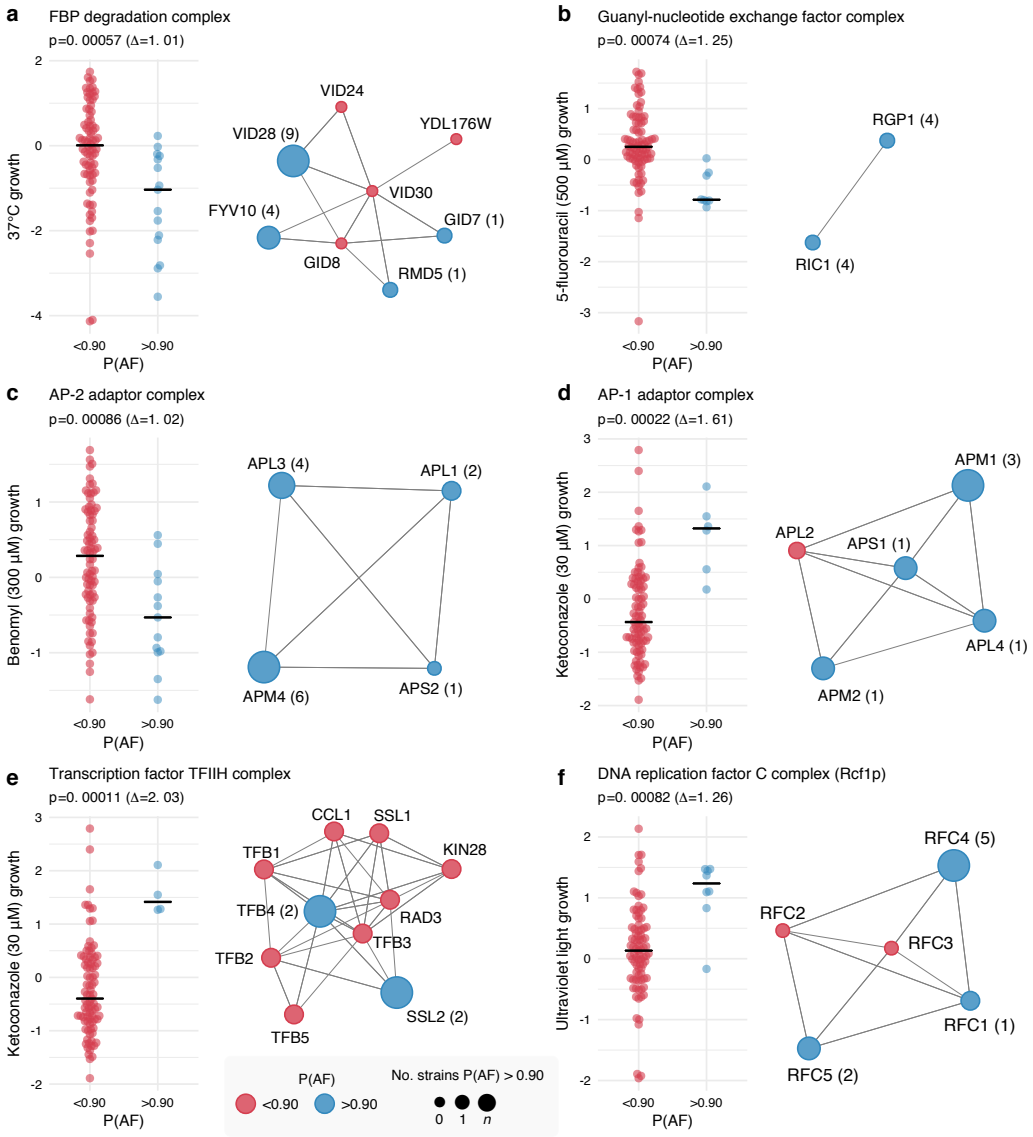



Fig. 5.6 Aggregating effects on a complex level. (a-f) Examples of complex level associations. For each example, the growth of strains predicted to have at least one complex member altered (blue) is compared against those that do not (red). Complex members are shown connected by an edge if they are known in the BioGRID database [261] to physically interact. The values indicated on node labels denote the number of strains with a high $P_{AF}$ score for that gene.

## 5.3 Methods

### 5.3.1 Phenotyping of *S. cerevisiae* strains

Phenotyping for 166 *S. cerevisiae* strains across 43 conditions was performed by lab member Bede Busby. The screening was carried out in 1536 format on synthetic complete media with the addition of the appropriate chemical at a specific concentration. The Singer RoToR (Singer instruments, UK) was used to replicate screening plates in 1536 format. Agar plates were pinned onto the conditioned media in quadruples and allowed to grow for 48 or 72 hours at 30 degrees centigrade (unless specified otherwise). A total of four technical replicates were carried out. For each technical replicate, up to 12 biological replicates were carried out.

After incubation, plates were imaged and the processing of plate images was carried out by lab member Marco Galardini. Colony sizes were extracted using IRIS version v0.9.7 [407] with the "Colony growth" profile, which extracts colony size, circularity and opacity from each colony in each plate. Individual strains were scored using the E-MAP software, which transforms colony sizes into S-scores [421]. In brief, a surface correction algorithm is applied to each plate, the outer frame effect is corrected by bringing the two outermost rows and columns to the plate middle median. All the plates are then normalized to the overall median, followed by a variance correction (Figure 5.7a). Finally, the S-score is calculated based on a modified t-test, as defined by Collins et al. [421]:

$$t = \frac{\hat{\mu} - \hat{\mu_0}}{\sqrt{var(\hat{\mu}) + var(\hat{\mu_0})}}. \tag{5.1}$$

Here, $\hat{\mu}$ and $\hat{\mu_0}$ are the median observed and expected colony sizes across technical replicates, respectively. Expected colony sizes are computed as growth for a strain across all conditions. The resulting S-scores are quantile normalized in each condition separately (Figure 5.7b). Final S-scores are then computed by averaging S-scores of biological replicates.

### 5.3.2 Genetic variants for *S. cerevisiae* strains

Genetic variants for the 93 strains analysed here were called from whole genome reads obtained from Strope et al. [336]. Variants with a MAF > 20% were discarded. Additional information on data collection and variant calling can be found in section 4.3.1.

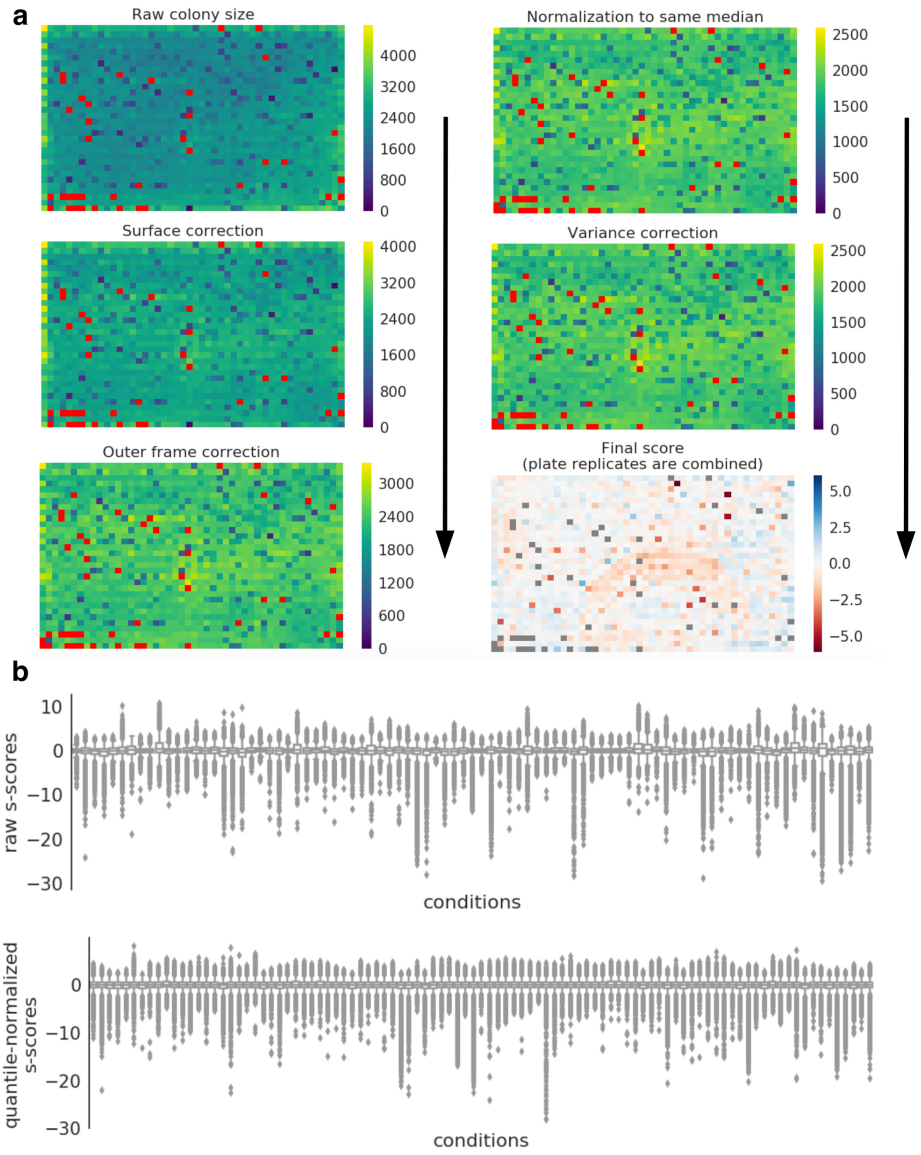



Fig. 5.7 S-score calculation and normalisation. (a) Normalisation of raw quantified colony sizes and calculation of the S-score. (b) Quantile normalization of S-scores.

### 5.3.3 Chemical genetic data

Chemical genetic data for heterozygous and homozygous knockout reference strains were obtained from Hillenmeyer et al [409] for 5,900 genes across 568 conditions. The files `hom.z_tdist_pval_nm.pub` and `het.z_tdist_pval_nm.goodbatch.pub` were used. Conditions were manually matched to assayed conditions in section 5.3.1 and genes were considered associated with the condition if their $p$-value was below $1 \times 10^{-5}$, as recommended by the authors. For cases where data for multiple generations or replicates are

reported or the phenotype is observed for both heterozygous and homozygous knockdowns, the association with the most significant $p$-value is retained. The final set contained 7,944 gene-phenotype associations across 5,883 genes and 28 conditions.

Chemical genetic data for null strain mutants from various other high-throughput and low-throughout studies was obtained from the *Saccharomyces* Genome Database [422] through the `phenotype_data_sgd.tab` file. The "phenotype" and "chemical" columns were manually matched to assayed conditions section 5.3.1 resulting in a total of 15,194 gene-phenotype associations across 3,508 genes and 30 conditions.

### 5.3.4 Computing gene and complex-burden scores and associations

For a gene $g$, given a list of $n$ variants $v_1, \cdots, v_n$ with $P_{del}$ scores of $P_{del}^{v_1}, \cdots, P_{del}^{v_n}$ the gene-level burden scores are computed as:

$$P_{AF} = \max_{1 \leq i \leq n} P_{del}^{i} \tag{5.2}$$

Similarly, given a set of $k$ genes $g_1, ..., g_k$ in a complex with $P_{AF}$ scores of $P_{AF}^{g_1}, \cdots, P_{AF}^{g_k}$, complex-burden scores are computed as:

$$P_{CAF} = \max_{1 \leq i \leq k} P_{AF}^{g_i} \tag{5.3}$$

All associations were carried out using the MatrixEQTL R package [423] with the `modelLINEAR` mode. The package was developed for rapidly conducting hundreds of thousands of associations between variants and gene expression, although the methods are generally applicable. The significance of the association is then measured using a $t$-statistic. Here, binarised $P_{AF}$ scores are used as genotypes where a $P_{AF}$ score above or below 0.9 is given a value 1 and 0, respectively and growth phenotypes are used in lieu of gene expression. A $p$-value threshold of 0.001 is used for all associations and multiple testing correction is carried out using the false discovery rate method.

The effect size was computed using Glass' $\Delta$. For the case ($p$) and control ($n$) group, differences in the mean is computed relative to the standard deviation of one of the groups. Given the mean ($\mu_i$) and standard deviation ($\sigma_i$) for a given group $i$ this is computed as:

$$\Delta_i = \frac{\mu_p - \mu_n}{\sigma_i} \tag{5.4}$$

To ensure sufficient effect size in either direction, this is computed in either direction and the final effect size, $\Delta$, is reported as the minimum absolute value of effect sizes in both directions, signed by the direction of the association:

$$\Delta = \varepsilon \cdot min\{\Delta_p, \Delta_n\}$$
$$\varepsilon = \begin{cases} -1, & \text{if } \mu_p < \mu_n \\ 1, & \text{otherwise} \end{cases} \tag{5.5}$$

## 5.4 Discussion

In this chapter, we describe an approach to combine effects of PTVs, SIFT-predicted deleteriousness and FoldX-predicted stability impact to collectively define a gene burden score.

While the maximum predicted variant impact is used here, there exist several other approaches at computing gene burden. For instance, Jelier et al. additively modelled gene burden by taking the product of probabilities that a variant is neutral [355]. Although it is true that the impact of variants accumulate additively, the measure of deleteriousness ($P_{del}$) as defined here does not directly reflect the intrinsic probability of deleteriousness. For instance, a $P_{del}$ score of 0.3 that is based on a protein stability predictor will not always indicate there is 30% chance of this variant impacting function. This would result in the inflation of gene $P_{AF}$ scores should the gene harbour numerous mutations with mild $P_{del}$ scores. Benchmarking which gene burden scoring metric is appropriate remains a challenging task due to the lack of gold-standard gene-phenotype associations. Chemical genetic screens available for *S. cerevisiae* is a potentially useful resource, yet, there are several inherent limitations. First, the chemical genetic data is often carried out in a reference strain which may not always reflect individual-specific growth behaviour. Second, the conditions in which these screens are carried out do not necessarily match growth phenotypes used to conduct the associations. These limitations would not allow us to accurately detect false positive associations. Approaches such as individual-specific gene knock-in experiments, where the endogenous gene is replaced with a wildtype copy of the reference strain gene, would provide a valuable way by which the accuracy gene burden scores could be measured.

There are several notable drawbacks to the use of *in silico* variant impact predictors for gene burden. First and foremost, there is the issue of false positives. Predictors used here measure variant impact independently relative to the reference individual. However, the genetic context of an individual has been shown to play significant roles in dictating variant impact. For instance, phenomena like epistatic interactions between variants commonly occur, where the occurrence of a variant can suppress or aggravate effects another variant [424]. Commonly used variant impact predictors such as SIFT and FoldX were not designed to account for such context-specific effects, resulting in the misclassification of many variants.

Second, predictors provide no information on whether an impactful variant confers gain or loss-of-function. Such information would improve both statistical power and interpretation of gene burden association tests. Future improvement of variant impact predictors will, therefore, be paramount to gene burden association tests.

In addition to predictors, there are several limitations with respect to the associations carried out. First, although the 93 strains utilised here are geographically and environmentally heterogeneous [336], it is possible that sub-populations of strains exist. This population structure could thus be a potential confounding factor. Strains that belong to the same sub-population are likely to exhibit similar phenotypes and therefore many identified associations may be due to strains belonging to the same sub-population. Second, the number of strains utilised here limits the number of genes that can be tested: for the 93 strains, we were able to perform associations for about 25% of the *S. cerevisiae* genome (1,446) that had at least three high $P_{AF}$ scores. Relative to *H. sapiens*, the number sequenced *S. cerevisiae* strains still remains low at 500 strains, and even fewer with sufficient phenotypic growth data [336–339]. A larger number of strains would provide additional variation and allow for associations to be carried out for a larger number of genes and allow for testing associations within sub-populations. Lastly, genes that do not harbour any genetic variation are assigned a $P_{AF}$ score of 0, which may not always be accurate since there are other types of mutations beyond SNVs, including indels, frameshift mutations, and copy number variations (CNVs) that were not accounted for here. The inclusion of such variation in the future could further improve the calculation of gene burden scores. Lastly, strains that are more distant from the reference strain are more likely to have suppressed effects of deleterious variants. Thus, approaches that are able to generate strain-specific variant impact scores or normalise against distance from the reference strain would allow for more accurate individual $P_{AF}$ scores.

Although there are several aspects of the described approach that could be improved, this study offers initial insight into the use of mechanistic variant impact, particularly protein stability, in the calculation of gene burden scores. Incorporation of additional mechanisms in the future will allow for us to comprehensively and accurately carry out hypothesis-driven gene burden associations that can be traced back to individual variants for which we could then mechanistically explain. Future expansion to human would offer unprecedented insight into disease risk and potential therapeutics.

# Chapter 6

# Summary and future directions

The arrival of next-generation sequencing technologies has brought about an abundance of individual genome data and a new era of genetics. The analysis of genomic data is, however, bottlenecked by variant interpretation, where an individual's genetic variants can be classified as pathogenic or benign. This has prompted an out pour of bioinformatic tools aimed at aiding variant interpretation. The abundance of corresponding molecular and organism-level phenotype data corresponding to an individual genome has also driven many association-based studies as a means of identifying causal variants. These approaches have been routinely applied to large-scale datasets in both human and model organisms to both guide clinical validation of variants and aid drug development [425, 426]. Uncovering the biology behind genetic variants will, therefore, have many implications in personalised therapies and drug development

The purpose of this thesis was to explore cellular mechanisms that are regularly influenced by genetic variation. A significant portion of this thesis discusses the role of sequence specificity-mediated interactions, specifically TF binding and kinase-substrate phosphorylation. Because these interactions are mediated by short motifs, they play a key role in variant interpretation. In Chapter 2, we discussed a computational approach that was developed to predict kinase-substrate specificity without any prior information on kinase-specific target sites. By leveraging functional interaction data and the abundance of available phosphorylation data, we were able to uncover predicted specificities for over half the human kinases. We have also shown that this approach can also easily be expanded to other PTM-binding domains. Since kinase target sites are frequently mutated in disease, and specifically cancer [148], the ability to uncover additional specificities allows for a better understanding of how phosphorylation is altered in disease. Much like kinase-substrate phosphorylation, TF binding also depends on DNA sequence specificity. In Chapter 3, we explored five computational methods used for the modelling of TF-binding and assessing the variant impact on

TFBSs. Using variants known to alter TF-binding from allele-specific ChIP-seq data, we systematically assessed the performance of specificity models in the five methods at predicting variant impact across over 40 TFs. We defined and compare the performance several variant impact scoring metrics across the methods and show that machine learning-based significantly outperform the PWM, which is commonly used to assess variant impact [427–429]. We highlight differences in performance across different TFs and explore alternative mechanisms that may contribute to the inability to assess variant impact using sequence specificity information alone.

These methods, including others, have the ability to shed light on altered mechanisms caused by genetic variants. In Chapter 4, we compile predictors for mechanistic variant impact, including protein stability, PPIs interfaces, PTMs, kinase-substrate phosphorylation, TF-binding, short linear motifs, and start and stop codons. We use these tools to precompute variant impact predictions that we provide through the mutfunc web server. This interactive tool allows for rapid annotation and prioritisation of variants in a mechanistic light, without requiring the cumbersome set up of any of the individual predictors. We validate the predictions generated by analysing natural and disease variants in human and yeast genomes. We show that variants altering mechanisms are more likely rare than common and that they are depleted in essential genes. We also analysed known pathogenic variants to show that altered mechanisms are enriched in pathogenic variants and show that these mechanistic variant impact predictions can be used to shed light on clinical variants of uncertain significance.

Identifying variants associated with phenotypic trait differences using traditional GWAS requires the variant to be, to a certain degree, prevalent in the population. Since many disease phenotypes are driven by rare genetic variants, identifying such associations with traditional variant-based association methods is not always feasible [430]. In Chapter 5 we utilised the predictions in mutfunc to predict mechanistic consequences of rare variants in a panel of 93 yeast strains and define gene burden scores. We carried out phenotypic screening for the corresponding strains in over 40 different conditions and employed this data in combination with gene burden scores to perform gene-phenotype associations, uncovering many known and novel associations. We demonstrate the added benefit of generating gene burden scores using mechanistic variant impact predictions over variant-level associations and show that this can be taken a step further by assessing complex-level burden.

Understanding the mechanistic basis for variants is very much an open question in the field of genetics and has fundamental roles variant interpretation. As such, there are numerous future avenues that can be taken. Specifically, with the advent of deep learning methods and increasing abundance of molecular-level phenotypic data, there exist many possibilities

to employ deep learning in modelling cellular mechanisms for which current approaches do not perform well at such as protein localisation [431], splicing [432], PTMs [433], and more. These models could, in turn, be used for variant impact prediction. Furthermore, high throughput experimental assaying of variant impact on protein function, such deep mutational scanning, are increasingly generating data that could be utilised for the training of variant impact predictors.

A primary drawback of many existing variant impact predictors is their inability to account for differences in genetic background. More specifically, they revolve around the assumption that genetic variants act individually when many diseases can be driven by the additive effect of numerous variants [434, 435]. The increased development of methods that can quantitatively model the cooperative impact of variants will allow for more accurate individual-specific variant impact predictions. Furthermore, additional factors such as epigenetics can greatly affect the deleteriousness of a variant [436]. Thus, the incorporation of additional context-specific *omics* data such as epigenomics, proteomics and metabolomics is fundamental to better understanding the role of such factors in cellular processes and accurately modelling tissue and cell type-specific variant impact.

Interpretation of variant impact predictors output is another key area of future research. As genetic testing becomes more prevalent in the clinical setting, the ability for healthcare professionals to be able to interpret and act on variant impact predictions is becoming more of a concern. The standardisation of predictor output is one way to aid variant interpretation and simplify the comparison of variant impact across multiple methods. Such standardisation would further streamline the incorporation of data from multiple predictors and improve the utility of variant impact predictions in phenotype association methods. Developed methods must also be user-friendly and intuitive to facilitate the construction of variant interpretation pipelines and allow use by a wide range of expert and non-expert users.

Ultimately, a paradigm shift in variant interpretation and GWAS is required whereby more light is shed on the affected biological mechanisms. To do this, our understanding of biology must be incorporated into assessing variant impact through the development of mechanistic variant impact predictors, which can be employed to assess higher order impact on gene, complex and pathway levels. Such a layering approach will be able to capture the propagation of variant impact through cellular processes and improve our ability to associate genetic variation as a whole to phenotypic differences.

# List of publications

## Scientific articles

1. Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*

2. Yuan Chen, Rhys A Farrer, Charles Giamberardino, Sharadha Sakthikumar, Alexander Jones, Timothy Yang, Jennifer L Tenor, Omar Wagih, Marelize Van Wyk, Nelesh P Govender, Thomas G Mitchell, Anastasia P Litvintseva, Christina A Cuomo, John R Perfect (2017). Microevolution of serial clinical isolates of *Cryptococcus neoformans var. grubii* and *C. gattii*. *MBio*, 8(2):e00166–17

3. Marco Galardini, Alexandra Koumoutsi, Lucia Herrera-Dominguez, Juan Antonio Cordero Varela, Anja Telzerow, Omar Wagih, Morgane Wartel, Olivier Clermont, Erick Denamur, Athanasios Typas, Pedro Beltrao (2017). Phenotype prediction in an *Escherichia coli* strain panel. *bioRxiv*, page 141879

4. Haruna Imamura, Omar Wagih, Tomoya Niinae, Naoyuki Sugiyama, Pedro Beltrao, Yasushi Ishihama (2017). Identification of putative PKA substrates with quantitative phosphoproteomics and primary-sequence-based scoring. *Journal of Proteome Research*, 16(4):1825–1830

5. Emanuel Goncçalves, Zrinka Raguz Nakic, Mattia Zampieri, Omar Wagih, David Ochoa, Uwe Sauer, Pedro Beltrao, Julio Saez-Rodriguez (2017). Systematic analysis of transcriptional and post-transcriptional regulation of metabolism in yeast. *PLoS computational biology*, 13(1):e1005297

6. Justin D Smith, Sundari Suresh, Ulrich Schlecht, Manhong Wu, Omar Wagih, Gary Peltz, Ronald W Davis, Lars M Steinmetz, Leopold Parts, Robert P St Onge (2016). Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design. *Genome biology*, 17(1):45

7. Ashwani Kumar, Natalia Beloglazova, Cedoljub Bundalovic-Torma, Sadhna Phanse, Viktor Deineko, Alla Gagarinova, Gabriel Musso, James Vlasblom, Sofia Lemak, Mohsen Hooshyar, Zoran Minic, Omar Wagih, Roberto Mosca, Patrick Aloy, Ashkan Golshani, John Parkinson, Andrew Emili, Alexander F Yakunin, Mohan Babu (2016). Conditional epistatic interaction maps reveal global functional rewiring of genome integrity pathways in *Escherichia coli*. *Cell reports*, 14(3):648–661

8. Omar Wagih, Naoyuki Sugiyama, Yasushi Ishihama, Pedro Beltrao (2016). Uncovering phosphorylation-based specificities through functional interaction networks. *Molecular & Cellular Proteomics*, 15(1):236–245

9. Jeffrey R Johnson, Silvia D Santos, Tasha Johnson, Ursula Pieper, Marta Strumillo, Omar Wagih, Andrej Sali, Nevan J Krogan, Pedro Beltrao (2015). Prediction of functionally important phospho-regulatory events in *Xenopus laevis* oocytes. *PLoS computational biology*, 11(8):e1004362

## Software tools and packages

1. ggseqlogo: a 'ggplot2' extension for drawing publication-ready sequence logos (https://cran.r-project.org/web/packages/ggseqlogo)

2. rmotifx: discovery of biological sequence motifs in R (https://github.com/omarwagih/rmotifx)

3. siftr: predicting the impact of mutations on protein function (https://github.com/omarwagih/siftr)

# References


[1] The Genotype-Tissue Expression (GTEx) project. *Nature genetics*, 45(6):580–5, Jun 2013.

[2] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–20, Oct 2013.

[3] Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, Jun 2007.

[4] S T Sherry, M H Ward, M Kholodov, J Baker, L Phan, E M Smigielski, and K Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–11, Jan 2001.

[5] Z Wang and J Moult. SNPs, protein structure, and disease. *Human mutation*, 17(4):263–70, Apr 2001.

[6] Valeria Faa', Alessandra Coiana, Federica Incani, Lucy Costantino, Antonio Cao, and Maria Cristina Rosatelli. A synonymous mutation in the CFTR gene causes aberrant splicing in an italian patient affected by a mild form of cystic fibrosis. *The Journal of molecular diagnostics : JMD*, 12(3):380–3, May 2010.

[7] Jubao Duan, Mark S Wainwright, Josep M Comeron, Naruya Saitou, Alan R Sanders, Joel Gelernter, and Pablo V Gejman. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Human molecular genetics*, 12(3):205–16, Feb 2003.

[8] Gong Zhang, Magdalena Hubalewska, and Zoya Ignatova. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature structural & molecular biology*, 16(3):274–80, Mar 2009.

[9] Zuben E Sauna and Chava Kimchi-Sarfaty. Understanding the contribution of synonymous mutations to human disease. *Nature reviews. Genetics*, 12(10):683–91, Aug 2011.

[10] Yumi Yamaguchi-Kabata, Makoto K Shimada, Yosuke Hayakawa, Shinsei Minoshima, Ranajit Chakraborty, Takashi Gojobori, and Tadashi Imanishi. Distribution and effects of nonsense polymorphisms in human genes. *PloS one*, 3(10):e3393, 2008.

[11] A Beaudet, A Bowcock, M Buchwald, L Cavalli-Sforza, M Farrall, M C King, K Klinger, J M Lalouel, G Lathrop, and S Naylor. Linkage of cystic fibrosis to two tightly linked DNA markers: joint report from a collaborative study. *American journal of human genetics*, 39(6):681–93, Dec 1986.

[12] A C Jones, C E Daniells, R G Snell, M Tachataki, S A Idziaszczyk, M Krawczak, J R Sampson, and J P Cheadle. Molecular genetic and phenotypic analysis reveals differences between TSC1 and TSC2 associated familial and sporadic tuberous sclerosis. *Human molecular genetics*, 6(12):2155–61, Nov 1997.

[13] Pallav Bhatnagar, Shirley Purvis, Emily Barron-Casella, Michael R DeBaun, James F Casella, Dan E Arking, and Jeffrey R Keefer. Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *Journal of human genetics*, 56(4):316–23, Apr 2011.

[14] Julius Gudmundsson, Patrick Sulem, Valgerdur Steinthorsdottir, Jon T Bergthorsson, Gudmar Thorleifsson, Andrei Manolescu, Thorunn Rafnar, Daniel Gudbjartsson, Bjarni A Agnarsson, Adam Baker, Asgeir Sigurdsson, Kristrun R Benediktsdottir, Margret Jakobsdottir, Thorarinn Blondal, Simon N Stacey, Agnar Helgason, Steinunn Gunnarsdottir, Adalheidur Olafsdottir, Kari T Kristinsson, Birgitta Birgisdottir, Shyamali Ghosh, Steinunn Thorlacius, Dana Magnusdottir, Gerdur Stefansdottir, Kristleifur Kristjansson, Yu Bagger, Robert L Wilensky, Muredach P Reilly, Andrew D Morris, Charlotte H Kimber, Adebowale Adeyemo, Yuanxiu Chen, Jie Zhou, Wing-Yee So, Peter C Y Tong, Maggie C Y Ng, Torben Hansen, Gitte Andersen, Knut Borch-Johnsen, Torben Jorgensen, Alejandro Tres, Fernando Fuertes, Manuel Ruiz-Echarri, Laura Asin, Berta Saez, Erica van Boven, Siem Klaver, Dorine W Swinkels, Katja K Aben, Theresa Graif, John Cashy, Brian K Suarez, Onco van Vierssen Trip, Michael L Frigge, Carole Ober, Marten H Hofker, Cisca Wijmenga, Claus Christiansen, Daniel J Rader, Colin N A Palmer, Charles Rotimi, Juliana C N Chan, Oluf Pedersen, Gunnar Sigurdsson, Rafn Benediktsson, Eirikur Jonsson, Gudmundur V Einarsson, Jose I Mayordomo, William J Catalona, Lambertus A Kiemeney, Rosa B Barkardottir, Jeffrey R Gulcher, Unnur Thorsteinsdottir, Augustine Kong, and Kari Stefansson. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nature genetics*, 39(8):977–83, Aug 2007.

[15] Kai Wang, Robert Baldassano, Haitao Zhang, Hui-Qi Qu, Marcin Imielinski, Subra Kugathasan, Vito Annese, Marla Dubinsky, Jerome I Rotter, Richard K Russell, Jonathan P Bradfield, Patrick M A Sleiman, Joseph T Glessner, Thomas Walters, Cuiping Hou, Cecilia Kim, Edward C Frackelton, Maria Garris, James Doran, Claudio Romano, Carlo Catassi, Johan Van Limbergen, Stephen L Guthery, Lee Denson, David Piccoli, Mark S Silverberg, Charles A Stanley, Dimitri Monos, David C Wilson, Anne Griffiths, Struan F A Grant, Jack Satsangi, Constantin Polychronakos, and Hakon Hakonarson. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Human molecular genetics*, 19(10):2059–67, May 2010.

[16] Till F M Andlauer, Dorothea Buck, Gisela Antony, Antonios Bayas, Lukas Bechmann, Achim Berthele, Andrew Chan, Christiane Gasperi, Ralf Gold, Christiane Graetz, Jürgen Haas, Michael Hecker, Carmen Infante-Duarte, Matthias Knop, Tania Kümpfel, Volker Limmroth, Ralf A Linker, Verena Loleit, Felix Luessi, Sven G

Meuth, Mark Mühlau, Sandra Nischwitz, Friedemann Paul, Michael Pütz, Tobias Ruck, Anke Salmen, Martin Stangel, Jan-Patrick Stellmann, Klarissa H Stürner, Björn Tackenberg, Florian Then Bergh, Hayrettin Tumani, Clemens Warnke, Frank Weber, Heinz Wiendl, Brigitte Wildemann, Uwe K Zettl, Ulf Ziemann, Frauke Zipp, Janine Arloth, Peter Weber, Milena Radivojkov-Blagojevic, Markus O Scheinhardt, Theresa Dankowski, Thomas Bettecken, Peter Lichtner, Darina Czamara, Tania Carrillo-Roa, Elisabeth B Binder, Klaus Berger, Lars Bertram, Andre Franke, Christian Gieger, Stefan Herms, Georg Homuth, Marcus Ising, Karl-Heinz Jöckel, Tim Kacprowski, Stefan Kloiber, Matthias Laudes, Wolfgang Lieb, Christina M Lill, Susanne Lucae, Thomas Meitinger, Susanne Moebus, Martina Müller-Nurasyid, Markus M Nöthen, Astrid Petersmann, Rajesh Rawal, Ulf Schminke, Konstantin Strauch, Henry Völzke, Melanie Waldenberger, Jürgen Wellmann, Eleonora Porcu, Antonella Mulas, Maristella Pitzalis, Carlo Sidore, Ilenia Zara, Francesco Cucca, Magdalena Zoledziewska, Andreas Ziegler, Bernhard Hemmer, and Bertram Müller-Myhsok. Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Science advances*, 2(6):e1501678, 06 2016.

[17] Yuan Zhou, Gu Zhu, Jac C Charlesworth, Steve Simpson, Rohina Rubicz, Harald Hh Göring, Nikolaos A Patsopoulos, Caroline Laverty, Feitong Wu, Anjali Henders, Jonathan J Ellis, Ingrid van der Mei, Grant W Montgomery, John Blangero, Joanne E Curran, Matthew P Johnson, Nicholas G Martin, Dale R Nyholt, and Bruce V Taylor. Genetic loci for Epstein-Barr virus nuclear antigen-1 are associated with risk of multiple sclerosis. *Multiple sclerosis (Houndmills, Basingstoke, England)*, 22(13):1655–1664, Nov 2016.

[18] Kai Wang, Wei-Dong Li, Clarence K Zhang, Zuoheng Wang, Joseph T Glessner, Struan F A Grant, Hongyu Zhao, Hakon Hakonarson, and R Arlen Price. A genome-wide association study on obesity and obesity-related traits. *PloS one*, 6(4):e18939, Apr 2011.

[19] R Dorajoo, A I F Blakemore, X Sim, R T-H Ong, D P K Ng, M Seielstad, T-Y Wong, S-M Saw, P Froguel, J Liu, and E-S Tai. Replication of 13 obesity loci among Singaporean Chinese, Malay and Asian-Indian populations. *International journal of obesity (2005)*, 36(1):159–63, Jan 2012.

[20] Diana L Cousminer, Diane J Berry, Nicholas J Timpson, Wei Ang, Elisabeth Thiering, Enda M Byrne, H Rob Taal, Ville Huikari, Jonathan P Bradfield, Marjan Kerkhof, Maria M Groen-Blokhuis, Eskil Kreiner-Møller, Marcella Marinelli, Claus Holst, Jaakko T Leinonen, John R B Perry, Ida Surakka, Olli Pietiläinen, Johannes Kettunen, Verneri Anttila, Marika Kaakinen, Ulla Sovio, Anneli Pouta, Shikta Das, Vasiliki Lagou, Chris Power, Inga Prokopenko, David M Evans, John P Kemp, Beate St Pourcain, Susan Ring, Aarno Palotie, Eero Kajantie, Clive Osmond, Terho Lehtimäki, Jorma S Viikari, Mika Kähönen, Nicole M Warrington, Stephen J Lye, Lyle J Palmer, Carla M T Tiesler, Claudia Flexeder, Grant W Montgomery, Sarah E Medland, Albert Hofman, Hakon Hakonarson, Mònica Guxens, Meike Bartels, Veikko Salomaa, Joanne M Murabito, Jaakko Kaprio, Thorkild I A Sørensen, Ferran Ballester, Hans Bisgaard, Dorret I Boomsma, Gerard H Koppelman, Struan F A Grant, Vincent W V Jaddoe, Nicholas G Martin, Joachim Heinrich, Craig E Pennell, Olli T Raitakari, Johan G Eriksson, George Davey Smith, Elina Hyppönen, Marjo-Riitta Järvelin, Mark I

McCarthy, Samuli Ripatti, and Elisabeth Widén. Genome-wide association and longitudinal analyses reveal genetic loci linking pubertal height growth, pubertal timing and childhood adiposity. *Human molecular genetics*, 22(13):2735–47, Jul 2013.

[21] Jiali Han, Peter Kraft, Hongmei Nan, Qun Guo, Constance Chen, Abrar Qureshi, Susan E Hankinson, Frank B Hu, David L Duffy, Zhen Zhen Zhao, Nicholas G Martin, Grant W Montgomery, Nicholas K Hayward, Gilles Thomas, Robert N Hoover, Stephen Chanock, and David J Hunter. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS genetics*, 4(5):e1000074, May 2008.

[22] Myoung Keun Lee, John R Shaffer, Elizabeth J Leslie, Ekaterina Orlova, Jenna C Carlson, Eleanor Feingold, Mary L Marazita, and Seth M Weinberg. Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. *PloS one*, 12(4):e0176566, 2017.

[23] James N Ingle, Daniel J Schaid, Paul E Goss, Mohan Liu, Taisei Mushiroda, Judy-Anne W Chapman, Michiaki Kubo, Gregory D Jenkins, Anthony Batzler, Lois Shepherd, Joseph Pater, Liewei Wang, Matthew J Ellis, Vered Stearns, Daniel C Rohrer, Matthew P Goetz, Kathleen I Pritchard, David A Flockhart, Yusuke Nakamura, and Richard M Weinshilboum. Genome-wide associations and functional genomic studies of musculoskeletal adverse events in women receiving aromatase inhibitors. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(31):4674–82, Nov 2010.

[24] Kazuma Kiyotani, Taisei Mushiroda, Tatsuhiko Tsunoda, Takashi Morizono, Naoya Hosono, Michiaki Kubo, Yusuke Tanigawara, Chiyo K Imamura, David A Flockhart, Fuminori Aki, Koichi Hirata, Yuichi Takatsuka, Minoru Okazaki, Shozo Ohsumi, Takashi Yamakawa, Mitsunori Sasa, Yusuke Nakamura, and Hitoshi Zembutsu. A genome-wide association study identifies locus at 10q22 associated with clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients in Japanese. *Human molecular genetics*, 21(7):1665–72, Apr 2012.

[25] Suyoun Chung, Siew-Kee Low, Hitoshi Zembutsu, Atsushi Takahashi, Michiaki Kubo, Mitsunori Sasa, and Yusuke Nakamura. A genome-wide association study of chemotherapy-induced alopecia in breast cancer patients. *Breast cancer research : BCR*, 15(5):R81, 2013.

[26] Susanna Atwell, Yu S Huang, Bjarni J Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M Tarone, Tina T Hu, Rong Jiang, N Wayan Muliyati, Xu Zhang, Muhammad Ali Amer, Ivan Baxter, Benjamin Brachi, Joanne Chory, Caroline Dean, Marilyne Debieu, Juliette de Meaux, Joseph R Ecker, Nathalie Faure, Joel M Kniskern, Jonathan D G Jones, Todd Michael, Adnane Nemri, Fabrice Roux, David E Salt, Chunlao Tang, Marco Todesco, M Brian Traw, Detlef Weigel, Paul Marjoram, Justin O Borevitz, Joy Bergelson, and Magnus Nordborg. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–31, Jun 2010.

[27] Feng Tian, Peter J Bradbury, Patrick J Brown, Hsiaoyi Hung, Qi Sun, Sherry Flint-Garcia, Torbert R Rocheford, Michael D McMullen, James B Holland, and Edward S

Buckler. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics*, 43(2):159–62, Feb 2011.

[28] Wei-Xuan Fu, Yang Liu, Xin Lu, Xiao-Yan Niu, Xiang-Dong Ding, Jian-Feng Liu, and Qin Zhang. A genome-wide association study identifies two novel promising candidate genes affecting Escherichia coli F4ab/F4ac susceptibility in swine. *PloS one*, 7(3):e32127, 2012.

[29] T Ohta. Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, 79(6):1940–4, Mar 1982.

[30] Montgomery Slatkin. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics*, 9(6):477–85, Jun 2008.

[31] David M Altshuler, Richard A Gibbs, Leena Peltonen, David M Altshuler, Richard A Gibbs, Leena Peltonen, Emmanouil Dermitzakis, Stephen F Schaffner, Fuli Yu, Leena Peltonen, Emmanouil Dermitzakis, Penelope E Bonnen, David M Altshuler, Richard A Gibbs, Paul I W de Bakker, Panos Deloukas, Stacey B Gabriel, Rhian Gwilliam, Sarah Hunt, Michael Inouye, Xiaoming Jia, Aarno Palotie, Melissa Parkin, Pamela Whittaker, Fuli Yu, Kyle Chang, Alicia Hawes, Lora R Lewis, Yanru Ren, David Wheeler, Richard A Gibbs, Donna Marie Muzny, Chris Barnes, Katayoon Darvishi, Matthew Hurles, Joshua M Korn, Kati Kristiansson, Charles Lee, Steven A McCarrol, James Nemesh, Emmanouil Dermitzakis, Alon Keinan, Stephen B Montgomery, Samuela Pollack, Alkes L Price, Nicole Soranzo, Penelope E Bonnen, Richard A Gibbs, Claudia Gonzaga-Jauregui, Alon Keinan, Alkes L Price, Fuli Yu, Verneri Anttila, Wendy Brodeur, Mark J Daly, Stephen Leslie, Gil McVean, Loukas Moutsianas, Huy Nguyen, Stephen F Schaffner, Qingrun Zhang, Mohammed J R Ghori, Ralph McGinnis, William McLaren, Samuela Pollack, Alkes L Price, Stephen F Schaffner, Fumihiko Takeuchi, Sharon R Grossman, Ilya Shlyakhter, Elizabeth B Hostetter, Pardis C Sabeti, Clement A Adebamowo, Morris W Foster, Deborah R Gordon, Julio Licinio, Maria Cristina Manca, Patricia A Marshall, Ichiro Matsuda, Duncan Ngare, Vivian Ota Wang, Deepa Reddy, Charles N Rotimi, Charmaine D Royal, Richard R Sharp, Changqing Zeng, Lisa D Brooks, and Jean E McEwen. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, Sep 2010.

[32] Joanna K Ledwoń, Ewa E Hennig, Natalia Maryan, Krzysztof Goryca, Dorota Nowakowska, Anna Niwińska, and Jerzy Ostrowski. Common low-penetrance risk variants associated with breast cancer in Polish women. *BMC cancer*, 13:510, Oct 2013.

[33] Stacey L Edwards, Jonathan Beesley, Juliet D French, and Alison M Dunning. Beyond GWASs: illuminating the dark road from association to function. *American journal of human genetics*, 93(5):779–97, Nov 2013.

[34] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, N.Y.)*, 296(5568):752–5, Apr 2002.

[35] Andreas Massouras, Sebastian M Waszak, Monica Albarca-Aguilera, Korneel Hens, Wiebke Holcombe, Julien F Ayroles, Emmanouil T Dermitzakis, Eric A Stone, Jeffrey D Jensen, Trudy F C Mackay, and Bart Deplancke. Genomic variation and its impact on gene expression in Drosophila melanogaster. *PLoS genetics*, 8(11):e1003055, 2012.

[36] Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, 348(6235):648–60, May 2015.

[37] David Melzer, John R B Perry, Dena Hernandez, Anna-Maria Corsi, Kara Stevens, Ian Rafferty, Fulvio Lauretani, Anna Murray, J Raphael Gibbs, Giuseppe Paolisso, Sajjad Rafiq, Javier Simon-Sanchez, Hana Lango, Sonja Scholz, Michael N Weedon, Sampath Arepalli, Neil Rice, Nicole Washecka, Alison Hurst, Angela Britton, William Henley, Joyce van de Leemput, Rongling Li, Anne B Newman, Greg Tranah, Tamara Harris, Vijay Panicker, Colin Dayan, Amanda Bennett, Mark I McCarthy, Aimo Ruokonen, Marjo-Riitta Jarvelin, Jack Guralnik, Stefania Bandinelli, Timothy M Frayling, Andrew Singleton, and Luigi Ferrucci. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS genetics*, 4(5):e1000072, May 2008.

[38] Fabian Grubert, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, Alicia R Martin, Peyton Greenside, Rohith Srivas, Doug H Phanstiel, Aleksandra Pekowska, Nastaran Heidari, Ghia Euskirchen, Wolfgang Huber, Jonathan K Pritchard, Carlos D Bustamante, Lars M Steinmetz, Anshul Kundaje, and Michael Snyder. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–65, Aug 2015.

[39] Nicholas E Banovich, Xun Lan, Graham McVicker, Bryce van de Geijn, Jacob F Degner, John D Blischak, Julien Roux, Jonathan K Pritchard, and Yoav Gilad. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS genetics*, 10(9):e1004663, Sep 2014.

[40] Thomas A Pearson and Teri A Manolio. How to interpret a genome-wide association study. *JAMA*, 299(11):1335–44, Mar 2008.

[41] M Fried and D M Crothers. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic acids research*, 9(23):6505–25, Dec 1981.

[42] Michal Levo and Eran Segal. In pursuit of design principles of regulatory sequences. *Nature reviews. Genetics*, 15(7):453–68, Jul 2014.

[43] T K Blackwell and H Weintraub. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science (New York, N.Y.)*, 250(4984):1104–10, Nov 1990.

[44] E Kowalska, F Bartnicki, K Pels, and W Strzalka. The impact of immobilized metal affinity chromatography (IMAC) resins on DNA aptamer selection. *Analytical and bioanalytical chemistry*, 406(22):5495–9, Sep 2014.

[45] Marko Djordjevic. SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular engineering*, 24(2):179–89, Jun 2007.

[46] Emmanuelle Roulet, Stéphane Busso, Anamaria A Camargo, Andrew J G Simpson, Nicolas Mermod, and Philipp Bucher. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nature biotechnology*, 20(8):831–5, Aug 2002.

[47] Artem Zykovich, Ian Korf, and David J Segal. Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. *Nucleic acids research*, 37(22):e151, Dec 2009.

[48] Tatjana Schütze, Barbara Wilhelm, Nicole Greiner, Hannsjörg Braun, Franziska Peter, Mario Mörl, Volker A Erdmann, Hans Lehrach, Zoltán Konthur, Marcus Menger, Peter F Arndt, and Jörn Glökler. Probing the SELEX process with next-generation sequencing. *PloS one*, 6(12):e29604, 2011.

[49] M L Bulyk, E Gentalen, D J Lockhart, and G M Church. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nature biotechnology*, 17(6):573–7, Jun 1999.

[50] Michael F Berger and Martha L Bulyk. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods in molecular biology (Clifton, N.J.)*, 338:245–60, 2006.

[51] Sonali Mukherjee, Michael F Berger, Ghil Jona, Xun S Wang, Dale Muzzey, Michael Snyder, Richard A Young, and Martha L Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature genetics*, 36(12):1331–9, Dec 2004.

[52] Eugene Bolotin, Hailing Liao, Tuong Chi Ta, Chuhu Yang, Wendy Hwang-Verslues, Jane R Evans, Tao Jiang, and Frances M Sladek. Integrated approach for the identification of human hepatocyte nuclear factor 4alpha target genes using protein binding microarrays. *Hepatology (Baltimore, Md.)*, 51(2):642–53, Feb 2010.

[53] Colin R Lickwar, Florian Mueller, and Jason D Lieb. Genome-wide measurement of protein-DNA binding dynamics using competition ChIP. *Nature protocols*, 8(7):1337–53, 2013.

[54] Peter J Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80, Oct 2009.

[55] C W Garvie and C Wolberger. Recognition of specific DNA sequences. *Molecular cell*, 8(5):937–46, Nov 2001.

[56] N M Luscombe, R A Laskowski, and J M Thornton. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic acids research*, 29(13):2860–74, Jul 2001.

[57] Andrei Lihu and Ştefan Holban. A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Briefings in bioinformatics*, 16(6):964–73, Nov 2015.

[58] G D Stormo, T D Schneider, L Gold, and A Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic acids research*, 10(9):2997–3011, May 1982.

[59] G D Stormo. DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, 16(1):16–23, Jan 2000.

[60] Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics*, 5(4):276–87, Apr 2004.

[61] T K Man and G D Stormo. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic acids research*, 29(12):2471–8, Jun 2001.

[62] Sebastian J Maerkl and Stephen R Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science (New York, N.Y.)*, 315(5809):233–7, Jan 2007.

[63] Martha L Bulyk, Philip L F Johnson, and George M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, 30(5):1255–61, Mar 2002.

[64] Jessica L Stringham, Adam S Brown, Robert A Drewell, and Jacqueline M Dresch. Flanking sequence context-dependent transcription factor binding in early Drosophila development. *BMC bioinformatics*, 14:298, Oct 2013.

[65] Alexander E Kel, Monika Niehof, Volker Matys, Rüdiger Zemlin, and Jürgen Borlak. Genome wide prediction of HNF4alpha functional binding sites by the use of local and global sequence context. *Genome biology*, 9(2):R36, 2008.

[66] Rahul Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PloS one*, 5(3):e9722, Mar 2010.

[67] Christopher Fletez-Brant, Dongwon Lee, Andrew S McCallion, and Michael A Beer. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic acids research*, 41(Web Server issue):W544–56, Jul 2013.

[68] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–8, Aug 2015.

[69] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, 12(10):931–4, Oct 2015.

[70] Bart Hooghe, Stefan Broos, Frans van Roy, and Pieter De Bleser. A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic acids research*, 40(14):e106, Aug 2012.

[71] Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W Wasserman. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell systems*, 3(3):278–286.e4, Sep 2016.

[72] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K Das, Teemu Kivioja, Kashyap Dave, Fan Zhong, Kazuhiro R Nitta, Minna Taipale, Alexander Popov, Paul A Ginno, Silvia Domcke, Jian Yan, Dirk Schübeler, Charles Vinson, and Jussi Taipale. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (New York, N.Y.)*, 356(6337), 05 2017.

[73] Sheng Liu, Cristina Zibetti, Jun Wan, Guohua Wang, Seth Blackshaw, and Jiang Qian. Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility. *BMC bioinformatics*, 18(1):355, Jul 2017.

[74] Matthew Slattery, Todd Riley, Peng Liu, Namiko Abe, Pilar Gomez-Alcala, Iris Dror, Tianyin Zhou, Remo Rohs, Barry Honig, Harmen J Bussemaker, and Richard S Mann. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, 147(6):1270–82, Dec 2011.

[75] Anthony Mathelier, Calvin Lefebvre, Allen W Zhang, David J Arenillas, Jiarui Ding, Wyeth W Wasserman, and Sohrab P Shah. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome biology*, 16:84, Apr 2015.

[76] Collin Melton, Jason A Reuter, Damek V Spacek, and Michael Snyder. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature genetics*, 47(7):710–6, Jul 2015.

[77] Frederick Kinyua Kamanu, Yulia A Medvedeva, Ulf Schaefer, Boris R Jankovic, John A C Archer, and Vladimir B Bajic. Mutations and binding sites of human transcription factors. *Frontiers in genetics*, 3:100, 2012.

[78] M J Reijnen, F M Sladek, R M Bertina, and P H Reitsma. Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. *Proceedings of the National Academy of Sciences of the United States of America*, 89(14):6300–3, Jul 1992.

[79] Shizhi Wang, Shenshen Wu, Qingtao Meng, Xiaobo Li, Jinchun Zhang, Rui Chen, and Meilin Wang. FAS rs2234767 and rs1800682 polymorphisms jointly contributed to risk of colorectal cancer by affecting SP1/STAT1 complex recruitment to chromatin. *Scientific reports*, 6:19229, Jan 2016.

[80] M Matsuda, N Sakamoto, and Y Fukumaki. Delta-thalassemia caused by disruption of the site for an erythroid-specific transcription factor, GATA-1, in the delta-globin gene promoter. *Blood*, 80(5):1347–51, Sep 1992.

[81] Juliet D French, Maya Ghoussaini, Stacey L Edwards, Kerstin B Meyer, Kyriaki Michailidou, Shahana Ahmed, Sofia Khan, Mel J Maranian, Martin O'Reilly, Kristine M Hillman, Joshua A Betts, Thomas Carroll, Peter J Bailey, Ed Dicks, Jonathan Beesley, Jonathan Tyrer, Ana-Teresa Maia, Andrew Beck, Nicholas W Knoblauch, Constance Chen, Peter Kraft, Daniel Barnes, Anna González-Neira, M Rosario Alonso, Daniel Herrero, Daniel C Tessier, Daniel Vincent, Francois Bacot, Craig Luccarini, Caroline Baynes, Don Conroy, Joe Dennis, Manjeet K Bolla, Qin Wang, John L Hopper, Melissa C Southey, Marjanka K Schmidt, Annegien Broeks, Senno Verhoef, Sten Cornelissen, Kenneth Muir, Artitaya Lophatananon, Sarah Stewart-Brown,

Pornthep Siriwanarangsan, Peter A Fasching, Christian R Loehberg, Arif B Ekici, Matthias W Beckmann, Julian Peto, Isabel dos Santos Silva, Nichola Johnson, Zoe Aitken, Elinor J Sawyer, Ian Tomlinson, Michael J Kerin, Nicola Miller, Frederik Marme, Andreas Schneeweiss, Christof Sohn, Barbara Burwinkel, Pascal Guénel, Thérèse Truong, Pierre Laurent-Puig, Florence Menegaux, Stig E Bojesen, Børge G Nordestgaard, Sune F Nielsen, Henrik Flyger, Roger L Milne, M Pilar Zamora, Jose Ignacio Arias Perez, Javier Benitez, Hoda Anton-Culver, Hermann Brenner, Heiko Müller, Volker Arndt, Christa Stegmaier, Alfons Meindl, Peter Lichtner, Rita K Schmutzler, Christoph Engel, Hiltrud Brauch, Ute Hamann, Christina Justenhoven, Kirsimari Aaltonen, Päivi Heikkilä, Kristiina Aittomäki, Carl Blomqvist, Keitaro Matsuo, Hidemi Ito, Hiroji Iwata, Aiko Sueta, Natalia V Bogdanova, Natalia N Antonenkova, Thilo Dörk, Annika Lindblom, Sara Margolin, Arto Mannermaa, Vesa Kataja, Veli-Matti Kosma, Jaana M Hartikainen, Anna H Wu, Chiu-chen Tseng, David Van Den Berg, Daniel O Stram, Diether Lambrechts, Stephanie Peeters, Ann Smeets, Giuseppe Floris, Jenny Chang-Claude, Anja Rudolph, Stefan Nickels, Dieter Flesch-Janys, Paolo Radice, Paolo Peterlongo, Bernardo Bonanni, Domenico Sardella, Fergus J Couch, Xianshu Wang, Vernon S Pankratz, Adam Lee, Graham G Giles, Gianluca Severi, Laura Baglietto, Christopher A Haiman, Brian E Henderson, Fredrick Schumacher, Loic Le Marchand, Jacques Simard, Mark S Goldberg, France Labrèche, Martine Dumont, Soo Hwang Teo, Cheng Har Yip, Char-Hong Ng, Eranga Nishanthie Vithana, Vessela Kristensen, Wei Zheng, Sandra Deming-Halverson, Martha Shrubsole, Jirong Long, Robert Winqvist, Katri Pylkäs, Arja Jukkola-Vuorinen, Mervi Grip, Irene L Andrulis, Julia A Knight, Gord Glendon, Anna Marie Mulligan, Peter Devilee, Caroline Seynaeve, Montserrat García-Closas, Jonine Figueroa, Stephen J Chanock, Jolanta Lissowska, Kamila Czene, Daniel Klevebring, Nils Schoof, Maartje J Hooning, John W M Martens, J Margriet Collée, Madeleine Tilanus-Linthorst, Per Hall, Jingmei Li, Jianjun Liu, Keith Humphreys, Xiao-Ou Shu, Wei Lu, Yu-Tang Gao, Hui Cai, Angela Cox, Sabapathy P Balasubramanian, William Blot, Lisa B Signorello, Qiuyin Cai, Paul D P Pharoah, Catherine S Healey, Mitul Shah, Karen A Pooley, Daehee Kang, Keun-Young Yoo, Dong-Young Noh, Mikael Hartman, Hui Miao, Jen-Hwei Sng, Xueling Sim, Anna Jakubowska, Jan Lubinski, Katarzyna Jaworska-Bieniek, Katarzyna Durda, Suleeporn Sangrajrang, Valerie Gaborieau, James McKay, Amanda E Toland, Christine B Ambrosone, Drakoulis Yannoukakos, Andrew K Godwin, Chen-Yang Shen, Chia-Ni Hsiung, Pei-Ei Wu, Shou-Tung Chen, Anthony Swerdlow, Alan Ashworth, Nick Orr, Minouk J Schoemaker, Bruce A J Ponder, Heli Nevanlinna, Melissa A Brown, Georgia Chenevix-Trench, Douglas F Easton, and Alison M Dunning. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *American journal of human genetics*, 92(4):489–503, Apr 2013.

[82] Nils Weinhold, Anders Jacobsen, Nikolaus Schultz, Chris Sander, and William Lee. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, 46(11):1160–5, Nov 2014.

[83] C Solis, G I Aizencang, K H Astrin, D F Bishop, and R J Desnick. Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria. *The Journal of clinical investigation*, 107(6):753–62, Mar 2001.

[84] C Gragnoli, T Lindner, B N Cockburn, P J Kaisaki, F Gragnoli, G Marozzi, and G I Bell. Maturity-onset diabetes of the young due to a mutation in the hepatocyte nuclear factor-4 alpha binding site in the promoter of the hepatocyte nuclear factor-1 alpha gene. *Diabetes*, 46(10):1648–51, Oct 1997.

[85] L B Ludlow, B P Schick, M L Budarf, D A Driscoll, E H Zackai, A Cohen, and B A Konkle. Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibbeta promoter resulting in the Bernard-Soulier syndrome. *The Journal of biological chemistry*, 271(36):22076–80, Sep 1996.

[86] Marco De Gobbi, Vip Viprakasit, Jim R Hughes, Chris Fisher, Veronica J Buckle, Helena Ayyub, Richard J Gibbons, Douglas Vernimmen, Yuko Yoshinaga, Pieter de Jong, Jan-Fang Cheng, Edward M Rubin, William G Wood, Don Bowden, and Douglas R Higgs. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science (New York, N.Y.)*, 312(5777):1215–7, May 2006.

[87] W S Yang, D N Nevin, R Peng, J D Brunzell, and S S Deeb. A mutation in the promoter of the lipoprotein lipase (LPL) gene in a patient with familial combined hyperlipidemia and low LPL activity. *Proceedings of the National Academy of Sciences of the United States of America*, 92(10):4462–6, May 1995.

[88] Jianming Wu, Christine Metz, Xiulong Xu, Riichiro Abe, Andrew W Gibson, Jeffrey C Edberg, Jennifer Cooke, Fenglong Xie, Glinda S Cooper, and Robert P Kimberly. A novel polymorphic CAAT/enhancer-binding protein beta element in the FasL gene promoter alters Fas ligand expression: a candidate background gene in African American systemic lupus erythematosus patients. *Journal of immunology (Baltimore, Md. : 1950)*, 170(1):132–8, Jan 2003.

[89] Andrew W Dodd, Catherine M Syddall, and John Loughlin. A rare variant in the osteoarthritis-associated locus GDF5 is functional and reveals a site that can be manipulated to modulate GDF5 expression. *European journal of human genetics : EJHG*, 21(5):517–21, May 2013.

[90] M Crossley, M Ludwig, K M Stowell, P De Vos, K Olek, and G G Brownlee. Recovery from hemophilia B Leyden: an androgen-responsive element in the factor IX promoter. *Science (New York, N.Y.)*, 257(5068):377–9, Jul 1992.

[91] Xing-Wu Zheng, Rama Kudaravalli, Theresa T Russell, Donna M DiMichele, Constance Gibb, J Eric Russell, Paris Margaritis, and Eleanor S Pollak. Mutation in the factor VII hepatocyte nuclear factor 4$\alpha$-binding site contributes to factor VII deficiency. *Blood coagulation & fibrinolysis : an international journal in haemostasis and thrombosis*, 22(7):624–7, Oct 2011.

[92] Melina Claussnitzer, Simon N Dankel, Bernward Klocke, Harald Grallert, Viktoria Glunk, Tea Berulava, Heekyoung Lee, Nikolay Oskolkov, Joao Fadista, Kerstin Ehlers, Simone Wahl, Christoph Hoffmann, Kun Qian, Tina Rönn, Helene Riess, Martina Müller-Nurasyid, Nancy Bretschneider, Timm Schroeder, Thomas Skurk, Bernhard Horsthemke, Derek Spieler, Martin Klingenspor, Martin Seifert, Michael J Kern, Niklas Mejhert, Ingrid Dahlman, Ola Hansson, Stefanie M Hauck, Matthias Blüher, Peter Arner, Leif Groop, Thomas Illig, Karsten Suhre, Yi-Hsiang Hsu, Gunnar

Mellgren, Hans Hauner, and Helmut Laumen. Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell*, 156(1-2):343–58, Jan 2014.

[93] Li Jia, Gilad Landan, Mark Pomerantz, Rami Jaschek, Paula Herman, David Reich, Chunli Yan, Omar Khalid, Phil Kantoff, William Oh, J Robert Manak, Benjamin P Berman, Brian E Henderson, Baruch Frenkel, Christopher A Haiman, Matthew Freedman, Amos Tanay, and Gerhard A Coetzee. Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS genetics*, 5(8):e1000597, Aug 2009.

[94] D I Martin, S F Tsai, and S H Orkin. Increased gamma-globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding factor. *Nature*, 338(6214):435–8, Mar 1989.

[95] Michael N Weedon, Ines Cebola, Ann-Marie Patch, Sarah E Flanagan, Elisa De Franco, Richard Caswell, Santiago A Rodríguez-Seguí, Charles Shaw-Smith, Candy H-H Cho, Hana Lango Allen, Jayne Al Houghton, Christian L Roth, Rongrong Chen, Khalid Hussain, Phil Marsh, Ludovic Vallier, Anna Murray, Sian Ellard, Jorge Ferrer, and Andrew T Hattersley. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nature genetics*, 46(1):61–64, Jan 2014.

[96] Jennifer R Kulzer, Michael L Stitzel, Mario A Morken, Jeroen R Huyghe, Christian Fuchsberger, Johanna Kuusisto, Markku Laakso, Michael Boehnke, Francis S Collins, and Karen L Mohlke. A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *American journal of human genetics*, 94(2):186–97, Feb 2014.

[97] Dominique J Verlaan, Soizik Berlivet, Gary M Hunninghake, Anne-Marie Madore, Mathieu Larivière, Sanny Moussette, Elin Grundberg, Tony Kwan, Manon Ouimet, Bing Ge, Rose Hoberman, Marcin Swiatek, Joana Dias, Kevin C L Lam, Vonda Koka, Eef Harmsen, Manuel Soto-Quiros, Lydiana Avila, Juan C Celedón, Scott T Weiss, Ken Dewar, Daniel Sinnett, Catherine Laprise, Benjamin A Raby, Tomi Pastinen, and Anna K Naumova. Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. *American journal of human genetics*, 85(3):377–93, Sep 2009.

[98] Sabina Benko, Judy A Fantes, Jeanne Amiel, Dirk-Jan Kleinjan, Sophie Thomas, Jacqueline Ramsay, Negar Jamshidi, Abdelkader Essafi, Simon Heaney, Christopher T Gordon, David McBride, Christelle Golzio, Malcolm Fisher, Paul Perry, Véronique Abadie, Carmen Ayuso, Muriel Holder-Espinasse, Nicky Kilpatrick, Melissa M Lees, Arnaud Picard, I Karen Temple, Paul Thomas, Marie-Paule Vazquez, Michel Vekemans, Hugues Roest Crollius, Nicholas D Hastie, Arnold Munnich, Heather C Etchevers, Anna Pelet, Peter G Farlie, David R Fitzpatrick, and Stanislas Lyonnet. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nature genetics*, 41(3):359–64, Mar 2009.

[99] Laure Lecerf, Anthula Kavo, Macarena Ruiz-Ferrer, Viviane Baral, Yuli Watanabe, Asma Chaoui, Veronique Pingault, Salud Borrego, and Nadege Bondurand. An impairment of long distance SOX10 regulatory elements underlies isolated Hirschsprung disease. *Human mutation*, 35(3):303–7, Mar 2014.

[100] Shoaib Al Zadjali, Yasser Wali, Fatma Al Lawatiya, David Gravell, Salam Alkindi, Kareema Al Falahi, Rajagopal Krishnamoorthy, and Shahina Daar. The $\beta$-globin promoter -71 C>T mutation is a $\beta$+ thalassemic allele. *European journal of haematology*, 87(5):457–60, Nov 2011.

[101] Cibele Masotti, Lucia M Armelin-Correa, Alessandra Splendore, Chin J Lin, Angela Barbosa, Mari C Sogayar, and Maria Rita Passos-Bueno. A functional SNP in the promoter region of TCOF1 is associated with reduced gene expression and YY1 DNA-protein interaction. *Gene*, 359:44–52, Oct 2005.

[102] Fedik Rahimov, Mary L Marazita, Axel Visel, Margaret E Cooper, Michael J Hitchler, Michele Rubini, Frederick E Domann, Manika Govil, Kaare Christensen, Camille Bille, Mads Melbye, Astanand Jugessur, Rolv T Lie, Allen J Wilcox, David R Fitzpatrick, Eric D Green, Peter A Mossey, Julian Little, Regine P Steegers-Theunissen, Len A Pennacchio, Brian C Schutte, and Jeffrey C Murray. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nature genetics*, 40(11):1341–7, Nov 2008.

[103] Dongwon Lee, David U Gorkin, Maggie Baker, Benjamin J Strober, Alessandro L Asoni, Andrew S McCallion, and Michael A Beer. A method to predict the impact of regulatory variants from DNA sequence. *Nature genetics*, 47(8):955–61, Aug 2015.

[104] Jieming Chen, Joel Rozowsky, Timur R Galeev, Arif Harmanci, Robert Kitchen, Jason Bedford, Alexej Abyzov, Yong Kong, Lynne Regan, and Mark Gerstein. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nature communications*, 7:11101, Apr 2016.

[105] Wenqiang Shi, Oriol Fornes, Anthony Mathelier, and Wyeth W Wasserman. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic acids research*, 44(21):10106–10116, Dec 2016.

[106] Ines de Santiago, Wei Liu, Ke Yuan, Martin O'Reilly, Chandra Sekhar Reddy Chilamakuri, Bruce A J Ponder, Kerstin B Meyer, and Florian Markowetz. BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. *Genome biology*, 18(1):39, Feb 2017.

[107] Heng Tao, David R Cox, and Kelly A Frazer. Allele-specific KRT1 expression is a complex trait. *PLoS genetics*, 2(6):e93, Jun 2006.

[108] Ruslan Strogantsev, Felix Krueger, Kazuki Yamazawa, Hui Shi, Poppy Gould, Megan Goldman-Roberts, Kirsten McEwen, Bowen Sun, Roger Pedersen, and Anne C Ferguson-Smith. Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome biology*, 16:112, May 2015.

[109] Roel Nusse. Wnt signaling and stem cell control. *Cell research*, 18(5):523–7, May 2008.

[110] S Wickner, M R Maurizi, and S Gottesman. Posttranslational quality control: folding, refolding, and degrading proteins. *Science (New York, N.Y.)*, 286(5446):1888–93, Dec 1999.

[111] Shimin Zhao, Wei Xu, Wenqing Jiang, Wei Yu, Yan Lin, Tengfei Zhang, Jun Yao, Li Zhou, Yaxue Zeng, Hong Li, Yixue Li, Jiong Shi, Wenlin An, Susan M Hancock, Fuchu He, Lunxiu Qin, Jason Chin, Pengyuan Yang, Xian Chen, Qunying Lei, Yue Xiong, and Kun-Liang Guan. Regulation of cellular metabolism by protein lysine acetylation. *Science (New York, N.Y.)*, 327(5968):1000–4, Feb 2010.

[112] Fuxiao Xin and Predrag Radivojac. Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics (Oxford, England)*, 28(22):2905–13, Nov 2012.

[113] Guangyou Duan and Dirk Walther. The roles of post-translational modifications in the context of protein interaction networks. *PLoS computational biology*, 11(2):e1004049, Feb 2015.

[114] Pablo Minguez, Luca Parca, Francesca Diella, Daniel R Mende, Runjun Kumar, Manuela Helmer-Citterich, Anne-Claude Gavin, Vera van Noort, and Peer Bork. Deciphering a global network of functionally associated post-translational modifications. *Molecular systems biology*, 8:599, Jul 2012.

[115] UniProt: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169, Jan 2017.

[116] Tzong-Yi Lee, Hsien-Da Huang, Jui-Hung Hung, Hsi-Yuan Huang, Yuh-Shyong Yang, and Tzu-Hao Wang. dbPTM: an information repository of protein post-translational modification. *Nucleic acids research*, 34(Database issue):D622–7, Jan 2006.

[117] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research*, 43(Database issue):D512–20, Jan 2015.

[118] Kai-Yao Huang, Min-Gang Su, Hui-Ju Kao, Yun-Chung Hsieh, Jhih-Hua Jhong, Kuang-Hao Cheng, Hsien-Da Huang, and Tzong-Yi Lee. dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic acids research*, 44(D1):D435–46, Jan 2016.

[119] Monzy Thomas, Nahid Dadgar, Abhishek Aphale, Jennifer M Harrell, Robin Kunkel, William B Pratt, and Andrew P Lieberman. Androgen receptor acetylation site mutations cause trafficking defects, misfolding, and aggregation similar to expanded glutamine tracts. *The Journal of biological chemistry*, 279(9):8389–95, Feb 2004.

[120] E Grasbon-Frodl, Holger Lorenz, U Mann, R M Nitsch, Otto Windl, and H A Kretzschmar. Loss of glycosylation associated with the T183A mutation in human prion disease. *Acta neuropathologica*, 108(6):476–84, Dec 2004.

[121] Jüri Reimand, Omar Wagih, and Gary D Bader. Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLoS genetics*, 11(1):e1004919, Jan 2015.

[122] G Manning, D B Whyte, R Martinez, T Hunter, and S Sudarsanam. The protein kinase complement of the human genome. *Science (New York, N.Y.)*, 298(5600):1912–34, 12 2002.

[123] Jeffrey A Ubersax and James E Ferrell. Mechanisms of specificity in protein phosphorylation. *Nature reviews. Molecular cell biology*, 8(7):530–41, Jul 2007.

[124] B E Kemp and R B Pearson. Protein kinase recognition sequence motifs. *Trends in biochemical sciences*, 15(9):342–6, Sep 1990.

[125] Martin Lee Miller, Lars Juhl Jensen, Francesca Diella, Claus Jørgensen, Michele Tinti, Lei Li, Marilyn Hsiung, Sirlester A Parker, Jennifer Bordeaux, Thomas Sicheritz-Ponten, Marina Olhovsky, Adrian Pasculescu, Jes Alexander, Stefan Knapp, Nikolaj Blom, Peer Bork, Shawn Li, Gianni Cesareni, Tony Pawson, Benjamin E Turk, Michael B Yaffe, Søren Brunak, and Rune Linding. Linear motif atlas for phosphorylation-dependent signaling. *Science signaling*, 1(35):ra2, Sep 2008.

[126] Tzong-Yi Lee, Justin Bo-Kai Hsu, Wen-Chi Chang, and Hsien-Da Huang. RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans. *Nucleic acids research*, 39(Database issue):D777–87, Jan 2011.

[127] Hamid D Ismail, Ahoi Jones, Jung H Kim, Robert H Newman, and Dukka B Kc. RF-Phos: A Novel General Phosphorylation Site Prediction Tool Based on Random Forest. *BioMed research international*, 2016:3281590, 2016.

[128] M Ristow, D Müller-Wieland, A Pfeiffer, W Krone, and C R Kahn. Obesity associated with a mutation in a genetic regulator of adipocyte differentiation. *The New England journal of medicine*, 339(14):953–9, Oct 1998.

[129] S Dupuis, C Dargemont, C Fieschi, N Thomassin, S Rosenzweig, J Harris, S M Holland, R D Schreiber, and J L Casanova. Impairment of mycobacterial but not viral immunity by a germline human STAT1 mutation. *Science (New York, N.Y.)*, 293(5528):300–3, Jul 2001.

[130] K L Toh, C R Jones, Y He, E J Eide, W A Hinz, D M Virshup, L J Ptácek, and Y H Fu. An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science (New York, N.Y.)*, 291(5506):1040–3, Feb 2001.

[131] Søren M Echwald, Helle Bach, Henrik Vestergaard, Bjørn Richelsen, Kurt Kristensen, Thomas Drivsholm, Knut Borch-Johnsen, Torben Hansen, and Oluf Pedersen. A P387L variant in protein tyrosine phosphatase-1B (PTP-1B) is associated with type 2 diabetes and impaired serine phosphorylation of PTP-1B in vitro. *Diabetes*, 51(1):1–6, Jan 2002.

[132] Edward P Gelmann, David J Steadman, Jing Ma, Natalie Ahronovitz, H James Voeller, Sheridan Swope, Mohammed Abbaszadegan, Kevin M Brown, Kate Strand, Richard B Hayes, and Meir J Stampfer. Occurrence of NKX3.1 C154T polymorphism in men with and without prostate cancer and studies of its effect on protein function. *Cancer research*, 62(9):2654–9, May 2002.

[133] Xiaoxian Li, Patrick Dumont, Anthony Della Pietra, Cory Shetler, and Maureen E Murphy. The codon 47 polymorphism in p53 is functionally significant. *The Journal of biological chemistry*, 280(25):24245–51, Jun 2005.

[134] Luisa Luna, Veslemøy Rolseth, Gunn A Hildrestrand, Marit Otterlei, Françoise Dantzer, Magnar Bjørås, and Erling Seeberg. Dynamic relocalization of hOGG1 during the cell cycle is disrupted in cells harbouring the hOGG1-Cys326 polymorphic variant. *Nucleic acids research*, 33(6):1813–24, 2005.

[135] S Benzeno, F Lu, M Guo, O Barbash, F Zhang, J G Herman, P S Klein, A Rustgi, and J A Diehl. Identification of mutations that disrupt phosphorylation-dependent nuclear export of cyclin D1. *Oncogene*, 25(47):6291–303, Oct 2006.

[136] You-Take Oh, Kwang Hoon Chun, Byoung Duck Park, Joon-Seok Choi, and Seung Ki Lee. Regulation of cyclin-dependent kinase inhibitor p21WAF1/CIP1 by protein kinase Cdelta-mediated phosphorylation. *Apoptosis : an international journal on programmed cell death*, 12(7):1339–47, Jul 2007.

[137] Saverio Gentile, Negin Martin, Erica Scappini, Jason Williams, Christian Erxleben, and David L Armstrong. The human ERG1 channel polymorphism, K897T, creates a phosphorylation site that inhibits channel activity. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38):14704–8, Sep 2008.

[138] Delaine K Ceholski, Catharine A Trieber, Charles F B Holmes, and Howard S Young. Lethal, hereditary mutants of phospholamban elude phosphorylation by protein kinase A. *The Journal of biological chemistry*, 287(32):26596–605, Aug 2012.

[139] William H Lagarde, Amanda J Blackwelder, John T Minges, Andrew T Hnat, Frank S French, and Elizabeth M Wilson. Androgen receptor exon 1 mutation causes androgen insensitivity by creating phosphorylation site and inhibiting melanoma antigen-A11 activation of NH2- and carboxyl-terminal interaction-dependent transactivation. *The Journal of biological chemistry*, 287(14):10905–15, Mar 2012.

[140] Fei-Yan Deng, Li-Jun Tan, Hui Shen, Yong-Jun Liu, Yao-Zhong Liu, Jian Li, Xue-Zhen Zhu, Xiang-Ding Chen, Qing Tian, Ming Zhao, and Hong-Wen Deng. SNP rs6265 regulates protein phosphorylation and osteoblast differentiation and influences BMD in humans. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, 28(12):2498–507, Dec 2013.

[141] Daniel H Ebert, Harrison W Gabel, Nathaniel D Robinson, Nathaniel R Kastan, Linda S Hu, Sonia Cohen, Adrija J Navarro, Matthew J Lyst, Robert Ekiert, Adrian P Bird, and Michael E Greenberg. Activity-dependent phosphorylation of MeCP2 threonine 308 regulates interaction with NCoR. *Nature*, 499(7458):341–5, Jul 2013.

[142] C Ortiz-Padilla, D Gallego-Ortega, B C Browne, F Hochgräfe, C E Caldon, R J Lyons, D R Croucher, D Rickwood, C J Ormandy, T Brummer, and R J Daly. Functional characterization of cancer-associated Gab1 mutations. *Oncogene*, 32(21):2696–702, May 2013.

[143] Julien Gautherot, Danièle Delautier, Marie-Anne Maubert, Tounsia Aït-Slimane, Gérard Bolbach, Jean-Louis Delaunay, Anne-Marie Durand-Schneider, Delphine Firrincieli, Véronique Barbu, Nicolas Chignard, Chantal Housset, Michèle Maurice, and Thomas Falguières. Phosphorylation of ABCB4 impacts its function: insights from disease-causing mutations. *Hepatology (Baltimore, Md.)*, 60(2):610–21, Aug 2014.

[144] Eunice E Lee, Jing Ma, Anastasia Sacharidou, Wentao Mi, Valerie K Salato, Nam Nguyen, Youxing Jiang, Juan M Pascual, Paula E North, Philip W Shaul, Marcel Mettlen, and Richard C Wang. A Protein Kinase C Phosphorylation Motif in GLUT1 Affects Glucose Transport and is Mutated in GLUT1 Deficiency Syndrome. *Molecular cell*, 58(5):845–53, Jun 2015.

[145] Marcello Niceta, Emilia Stellacci, Karen W Gripp, Giuseppe Zampino, Maria Kousi, Massimiliano Anselmi, Alice Traversa, Andrea Ciolfi, Deborah Stabley, Alessandro Bruselles, Viviana Caputo, Serena Cecchetti, Sabrina Prudente, Maria T Fiorenza, Carla Boitani, Nicole Philip, Dmitriy Niyazov, Chiara Leoni, Takaya Nakane, Kim Keppler-Noreuil, Stephen R Braddock, Gabriele Gillessen-Kaesbach, Antonio Palleschi, Philippe M Campeau, Brendan H L Lee, Celio Pouponnot, Lorenzo Stella, Gianfranco Bocchinfuso, Nicholas Katsanis, Katia Sol-Church, and Marco Tartaglia. Mutations Impairing GSK3-Mediated MAF Phosphorylation Cause Cataract, Deafness, Intellectual Disability, Seizures, and a Down Syndrome-like Facies. *American journal of human genetics*, 96(5):816–25, May 2015.

[146] Omar Wagih, Jüri Reimand, and Gary D Bader. MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nature methods*, 12(6):531–3, Jun 2015.

[147] Jason J Yi, Janet Berrios, Jason M Newbern, William D Snider, Benjamin D Philpot, Klaus M Hahn, and Mark J Zylka. An Autism-Linked Mutation Disables Phosphorylation Control of UBE3A. *Cell*, 162(4):795–807, Aug 2015.

[148] Jüri Reimand, Omar Wagih, and Gary D Bader. The mutational landscape of phosphorylation signaling in cancer. *Scientific reports*, 3:2651, Oct 2013.

[149] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, 32(Database issue):D115–9, Jan 2004.

[150] Yul Kim, Chiyong Kang, Bumki Min, and Gwan-Su Yi. Detection and analysis of disease-associated single nucleotide polymorphism influencing post-translational modification. *BMC medical genomics*, 8 Suppl 2:S7, 2015.

[151] Gil-Mi Ryu, Pamela Song, Kyu-Won Kim, Kyung-Soo Oh, Keun-Joon Park, and Jong Hun Kim. Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. *Nucleic acids research*, 37(4):1297–307, Mar 2009.

[152] Yu Xue, Jian Ren, Xinjiao Gao, Changjiang Jin, Longping Wen, and Xuebiao Yao. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Molecular & cellular proteomics : MCP*, 7(9):1598–608, Sep 2008.

[153] Jian Ren, Chunhui Jiang, Xinjiao Gao, Zexian Liu, Zineng Yuan, Changjiang Jin, Longping Wen, Zhaolei Zhang, Yu Xue, and Xuebiao Yao. PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Molecular & cellular proteomics : MCP*, 9(4):623–34, Apr 2010.

[154] Rune Linding, Lars Juhl Jensen, Gerard J Ostheimer, Marcel A T M van Vugt, Claus Jørgensen, Ioana M Miron, Francesca Diella, Karen Colwill, Lorne Taylor, Kelly Elder, Pavel Metalnikov, Vivian Nguyen, Adrian Pasculescu, Jing Jin, Jin Gyoon Park, Leona D Samson, James R Woodgett, Robert B Russell, Peer Bork, Michael B Yaffe, and Tony Pawson. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–26, Jun 2007.

[155] Pau Creixell, Erwin M Schoof, Craig D Simpson, James Longden, Chad J Miller, Hua Jane Lou, Lara Perryman, Thomas R Cox, Nevena Zivanovic, Antonio Palmeri, Agata Wesolowska-Andersen, Manuela Helmer-Citterich, Jesper Ferkinghoff-Borg, Hiroaki Itamochi, Bernd Bodenmiller, Janine T Erler, Benjamin E Turk, and Rune Linding. Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell*, 163(1):202–17, Sep 2015.

[156] Ralph Patrick, Bostjan Kobe, Kim-Anh Lê Cao, and Mikael Bodén. PhosphoPICK-SNP: quantifying the effect of amino acid variants on protein phosphorylation. *Bioinformatics (Oxford, England)*, 33(12):1773–1781, Jun 2017.

[157] Norman E Davey, Kim Van Roey, Robert J Weatheritt, Grischa Toedt, Bora Uyar, Brigitte Altenberg, Aidan Budd, Francesca Diella, Holger Dinkel, and Toby J Gibson. Attributes of short linear motifs. *Molecular bioSystems*, 8(1):268–81, Jan 2012.

[158] Debasree Sarkar, Tanmoy Jana, and Sudipto Saha. LMPID: a manually curated database of linear motifs mediating protein-protein interactions. *Database : the journal of biological databases and curation*, 2015, 2015.

[159] Tian Mi, Jerlin Camilus Merlin, Sandeep Deverasetty, Michael R Gryk, Travis J Bill, Andrew W Brooks, Logan Y Lee, Viraj Rathnayake, Christian A Ross, David P Sargeant, Christy L Strong, Paula Watts, Sanguthevar Rajasekaran, and Martin R Schiller. Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic acids research*, 40(Database issue):D252–60, Jan 2012.

[160] John C Obenauer, Lewis C Cantley, and Michael B Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic acids research*, 31(13):3635–41, Jul 2003.

[161] Holger Dinkel, Sushama Michael, Robert J Weatheritt, Norman E Davey, Kim Van Roey, Brigitte Altenberg, Grischa Toedt, Bora Uyar, Markus Seiler, Aidan Budd, Lisa Jödicke, Marcel A Dammert, Christian Schroeter, Maria Hammer, Tobias Schmidt, Peter Jehl, Caroline McGuigan, Magdalena Dymecka, Claudia Chica, Katja Luck, Allegra Via, Andrew Chatr-Aryamontri, Niall Haslam, Gleb Grebnev, Richard J Edwards, Michel O Steinmetz, Heike Meiselbach, Francesca Diella, and Toby J Gibson. ELM–the database of eukaryotic linear motifs. *Nucleic acids research*, 40(Database issue):D242–51, Jan 2012.

[162] A J Muslin, J W Tanner, P M Allen, and A S Shaw. Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine. *Cell*, 84(6):889–97, Mar 1996.

[163] Bhaswati Pandit, Anna Sarkozy, Len A Pennacchio, Claudio Carta, Kimihiko Oishi, Simone Martinelli, Edgar A Pogna, Wendy Schackwitz, Anna Ustaszewska, Andrew Landstrom, J Martijn Bos, Steve R Ommen, Giorgia Esposito, Francesca Lepri, Christian Faul, Peter Mundel, Juan P López Siguero, Romano Tenconi, Angelo Selicorni, Cesare Rossi, Laura Mazzanti, Isabella Torrente, Bruno Marino, Maria C Digilio, Giuseppe Zampino, Michael J Ackerman, Bruno Dallapiccola, Marco Tartaglia, and Bruce D Gelb. Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nature genetics*, 39(8):1007–12, Aug 2007.

[164] E Kalay, A P M de Brouwer, R Caylan, S B Nabuurs, B Wollnik, A Karaguzel, J G A M Heister, H Erdol, F P M Cremers, C W R J Cremers, H G Brunner, and H Kremer. A novel D458V mutation in the SANS PDZ binding motif causes atypical Usher syndrome. *Journal of molecular medicine (Berlin, Germany)*, 83(12):1025–32, Dec 2005.

[165] C B Anfinsen. Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(4096):223–30, Jul 1973.

[166] C Clementi, H Nymeyer, and J N Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *Journal of molecular biology*, 298(5):937–53, May 2000.

[167] C Nick Pace, Hailong Fu, Katrina Lee Fryar, John Landua, Saul R Trevino, David Schell, Richard L Thurlkill, Satoshi Imura, J Martin Scholtz, Ketan Gajiwala, Jozef Sevcik, Lubica Urbanikova, Jeffery K Myers, Kazufumi Takano, Eric J Hebert, Bret A Shirley, and Gerald R Grimsley. Contribution of hydrogen bonds to protein stability. *Protein science : a publication of the Protein Society*, 23(5):652–61, May 2014.

[168] C Nick Pace, J Martin Scholtz, and Gerald R Grimsley. Forces stabilizing proteins. *FEBS letters*, 588(14):2177–84, Jun 2014.

[169] D E Anderson, W J Becktel, and F W Dahlquist. pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry*, 29(9):2403–8, Mar 1990.

[170] I K McDonald and J M Thornton. Satisfying hydrogen bonding potential in proteins. *Journal of molecular biology*, 238(5):777–93, May 1994.

[171] Catherine A Royer. Probing protein folding and conformational transitions with fluorescence. *Chemical reviews*, 106(5):1769–84, May 2006.

[172] Roy M Daniel and Michael J Danson. Temperature and the catalytic activity of enzymes: a fresh understanding. *FEBS letters*, 587(17):2738–43, Sep 2013.

[173] Romain A Studer, Pascal-Antoine Christin, Mark A Williams, and Christine A Orengo. Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proceedings of the National Academy of Sciences of the United States of America*, 111(6):2223–8, Feb 2014.

[174] A A Pakula and R T Sauer. Genetic analysis of protein stability and function. *Annual review of genetics*, 23:289–310, 1989.

[175] K Abdulla Bava, M Michael Gromiha, Hatsuho Uedaira, Koji Kitajima, and Akinori Sarai. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic acids research*, 32(Database issue):D120–1, Jan 2004.

[176] K Nishikawa, S Ishino, H Takenaka, N Norioka, T Hirai, T Yao, and Y Seto. Constructing a protein mutant database. *Protein engineering*, 7(5):733, May 1994.

[177] Peter D Stenson, Matthew Mort, Edward V Ball, Katy Shaw, Andrew Phillips, and David N Cooper. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human genetics*, 133(1):1–9, Jan 2014.

[178] S B Prusiner. Molecular biology of prion diseases. *Science (New York, N.Y.)*, 252(5012):1515–22, Jun 1991.

[179] J Collinge. Prion diseases of humans and animals: their causes and molecular basis. *Annual review of neuroscience*, 24:519–50, 2001.

[180] Sivakumar Boopathy, Tania V Silvas, Maeve Tischbein, Silvia Jansen, Shivender M Shandilya, Jill A Zitzewitz, John E Landers, Bruce L Goode, Celia A Schiffer, and Daryl A Bosco. Structural basis for mutation-induced destabilization of profilin 1 in ALS. *Proceedings of the National Academy of Sciences of the United States of America*, 112(26):7984–9, Jun 2015.

[181] Elizabeth P Rakoczy, Christina Kiel, Richard McKeone, François Stricher, and Luis Serrano. Analysis of disease-linked rhodopsin mutations based on structure, function, and protein stability calculations. *Journal of molecular biology*, 405(2):584–606, Jan 2011.

[182] William Lin and Un Jung Kang. Characterization of PINK1 processing, stability, and subcellular localization. *Journal of neurochemistry*, 106(1):464–74, Jul 2008.

[183] Karen Nuytemans, Jessie Theuns, Marc Cruts, and Christine Van Broeckhoven. Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: a mutation update. *Human mutation*, 31(7):763–80, Jul 2010.

[184] Shawn Witham, Kyoko Takano, Charles Schwartz, and Emil Alexov. A missense mutation in CLIC2 associated with intellectual disability is predicted by in silico modeling to affect protein stability and dynamics. *Proteins*, 79(8):2444–54, Aug 2011.

[185] Kyoko Takano, Dan Liu, Patrick Tarpey, Esther Gallant, Alex Lam, Shawn Witham, Emil Alexov, Alka Chaubey, Roger E Stevenson, Charles E Schwartz, Philip G Board, and Angela F Dulhunty. An X-linked channelopathy with cardiomegaly due to a CLIC2 mutation enhancing ryanodine receptor channel activity. *Human molecular genetics*, 21(20):4497–507, Oct 2012.

[186] Shaolei Teng, Anand K Srivastava, and Liangjiang Wang. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC genomics*, 11 Suppl 2:S5, Nov 2010.

[187] Lukas Folkman, Bela Stantic, Abdul Sattar, and Yaoqi Zhou. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. *Journal of molecular biology*, 428(6):1394–1405, Mar 2016.

[188] Liang-Tsung Huang, M Michael Gromiha, and Shinn-Ying Ho. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics (Oxford, England)*, 23(10):1292–3, May 2007.

[189] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. *Nucleic acids research*, 33(Web Server issue):W382–8, Jul 2005.

[190] Josef Laimer, Heidi Hofer, Marko Fritz, Stefan Wegenkittl, and Peter Lackner. MAESTRO–multi agent stability prediction upon point mutations. *BMC bioinformatics*, 16:116, Apr 2015.

[191] Shuangye Yin, Feng Ding, and Nikolay V Dokholyan. Eris: an automated estimator of protein stability. *Nature methods*, 4(6):466–7, Jun 2007.

[192] Catherine L Worth, Robert Preissner, and Tom L Blundell. SDM–a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research*, 39(Web Server issue):W215–22, Jul 2011.

[193] D Gilis and M Rooman. PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein engineering*, 13(12):849–56, Dec 2000.

[194] Wei Guan, Arkadas Ozakin, Alexander Gray, Jose Borreguero, Shashi Pandit, Anna Jagielska, Liliana Wroblewska, and Jeffrey Skolnick. Learning Protein Folding Energy Functions. *Proceedings. IEEE International Conference on Data Mining*, pages 1062–1067, Dec 2011.

[195] Emidio Capriotti, Piero Fariselli, and Rita Casadio. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research*, 33(Web Server issue):W306–10, Jul 2005.

[196] Giulia Gonnelli, Marianne Rooman, and Yves Dehouck. Structure-based mutant stability predictions on proteins of unknown structure. *Journal of biotechnology*, 161(3):287–93, Oct 2012.

[197] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H-Werner Mewes. CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic acids research*, 38(Database issue):D497–501, Jan 2010.

[198] D Xu, C J Tsai, and R Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein engineering*, 10(9):999–1012, Sep 1997.

[199] S Jones and J M Thornton. Analysis of protein-protein interaction sites using surface patches. *Journal of molecular biology*, 272(1):121–32, Sep 1997.

[200] L Lo Conte, C Chothia, and J Janin. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285(5):2177–98, Feb 1999.

[201] Changhui Yan, Feihong Wu, Robert L Jernigan, Drena Dobbs, and Vasant Honavar. Characterization of protein-protein interfaces. *The protein journal*, 27(1):59–70, Jan 2008.

[202] M C Lawrence and P M Colman. Shape complementarity at protein/protein interfaces. *Journal of molecular biology*, 234(4):946–50, Dec 1993.

[203] Jemima Hoskins, Simon Lovell, and Tom L Blundell. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein science : a publication of the Protein Society*, 15(5):1017–29, May 2006.

[204] S Jones and J M Thornton. Prediction of protein-protein interaction sites using patch analysis. *Journal of molecular biology*, 272(1):133–43, Sep 1997.

[205] Roberto Mosca, Arnaud Céol, and Patrick Aloy. Interactome3D: adding structural details to protein networks. *Nature methods*, 10(1):47–53, Jan 2013.

[206] Jose M Duarte, Adam Srebniak, Martin A Schärer, and Guido Capitani. Protein interface classification by evolutionary analysis. *BMC bioinformatics*, 13:334, Dec 2012.

[207] Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joël Janin. A dissection of specific and non-specific protein-protein interfaces. *Journal of molecular biology*, 336(4):943–55, Feb 2004.

[208] Emmanuel D Levy. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of molecular biology*, 403(4):660–70, Nov 2010.

[209] E Kube, T Becker, K Weber, and V Gerke. Protein-protein interaction studied by site-directed mutagenesis. Characterization of the annexin II-binding site on p11, a member of the S100 protein family. *The Journal of biological chemistry*, 267(20):14175–82, Jul 1992.

[210] Timothy A Whitehead, Aaron Chevalier, Yifan Song, Cyrille Dreyfus, Sarel J Fleishman, Cecilia De Mattos, Chris A Myers, Hetunandan Kamisetty, Patrick Blair, Ian A Wilson, and David Baker. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature biotechnology*, 30(6):543–8, May 2012.

[211] Douglas M Fowler, Carlos L Araya, Sarel J Fleishman, Elizabeth H Kellogg, Jason J Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence-function relationships. *Nature methods*, 7(9):741–6, Sep 2010.

[212] Carlos L Araya and Douglas M Fowler. Deep mutational scanning: assessing protein function on a massive scale. *Trends in biotechnology*, 29(9):435–42, Sep 2011.

[213] Peng Xiong, Chengxin Zhang, Wei Zheng, and Yang Zhang. BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts. *Journal of molecular biology*, 429(3):426–434, Feb 2017.

[214] Jeffrey R Brender and Yang Zhang. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLoS computational biology*, 11(10):e1004494, Oct 2015.

[215] Minghui Li, Franco L Simonetti, Alexander Goncearenco, and Anna R Panchenko. MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. *Nucleic acids research*, 44(W1):W494–501, Jul 2016.

[216] Iain H Moal and Juan Fernández-Recio. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics (Oxford, England)*, 28(20):2600–7, Oct 2012.

[217] Sherlyn Jemimah, K Yugandhar, and M Michael Gromiha. PROXiMATE: a database of mutant protein-protein complex thermodynamics and kinetics. *Bioinformatics (Oxford, England)*, 33(17):2787–2788, Sep 2017.

[218] Timothy R Siegert, Michael J Bird, Kamlesh M Makwana, and Joshua A Kritzer. Analysis of Loops that Mediate Protein-Protein Interactions and Translation into Submicromolar Inhibitors. *Journal of the American Chemical Society*, 138(39):12876–12884, Oct 2016.

[219] Ruben Marrero, Ramiro Rodríguez Limardo, Elisa Carrillo, Guido A König, and Adrián G Turjanski. A computational study of the interaction of the foot and mouth disease virus VP1 with monoclonal antibodies. *Journal of immunological methods*, 425:51–7, Oct 2015.

[220] Dmitry V Chouljenko, Nithya Jambunathan, Vladimir N Chouljenko, Misagh Naderi, Michal Brylinski, John R Caskey, and Konstantin G Kousoulas. Herpes Simplex Virus 1 UL37 Protein Tyrosine Residues Conserved among All Alphaherpesviruses Are Required for Interactions with Glycoprotein K, Cytoplasmic Virion Envelopment, and Infectious Virus Production. *Journal of virology*, 90(22):10351–10361, Nov 2016.

[221] Raik Grünberg, Julia V Burnier, Tony Ferrar, Violeta Beltran-Sastre, François Stricher, Almer M van der Sloot, Raquel Garcia-Olivas, Arrate Mallabiabarrena, Xavier Sanjuan, Timo Zimmermann, and Luis Serrano. Engineering of weak helper interactions for high-efficiency FRET probes. *Nature methods*, 10(10):1021–7, Oct 2013.

[222] Pelagia Deriziotis, Brian J O'Roak, Sarah A Graham, Sara B Estruch, Danai Dimitropoulou, Raphael A Bernier, Jennifer Gerdts, Jay Shendure, Evan E Eichler, and Simon E Fisher. De novo TBR1 mutations in sporadic autism disrupt protein functions. *Nature communications*, 5:4954, Sep 2014.

[223] Emma L Baple, Helen Chambers, Harold E Cross, Heather Fawcett, Yuka Nakazawa, Barry A Chioza, Gaurav V Harlalka, Sahar Mansour, Ajith Sreekantan-Nair, Michael A Patton, Martina Muggenthaler, Phillip Rich, Karin Wagner, Roselyn Coblentz, Constance K Stein, James I Last, A Malcolm R Taylor, Andrew P Jackson, Tomoo Ogi, Alan R Lehmann, Catherine M Green, and Andrew H Crosby. Hypomorphic PCNA mutation underlies a human DNA repair disorder. *The Journal of clinical investigation*, 124(7):3137–46, Jul 2014.

[224] Caroline M Duffy, Brendan J Hilbert, and Brian A Kelch. A Disease-Causing Variant in PCNA Disrupts a Promiscuous Protein Binding Site. *Journal of molecular biology*, 428(6):1023–1040, Mar 2016.

[225] Eduard Porta-Pardo, Luz Garcia-Alonso, Thomas Hrabe, Joaquin Dopazo, and Adam Godzik. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS computational biology*, 11(10):e1004518, Oct 2015.

[226] H Billur Engin, Jason F Kreisberg, and Hannah Carter. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PloS one*, 11(4):e0152929, 2016.

[227] F R Blattner, G Plunkett, C A Bloch, N T Perna, V Burland, M Riley, J Collado-Vides, J D Glasner, C K Rode, G F Mayhew, J Gregor, N W Davis, H A Kirkpatrick, M A Goeden, D J Rose, B Mau, and Y Shao. The complete genome sequence of Escherichia coli K-12. *Science (New York, N.Y.)*, 277(5331):1453–62, Sep 1997.

[228] A Alejandra Klauer and Ambro van Hoof. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. *Wiley interdisciplinary reviews. RNA*, 3(5):649–60, 2012.

[229] Mario H Bengtson and Claudio A P Joazeiro. Role of a ribosome-associated E3 ubiquitin ligase in protein quality control. *Nature*, 467(7314):470–3, Sep 2010.

[230] Lyudmila N Dimitrova, Kazushige Kuroha, Tsuyako Tatematsu, and Toshifumi Inada. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *The Journal of biological chemistry*, 284(16):10343–52, Apr 2009.

[231] Geurt Schilders, Erwin van Dijk, Reinout Raijmakers, and Ger J M Pruijn. Cell and molecular biology of the exosome: how to make or break an RNA. *International review of cytology*, 251:159–208, 2006.

[232] Yoon Ki Kim, Luc Furic, Luc Desgroseillers, and Lynne E Maquat. Mammalian Staufen1 recruits Upf1 to specific mRNA 3'UTRs so as to elicit mRNA decay. *Cell*, 120(2):195–208, Jan 2005.

[233] Olga Anczuków, Mark D Ware, Monique Buisson, Almoutassem B Zetoune, Dominique Stoppa-Lyonnet, Olga M Sinilnikova, and Sylvie Mazoyer. Does the nonsense-mediated mRNA decay mechanism prevent the synthesis of truncated BRCA1, CHK2, and p53 proteins? *Human mutation*, 29(1):65–73, Jan 2008.

[234] Maila Giannandrea, Fabrizia C Guarnieri, Niels H Gehring, Elena Monzani, Fabio Benfenati, Andreas E Kulozik, and Flavia Valtorta. Nonsense-mediated mRNA decay and loss-of-function of the protein underlie the X-linked epilepsy associated with the W356× mutation in synapsin I. *PloS one*, 8(6):e67724, 2013.

[235] Nadia Amrani, Matthew S Sachs, and Allan Jacobson. Early nonsense: mRNA decay solves a translational problem. *Nature reviews. Molecular cell biology*, 7(6):415–25, Jun 2006.

[236] Vinay K Nagarajan, Christopher I Jones, Sarah F Newbury, and Pamela J Green. XRN 5'→3' exoribonucleases: structure, mechanisms and functions. *Biochimica et biophysica acta*, 1829(6-7):590–603, 2013.

[237] M Schloesser, S Arleth, U Lenz, R M Bertele, and J Reiss. A cystic fibrosis patient with the nonsense mutation G542X and the splice site mutation 1717-1. *Journal of medical genetics*, 28(12):878–80, Dec 1991.

[238] Jie Sun, Ziqi Hao, Hunjin Luo, Chufeng He, Lingyun Mei, Yalan Liu, Xueping Wang, Zhijie Niu, Hongsheng Chen, Jia-Da Li, and Yong Feng. Functional analysis of a nonstop mutation in MITF gene identified in a patient with Waardenburg syndrome type 2. *Journal of human genetics*, 62(7):703–709, Jul 2017.

[239] Irene Sargiannidou, Gun-Ha Kim, Styliana Kyriakoudi, Baik-Lin Eun, and Kleopas A Kleopa. A start codon CMT1X mutation associated with transient encephalomyelitis causes complete loss of Cx32. *Neurogenetics*, 16(3):193–200, Jul 2015.

[240] Kim M Keeling and David M Bedwell. Suppression of nonsense mutations as a therapeutic approach to treat genetic diseases. *Wiley interdisciplinary reviews. RNA*, 2(6):837–52, 2011.

[241] Kim M Keeling. Nonsense Suppression as an Approach to Treat Lysosomal Storage Diseases. *Diseases (Basel, Switzerland)*, 4(4), Dec 2016.

[242] Rasmus Nielsen. Molecular signatures of natural selection. *Annual review of genetics*, 39:197–218, 2005.

[243] Thomas A Isenbarger, Christopher E Carr, Sarah Stewart Johnson, Michael Finney, George M Church, Walter Gilbert, Maria T Zuber, and Gary Ruvkun. The most conserved genome segments for life detection on Earth and other planets. *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life*, 38(6):517–33, Dec 2008.

[244] Pauline C Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–4, Jul 2003.

[245] Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, 6(12):e1001025, Dec 2010.

[246] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*, Chapter 7:Unit7.20, Jan 2013.

[247] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10, Oct 1990.

[248] P C Ng and S Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–74, May 2001.

[249] Yongwook Choi, Gregory E Sims, Sean Murphy, Jason R Miller, and Agnes P Chan. Predicting the functional effect of amino acid substitutions and indels. *PloS one*, 7(10):e46688, 2012.

[250] Steven Henikoff and Luca Comai. Single-nucleotide mutations for plant functional genomics. *Annual review of plant biology*, 54:375–401, 2003.

[251] Anthony G Doran, Kim Wong, Jonathan Flint, David J Adams, Kent W Hunter, and Thomas M Keane. Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome biology*, 17(1):167, Aug 2016.

[252] Peter V Hornbeck, Jon M Kornhauser, Sasha Tkachev, Bin Zhang, Elzbieta Skrzypek, Beth Murray, Vaughan Latham, and Michael Sullivan. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research*, 40(Database issue):D261–70, Jan 2012.

[253] Holger Dinkel, Claudia Chica, Allegra Via, Cathryn M Gould, Lars J Jensen, Toby J Gibson, and Francesca Diella. Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic acids research*, 39(Database issue):D261–7, Jan 2011.

[254] T S Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C J Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K Kashyap, Riaz Mohmood, Y L Ramachandra, V Krishna, B Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. *Nucleic acids research*, 37(Database issue):D767–72, Jan 2009.

[255] Bostjan Kobe, Thorsten Kampmann, Jade K Forwood, Pawel Listwan, and Ross I Brinkworth. Substrate specificity of protein kinases and computational prediction of substrates. *Biochimica et biophysica acta*, 1754(1-2):200–9, Dec 2005.

[256] Robert D Finn, Jody Clements, and Sean R Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39(Web Server issue):W29–37, Jul 2011.

[257] Victor Neduva, Rune Linding, Isabelle Su-Angrand, Alexander Stark, Federico de Masi, Toby J Gibson, Joe Lewis, Luis Serrano, and Robert B Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS biology*, 3(12):e405, Dec 2005.

[258] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(Database issue):D561–8, Jan 2011.

[259] Natarajan Kannan and Andrew F Neuwald. Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha. *Protein science : a publication of the Protein Society*, 13(8):2059–77, Aug 2004.

[260] Guozhi Zhu, Koichi Fujii, Natalya Belkina, Yin Liu, Michael James, Juan Herrero, and Stephen Shaw. Exceptional disfavor for proline at the P + 1 position among AGC and CAMK kinases establishes reciprocal specificity between them and the proline-directed kinases. *The Journal of biological chemistry*, 280(11):10743–8, Mar 2005.

[261] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue):D535–9, Jan 2006.

[262] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars J Jensen. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue):D808–15, Jan 2013.

[263] Christian von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research*, 33(Database issue):D433–7, Jan 2005.

[264] Suraj Peri, J Daniel Navarro, Troels Z Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, T K B Gandhi, K N Chandrika, Nandan Deshpande, Shubha Suresh, B P Rashmi, K Shanker, N Padma, Vidya Niranjan, H C Harsha, Naveen Talreja, B M Vrushabendra, M A Ramya, A J Yatish, Mary Joy, H N Shivashankar, M P Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Neelanjana Ghosh, R Saravana, Sreenath Chandran, Sujatha Mohan, Chandra Kiran Jonnalagadda, C K Prasad, Chandan Kumar-Sinha, Krishna S Deshpande, and Akhilesh Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic acids research*, 32(Database issue):D497–501, Jan 2004.

[265] David M A Martin, Diego Miranda-Saavedra, and Geoffrey J Barton. Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic acids research*, 37(Database issue):D244–50, Jan 2009.

[266] Michael F Chou and Daniel Schwartz. Biological sequence motif discovery using motif-x. *Current protocols in bioinformatics*, Chapter 13:Unit 13.15–24, Sep 2011.

[267] T Tanoue, M Adachi, T Moriguchi, and E Nishida. A conserved docking motif in MAP kinases common to substrates, activators and regulators. *Nature cell biology*, 2(2):110–6, Feb 2000.

[268] A D Sharrocks, S H Yang, and A Galanis. Docking domains and substrate-specificity determination for MAP kinases. *Trends in biochemical sciences*, 25(9):448–53, Sep 2000.

[269] M Wang, M Weiss, M Simonovic, G Haertinger, S P Schrimpf, M O Hengartner, and C von Mering. PaxDb, a database of protein abundance averages across all three domains of life. *Molecular & cellular proteomics : MCP*, 11(8):492–500, Aug 2012.

[270] Haruna Imamura, Naoyuki Sugiyama, Masaki Wakabayashi, and Yasushi Ishihama. Large-scale identification of phosphorylation sites for profiling protein kinase selectivity. *Journal of proteome research*, 13(7):3410–9, Jul 2014.

[271] Jesse J Lipp, Michael C Marvin, Kevan M Shokat, and Christine Guthrie. SR protein kinases promote splicing of nonconsensus introns. *Nature structural & molecular biology*, 22(8):611–7, Aug 2015.

[272] H Fu, R R Subramanian, and S C Masters. 14-3-3 proteins: structure, function, and regulation. *Annual review of pharmacology and toxicology*, 40:617–47, 2000.

[273] Catherine Johnson, Sandra Crowther, Margaret J Stafford, David G Campbell, Rachel Toth, and Carol MacKintosh. Bioinformatic and experimental survey of 14-3-3-binding sites. *The Biochemical journal*, 427(1):69–78, Mar 2010.

[274] Annika E Wallberg, Soichiro Yamamura, Sohail Malik, Bruce M Spiegelman, and Robert G Roeder. Coordination of p300-mediated chromatin remodeling and TRAP/-mediator function through coactivator PGC-1alpha. *Molecular cell*, 12(5):1137–49, Nov 2003.

[275] Manuela Delvecchio, Jonathan Gaucher, Carmen Aguilar-Gurrieri, Esther Ortega, and Daniel Panne. Structure of the p300 catalytic core and implications for chromatin targeting and HAT regulation. *Nature structural & molecular biology*, 20(9):1040–6, Sep 2013.

[276] S K Hanks, A M Quinn, and T Hunter. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science (New York, N.Y.)*, 241(4861):42–52, Jul 1988.

[277] Kevin P O'Brien, Maido Remm, and Erik L L Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research*, 33(Database issue):D476–80, Jan 2005.

[278] Edward L Huttlin, Mark P Jedrychowski, Joshua E Elias, Tapasree Goswami, Ramin Rad, Sean A Beausoleil, Judit Villén, Wilhelm Haas, Mathew E Sowa, and Steven P Gygi. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143(7):1174–89, Dec 2010.

[279] A E Kel, E Gössling, I Reuter, E Cheremushkin, O V Kel-Margoulis, and E Wingender. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic acids research*, 31(13):3576–9, Jul 2003.

[280] Naoyuki Sugiyama, Takeshi Masuda, Kosaku Shinoda, Akihiro Nakamura, Masaru Tomita, and Yasushi Ishihama. Phosphopeptide enrichment by aliphatic hydroxy acid-modified metal oxide chromatography for nano-LC-MS/MS in proteomics applications. *Molecular & cellular proteomics : MCP*, 6(6):1103–9, Jun 2007.

[281] Flavio Meggio and Lorenzo A Pinna. One-thousand-and-one substrates of protein kinase CK2? *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 17(3):349–68, Mar 2003.

[282] Mariana Lemos Duarte, Darlene Aparecida Pena, Felipe Augusto Nunes Ferraz, Denise Aparecida Berti, Tiago José Paschoal Sobreira, Helio Miranda Costa-Junior, Munira Muhammad Abdel Baqui, Marie-Hélène Disatnik, José Xavier-Neto, Paulo Sérgio Lopes de Oliveira, and Deborah Schechtman. Protein folding creates structure-based, noncontiguous consensus phosphorylation motifs recognized by kinases. *Science signaling*, 7(350):ra105, Nov 2014.

[283] Pedro Beltrao, Véronique Albanèse, Lillian R Kenner, Danielle L Swaney, Alma Burlingame, Judit Villén, Wendell A Lim, James S Fraser, Judith Frydman, and Nevan J Krogan. Systematic functional prioritization of protein posttranslational modifications. *Cell*, 150(2):413–25, Jul 2012.

[284] Christian R Landry, Emmanuel D Levy, and Stephen W Michnick. Weak functional constraints on phosphoproteomes. *Trends in genetics : TIG*, 25(5):193–7, May 2009.

[285] Alan M Moses and Christian R Landry. Moving from transcriptional to phosphoevolution: generalizing regulatory evolution? *Trends in genetics : TIG*, 26(11):462–7, Nov 2010.

[286] Christopher R Baker, Brian B Tuch, and Alexander D Johnson. Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proceedings of the National Academy of Sciences of the United States of America*, 108(18):7493–8, May 2011.

[287] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(Database issue):D91–4, Jan 2004.

[288] Ashley K Tehranchi, Marsha Myrthil, Trevor Martin, Brian L Hie, David Golan, and Hunter B Fraser. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell*, 165(3):730–41, Apr 2016.

[289] Helena Kilpinen, Sebastian M Waszak, Andreas R Gschwind, Sunil K Raghav, Robert M Witwicki, Andrea Orioli, Eugenia Migliavacca, Michaël Wiederkehr, Maria Gutierrez-Arcelus, Nikolaos I Panousis, Alisa Yurovsky, Tuuli Lappalainen, Luciana Romano-Palumbo, Alexandra Planchon, Deborah Bielser, Julien Bryois, Ismael Padioleau, Gilles Udin, Sarah Thurnheer, David Hacker, Leighton J Core, John T Lis,

Nouria Hernandez, Alexandre Reymond, Bart Deplancke, and Emmanouil T Dermitzakis. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science (New York, N.Y.)*, 342(6159):744–7, Nov 2013.

[290] Matthew T Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, Harmen J Bussemaker, Quaid D Morris, Martha L Bulyk, Gustavo Stolovitzky, and Timothy R Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology*, 31(2):126–34, Feb 2013.

[291] Narayan Jayaram, Daniel Usvyat, and Andrew C R Martin. Evaluating tools for transcription factor binding site prediction. *BMC bioinformatics*, Nov 2016.

[292] Alexander Kaplun, Mathias Krull, Karthick Lakshman, Volker Matys, Birgit Lewicki, and Jennifer D Hogan. Establishing and validating regulatory regions for variant annotation and expression analysis. *BMC genomics*, 17 Suppl 2:393, 06 2016.

[293] Mauro A A Castro, Ines de Santiago, Thomas M Campbell, Courtney Vaughn, Theresa E Hickey, Edith Ross, Wayne D Tilley, Florian Markowetz, Bruce A J Ponder, and Kerstin B Meyer. Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nature genetics*, 48(1):12–21, Jan 2016.

[294] Haoyang Zeng, Tatsunori Hashimoto, Daniel D Kang, and David K Gifford. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics (Oxford, England)*, 32(4):490–6, Feb 2016.

[295] Timothy E Reddy, Jason Gertz, Florencia Pauli, Katerina S Kucera, Katherine E Varley, Kimberly M Newberry, Georgi K Marinov, Ali Mortazavi, Brian A Williams, Lingyun Song, Gregory E Crawford, Barbara Wold, Huntington F Willard, and Richard M Myers. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research*, 22(5):860–9, May 2012.

[296] Yingying Wei, Xia Li, Qian-fei Wang, and Hongkai Ji. iASeq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. *BMC genomics*, 13:681, Nov 2012.

[297] Swneke D Bailey, Carl Virtanen, Benjamin Haibe-Kains, and Mathieu Lupien. ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. *Bioinformatics (Oxford, England)*, 31(18):3057–9, Sep 2015.

[298] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A Kosmicki, Laramie E Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M Peloso, Ryan Poplin, Manuel A Rivas, Valentin Ruano-Rubio, Samuel A Rose, Douglas M Ruderfer, Khalid Shakir, Peter D Stenson, Christine Stevens, Brett P Thomas, Grace Tiao, Maria T Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey

Donnelly, Roberto Elosua, Jose C Florez, Stacey B Gabriel, Gad Getz, Stephen J Glatt, Christina M Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M Neale, Aarno Palotie, Shaun M Purcell, Danish Saleheen, Jeremiah M Scharf, Pamela Sklar, Patrick F Sullivan, Jaakko Tuomilehto, Ming T Tsuang, Hugh C Watkins, James G Wilson, Mark J Daly, and Daniel G MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–91, 08 2016.

[299] Nayanah Siva. 1000 Genomes project. *Nature biotechnology*, 26(3):256, Mar 2008.

[300] Wenqing Fu, Timothy D O'Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M Leal, Stacey Gabriel, Mark J Rieder, David Altshuler, Jay Shendure, Deborah A Nickerson, Michael J Bamshad, and Joshua M Akey. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–20, Jan 2013.

[301] Graham R S Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of noncoding sequence variants. *Nature methods*, 11(3):294–6, Mar 2014.

[302] Iuliana Ionita-Laza, Kenneth McCallum, Bin Xu, and Joseph D Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics*, 48(2):214–20, Feb 2016.

[303] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–5, Mar 2014.

[304] Ekta Khurana, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner, Lucas Lochovsky, Jieming Chen, Arif Harmanci, Jishnu Das, Alexej Abyzov, Suganthi Balasubramanian, Kathryn Beal, Dimple Chakravarty, Daniel Challis, Yuan Chen, Declan Clarke, Laura Clarke, Fiona Cunningham, Uday S Evani, Paul Flicek, Robert Fragoza, Erik Garrison, Richard Gibbs, Zeynep H Gümüş, Javier Herrero, Naoki Kitabayashi, Yong Kong, Kasper Lage, Vaja Liluashvili, Steven M Lipkin, Daniel G MacArthur, Gabor Marth, Donna Muzny, Tune H Pers, Graham R S Ritchie, Jeffrey A Rosenfeld, Cristina Sisu, Xiaomu Wei, Michael Wilson, Yali Xue, Fuli Yu, Emmanouil T Dermitzakis, Haiyuan Yu, Mark A Rubin, Chris Tyler-Smith, and Mark Gerstein. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science (New York, N.Y.)*, 342(6154):1235587, Oct 2013.

[305] Yao Fu, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y Yip, Ekta Khurana, and Mark Gerstein. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome biology*, 15(10):480, 2014.

[306] Dilmi Perera, Diego Chacon, Julie A I Thoms, Rebecca C Poulos, Adam Shlien, Dominik Beck, Peter J Campbell, John E Pimanda, and Jason W H Wong. OncoCis: annotation of cis-regulatory mutations in cancer. *Genome biology*, 15(10):485, 2014.

[307] Joshua L Payne and Andreas Wagner. Mechanisms of mutational robustness in transcriptional regulation. *Frontiers in genetics*, 6:322, 2015.

[308] Philip Machanick and Timothy L Bailey. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics (Oxford, England)*, 27(12):1696–7, Jun 2011.

[309] Ulf Schaefer, Sebastian Schmeier, and Vladimir B Bajic. TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic acids research*, 39(Database issue):D106–10, Jan 2011.

[310] Cornelia G Spruijt, Felix Gnerlich, Arne H Smits, Toni Pfaffeneder, Pascal W T C Jansen, Christina Bauer, Martin Münzel, Mirko Wagner, Markus Müller, Fariha Khan, H Christian Eberl, Anneloes Mensinga, Arie B Brinkman, Konstantin Lephikov, Udo Müller, Jörn Walter, Rolf Boelens, Hugo van Ingen, Heinrich Leonhardt, Thomas Carell, and Michiel Vermeulen. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell*, 152(5):1146–59, Feb 2013.

[311] Igor N Zelko, Michael R Mueller, and Rodney J Folz. CpG methylation attenuates Sp1 and Sp3 binding to the human extracellular superoxide dismutase promoter and regulates its cell-specific expression. *Free radical biology & medicine*, 48(7):895–904, Apr 2010.

[312] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268):1248–53, Oct 2009.

[313] A J Whitmarsh and R J Davis. Regulation of transcription factor function by phosphorylation. *Cellular and molecular life sciences : CMLS*, 57(8-9):1172–83, Aug 2000.

[314] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome research*, 12(6):996–1006, Jun 2002.

[315] Ge Tan and Boris Lenhard. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics (Oxford, England)*, 32(10):1555–6, May 2016.

[316] H. Pagès, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms*, 2016. R package version 2.42.1.

[317] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164, Sep 2010.

[318] Jan Grau, Ivo Grosse, and Jens Keilwagen. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics (Oxford, England)*, 31(15):2595–7, Aug 2015.

[319] Raluca Gordân, Ning Shen, Iris Dror, Tianyin Zhou, John Horton, Remo Rohs, and Martha L Bulyk. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell reports*, 3(4):1093–104, Apr 2013.

[320] Silvia Domcke, Anaïs Flore Bardet, Paul Adrian Ginno, Dominik Hartl, Lukas Burger, and Dirk Schübeler. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, 528(7583):575–9, Dec 2015.

[321] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*, 21(3):447–55, Mar 2011.

[322] M J Waterman, E S Stavridi, J L Waterman, and T D Halazonetis. ATM-dependent activation of p53 involves dephosphorylation and association with 14-3-3 proteins. *Nature genetics*, 19(2):175–8, Jun 1998.

[323] Aaron Arvey, Phaedra Agius, William Stafford Noble, and Christina Leslie. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome research*, 22(9):1723–34, Sep 2012.

[324] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–63, 04 2009.

[325] Ivan V Kulakovskiy, Yulia A Medvedeva, Ulf Schaefer, Artem S Kasianov, Ilya E Vorontsov, Vladimir B Bajic, and Vsevolod J Makeev. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic acids research*, 41(Database issue):D195–202, Jan 2013.

[326] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(Database issue):D1001–6, Jan 2014.

[327] Natsuhiko Kumasaka, Andrew J Knights, and Daniel J Gaffney. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature genetics*, 48(2):206–13, Feb 2016.

[328] Hui Y Xiong, Babak Alipanahi, Leo J Lee, Hannes Bretschneider, Daniele Merico, Ryan K C Yuen, Yimin Hua, Serge Gueroussov, Hamed S Najafabadi, Timothy R Hughes, Quaid Morris, Yoseph Barash, Adrian R Krainer, Nebojsa Jojic, Stephen W Scherer, Benjamin J Blencowe, and Brendan J Frey. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science (New York, N.Y.)*, 347(6218):1254806, Jan 2015.

[329] Carola Rintisch, Matthias Heinig, Anja Bauerfeind, Sebastian Schafer, Christin Mieth, Giannino Patone, Oliver Hummel, Wei Chen, Stuart Cook, Edwin Cuppen, Maria Colomé-Tatché, Frank Johannes, Ritsert C Jansen, Helen Neil, Michel Werner, Michal Pravenec, Martin Vingron, and Norbert Hubner. Natural variation of histone modification and its impact on gene expression in the rat genome. *Genome research*, 24(6):942–53, Jun 2014.

[330] M Lorch, J M Mason, R B Sessions, and A R Clarke. Effects of mutations on the thermodynamics of a protein folding reaction: implications for the mechanism of formation of the intermediate and transition states. *Biochemistry*, 39(12):3480–5, Mar 2000.

[331] P Björses, M Halonen, J J Palvimo, M Kolmer, J Aaltonen, P Ellonen, J Perheentupa, I Ulmanen, and L Peltonen. Mutations in the AIRE gene: effects on subcellular location and transactivation function of the autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy protein. *American journal of human genetics*, 66(2):378–92, Feb 2000.

[332] Matthew R Nelson, Toby Johnson, Liling Warren, Arlene R Hughes, Stephanie L Chissoe, Chun-Fang Xu, and Dawn M Waterworth. The genetics of drug efficacy: opportunities and challenges. *Nature reviews. Genetics*, 17(4):197–206, Apr 2016.

[333] Richard Labaudinière. The increasing importance of genetic variation in drug discovery and development. *Current opinion in molecular therapeutics*, 4(6):559–64, Dec 2002.

[334] Stefan Lutz. Beyond directed evolution–semi-rational protein engineering and design. *Current opinion in biotechnology*, 21(6):734–43, Dec 2010.

[335] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome biology*, 17(1):122, Jun 2016.

[336] Pooja K Strope, Daniel A Skelly, Stanislav G Kozmin, Gayathri Mahadevan, Eric A Stone, Paul M Magwene, Fred S Dietrich, and John H McCusker. The 100-genomes strains, an S. cerevisiae resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome research*, 25(5):762–74, May 2015.

[337] Brigida Gallone, Jan Steensels, Troels Prahl, Leah Soriaga, Veerle Saels, Beatriz Herrera-Malaver, Adriaan Merlevede, Miguel Roncoroni, Karin Voordeckers, Loren Miraglia, Clotilde Teiling, Brian Steffy, Maryann Taylor, Ariel Schwartz, Toby Richardson, Christopher White, Guy Baele, Steven Maere, and Kevin J Verstrepen. Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell*, 166(6):1397–1410.e16, Sep 2016.

[338] Anders Bergström, Jared T Simpson, Francisco Salinas, Benjamin Barré, Leopold Parts, Amin Zia, Alex N Nguyen Ba, Alan M Moses, Edward J Louis, Ville Mustonen, Jonas Warringer, Richard Durbin, and Gianni Liti. A high-definition view of functional genetic variation from natural yeast genomes. *Molecular biology and evolution*, 31(4):872–88, Apr 2014.

[339] Yuan O Zhu, Gavin Sherlock, and Dmitri A Petrov. Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation. *G3 (Bethesda, Md.)*, 6(8):2421–34, Aug 2016.

[340] Hongyi Zhou and Yaoqi Zhou. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins*, 54(2):315–22, Feb 2004.

[341] Jinfeng Liu, Yan Zhang, Xingye Lei, and Zemin Zhang. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome biology*, 9(4):R69, Apr 2008.

[342] Lakshmipuram S Swapna, Ramachandra M Bhaskara, Jyoti Sharma, and Narayanaswamy Srinivasan. Roles of residues in the interface of transient protein-protein complexes before complexation. *Scientific reports*, 2:334, 2012.

[343] Tianji Zhang, Brandy L Fultz, Sapna Das-Bradoo, and Anja-Katrin Bielinsky. Mapping ubiquitination sites of S. cerevisiae Mcm10. *Biochemistry and biophysics reports*, 8:212–218, Dec 2016.

[344] M S Rodriguez, J M Desterro, S Lain, D P Lane, and R T Hay. Multiple C-terminal lysine residues target p53 for ubiquitin-proteasome-mediated degradation. *Molecular and cellular biology*, 20(22):8458–67, Nov 2000.

[345] Dianqing Wu and Weijun Pan. GSK3: a multifaceted kinase in Wnt signaling. *Trends in biochemical sciences*, 35(3):161–8, Mar 2010.

[346] Anthony Mathelier, Oriol Fornes, David J Arenillas, Chih-Yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, Casper Shyr, Ge Tan, Rebecca Worsley-Hunt, Allen W Zhang, François Parcy, Boris Lenhard, Albin Sandelin, and Wyeth W Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 44(D1):D110–5, Jan 2016.

[347] Amos Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome research*, 16(8):962–72, Aug 2006.

[348] Daphne Ezer, Nicolae Radu Zabet, and Boris Adryan. Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. *Computational and structural biotechnology journal*, 10(17):63–9, Jul 2014.

[349] Lydie Lane, Ghislaine Argoud-Puy, Aurore Britan, Isabelle Cusin, Paula D Duek, Olivier Evalet, Alain Gateau, Pascale Gaudet, Anne Gleizes, Alexandre Masselot, Catherine Zwahlen, and Amos Bairoch. neXtProt: a knowledge platform for human proteins. *Nucleic acids research*, 40(Database issue):D76–83, Jan 2012.

[350] Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L L Sonnhammer, John Tate, and Marco Punta. Pfam: the protein families database. *Nucleic acids research*, 42(Database issue):D222–30, Jan 2014.

[351] Emilio Potenza, Tomás Di Domenico, Ian Walsh, and Silvio C E Tosatto. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic acids research*, 43(Database issue):D315–20, Jan 2015.

[352] Vincent A Blomen, Peter Májek, Lucas T Jae, Johannes W Bigenzahn, Joppe Nieuwenhuis, Jacqueline Staring, Roberto Sacco, Ferdy R van Diemen, Nadine Olk, Alexey Stukalov, Caleb Marceau, Hans Janssen, Jan E Carette, Keiryn L Bennett, Jacques Colinge, Giulio Superti-Furga, and Thijn R Brummelkamp. Gene essentiality and synthetic lethality in haploid human cells. *Science (New York, N.Y.)*, 350(6264):1092–6, Nov 2015.

[353] Traver Hart, Megha Chandrashekhar, Michael Aregger, Zachary Steinhart, Kevin R Brown, Graham MacLeod, Monika Mis, Michal Zimmermann, Amelie Fradet-Turcotte, Song Sun, Patricia Mero, Peter Dirks, Sachdev Sidhu, Frederick P Roth, Olivia S Rissland, Daniel Durocher, Stephane Angers, and Jason Moffat. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, 163(6):1515–26, Dec 2015.

[354] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(Database issue):D980–5, Jan 2014.

[355] Rob Jelier, Jennifer I Semple, Rosa Garcia-Verdugo, and Ben Lehner. Predicting phenotypic variation in yeast from individual genome sequences. *Nature genetics*, 43(12):1270–4, Nov 2011.

[356] S D Cramer, P M Ferree, K Lin, D S Milliner, and R P Holmes. The gene encoding hydroxypyruvate reductase (GRHPR) is mutated in patients with primary hyperoxaluria type II. *Human molecular genetics*, 8(11):2063–9, Oct 1999.

[357] David P Cregeen, Emma L Williams, Sally Hulton, and Gill Rumsby. Molecular analysis of the glyoxylate reductase (GRHPR) gene and description of mutations underlying primary hyperoxaluria type 2. *Human mutation*, 22(6):497, Dec 2003.

[358] Michael P S Booth, R Conners, Gill Rumsby, and R Leo Brady. Structural basis of substrate specificity in human glyoxylate reductase/hydroxypyruvate reductase. *Journal of molecular biology*, 360(1):178–89, Jun 2006.

[359] N A Alam, A J Rowan, N C Wortham, P J Pollard, M Mitchell, J P Tyrer, E Barclay, E Calonje, S Manek, S J Adams, P W Bowers, N P Burrows, R Charles-Holmes, L J Cook, B M Daly, G P Ford, L C Fuller, S E Hadfield-Jones, N Hardwick, A S Highet, M Keefe, S P MacDonald-Hull, E D A Potts, M Crone, S Wilkinson, F Camacho-Martinez, S Jablonska, R Ratnavel, A MacDonald, R J Mann, K Grice, G Guillet, M S Lewis-Jones, H McGrath, D C Seukeran, P J Morrison, S Fleming, S Rahman, D Kelsell, I Leigh, S Olpin, and I P M Tomlinson. Genetic and functional analyses of FH mutations in multiple cutaneous and uterine leiomyomatosis, hereditary leiomyomatosis and renal cancer, and fumarate hydratase deficiency. *Human molecular genetics*, 12(11):1241–52, Jun 2003.

[360] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, Gemma Hoad, Mikyung Jang, Nima Pakseresht, Sheila Plaister, Rajesh Radhakrishnan, Kethi Reddy, Siamak Sobhany, Petra Ten Hoopen, Robert Vaughan, Vadim Zalunin, and Guy Cochrane. The European Nucleotide Archive. *Nucleic acids research*, 39(Database issue):D28–31, Jan 2011.

[361] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[362] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

[363] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–303, Sep 2010.

[364] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, Aug 2009.

[365] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*, 2012.

[366] Valerie Obenchain, Michael Lawrence, Vincent Carey, Stephanie Gogarten, Paul Shannon, and Martin Morgan. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics (Oxford, England)*, 30(14):2076–8, Jul 2014.

[367] Damian Smedley, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, Mohammad Hamza Awedh, Richard Baldock, Giulia Barbiera, Philippe Bardou, Tim Beck, Andrew Blake, Merideth Bonierbale, Anthony J Brookes, Gabriele Bucci, Iwan Buetti, Sarah Burge, Cédric Cabau, Joseph W Carlson, Claude Chelala, Charalambos Chrysostomou, Davide Cittaro, Olivier Collin, Raul Cordova, Rosalind J Cutts, Erik Dassi, Alex Di Genova, Anis Djari, Anthony Esposito, Heather Estrella, Eduardo Eyras, Julio Fernandez-Banet, Simon Forbes, Robert C Free, Takatomo Fujisawa, Emanuela Gadaleta, Jose M Garcia-Manteiga, David Goodstein, Kristian Gray, José Afonso Guerra-Assunção, Bernard Haggarty, Dong-Jin Han, Byung Woo Han, Todd Harris, Jayson Harshbarger, Robert K Hastings, Richard D Hayes, Claire Hoede, Shen Hu, Zhi-Liang Hu, Lucie Hutchins, Zhengyan Kan, Hideya Kawaji, Aminah Keliet, Arnaud Kerhornou, Sunghoon Kim, Rhoda Kinsella, Christophe Klopp, Lei Kong, Daniel Lawson, Dejan Lazarevic, Ji-Hyun Lee, Thomas Letellier, Chuan-Yun Li, Pietro Lio, Chu-Jun Liu, Jie Luo, Alejandro Maass, Jerome Mariette, Thomas Maurel, Stefania Merella, Azza Mostafa Mohamed, Francois Moreews, Ibounyamine Nabihoudine, Nelson Ndegwa, Céline Noirot, Cristian Perez-Llamas, Michael Primig, Alessandro Quattrone, Hadi Quesneville, Davide Rambaldi, James Reecy, Michela Riba, Steven Rosanoff, Amna Ali Saddiq, Elisa Salas, Olivier Sallou, Rebecca Shepherd, Reinhard Simon, Linda Sperling, William Spooner, Daniel M Staines, Delphine Steinbach, Kevin Stone, Elia Stupka, Jon W Teague, Abu Z Dayem Ullah, Jun Wang, Doreen Ware, Marie Wong-Erasmus, Ken Youens-Clark, Amonida Zadissa, Shi-Jian Zhang, and Arek Kasprzyk. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic acids research*, 43(W1):W589–98, Jul 2015.

[368] T Kawabata, M Ota, and K Nishikawa. The Protein Mutant Database. *Nucleic acids research*, 27(1):355–7, Jan 1999.

[369] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, Dianna G Fisk, Jodi E Hirschman, Benjamin C Hitz, Kalpana Karra, Cynthia J Krieger, Stuart R Miyasato, Rob S Nash, Julie Park, Marek S Skrzypek, Matt Simison, Shuai

Weng, and Edith D Wong. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research*, 40(Database issue):D700–5, Jan 2012.

[370] Gianni Liti, David M Carter, Alan M Moses, Jonas Warringer, Leopold Parts, Stephen A James, Robert P Davey, Ian N Roberts, Austin Burt, Vassiliki Koufopanou, Isheng J Tsai, Casey M Bergman, Douda Bensasson, Michael J T O'Kelly, Alexander van Oudenaarden, David B H Barton, Elizabeth Bailes, Alex N Nguyen, Matthew Jones, Michael A Quail, Ian Goodhead, Sarah Sims, Frances Smith, Anders Blomberg, Richard Durbin, and Edward J Louis. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–41, Mar 2009.

[371] G Matassi, P M Sharp, and C Gautier. Chromosomal location effects on gene sequence evolution in mammals. *Current biology : CB*, 9(15):786–91, 1999.

[372] Martin J Lercher and Laurence D Hurst. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in genetics : TIG*, 18(7):337–40, Jul 2002.

[373] Leonid Teytelman, Michael B Eisen, and Jasper Rine. Silent but not static: accelerated base-pair substitution in silenced chromatin of budding yeasts. *PLoS genetics*, 4(11):e1000247, Nov 2008.

[374] Guri Giaever, Angela M Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Véronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno André, Adam P Arkin, Anna Astromoff, Mohamed El-Bakkoury, Rhonda Bangham, Rocio Benito, Sophie Brachat, Stefano Campanaro, Matt Curtiss, Karen Davis, Adam Deutschbauer, Karl-Dieter Entian, Patrick Flaherty, Francoise Foury, David J Garfinkel, Mark Gerstein, Deanna Gotte, Ulrich Güldener, Johannes H Hegemann, Svenja Hempel, Zelek Herman, Daniel F Jaramillo, Diane E Kelly, Steven L Kelly, Peter Kötter, Darlene LaBonte, David C Lamb, Ning Lan, Hong Liang, Hong Liao, Lucy Liu, Chuanyun Luo, Marc Lussier, Rong Mao, Patrice Menard, Siew Loon Ooi, Jose L Revuelta, Christopher J Roberts, Matthias Rose, Petra Ross-Macdonald, Bart Scherens, Greg Schimmack, Brenda Shafer, Daniel D Shoemaker, Sharon Sookhai-Mahadeo, Reginald K Storms, Jeffrey N Strathern, Giorgio Valle, Marleen Voet, Guido Volckaert, Ching-yun Wang, Teresa R Ward, Julie Wilhelmy, Elizabeth A Winzeler, Yonghong Yang, Grace Yen, Elaine Youngman, Kexin Yu, Howard Bussey, Jef D Boeke, Michael Snyder, Peter Philippsen, Ronald W Davis, and Mark Johnston. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418(6896):387–91, Jul 2002.

[375] Warren L DeLano. The pymol molecular graphics system. *http://pymol. org*, 2002.

[376] Sameer Velankar, José M Dana, Julius Jacobsen, Glen van Ginkel, Paul J Gane, Jie Luo, Thomas J Oldfield, Claire O'Donovan, Maria-Jesus Martin, and Gerard J Kleywegt. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic acids research*, 41(Database issue):D483–9, Jan 2013.

[377] Ursula Pieper, Narayanan Eswar, Ben M Webb, David Eramian, Libusha Kelly, David T Barkan, Hannah Carter, Parminder Mankoo, Rachel Karchin, Marc A Marti-Renom, Fred P Davis, and Andrej Sali. MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research*, 37(Database issue):D347–54, Jan 2009.

[378] Simon J Hubbard and Janet M Thornton. Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, 2(1), 1993.

[379] Simon Mitternacht. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*, 5:189, 2016.

[380] Ivan Sadowski, Bobby-Joe Breitkreutz, Chris Stark, Ting-Cheng Su, Matthew Dahabieh, Sheetal Raithatha, Wendy Bernhard, Rose Oughtred, Kara Dolinski, Kris Barreto, and Mike Tyers. The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update. *Database : the journal of biological databases and curation*, 2013:bat026, 2013.

[381] Holger Dinkel, Kim Van Roey, Sushama Michael, Manjeet Kumar, Bora Uyar, Brigitte Altenberg, Vladislava Milchevskaya, Melanie Schneider, Helen Kühn, Annika Behrendt, Sophie Luise Dahl, Victoria Damerell, Sandra Diebel, Sara Kalman, Steffen Klein, Arne C Knudsen, Christina Mäder, Sabina Merrill, Angelina Staudt, Vera Thiel, Lukas Welti, Norman E Davey, Francesca Diella, and Toby J Gibson. ELM 2016–data update and new functionality of the eukaryotic linear motif resource. *Nucleic acids research*, 44(D1):D294–300, Jan 2016.

[382] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, and Cathy H Wu. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)*, 31(6):926–32, Mar 2015.

[383] Emanuel Gonçalves, Zrinka Raguz Nakic, Mattia Zampieri, Omar Wagih, David Ochoa, Uwe Sauer, Pedro Beltrao, and Julio Saez-Rodriguez. Systematic Analysis of Transcriptional and Post-transcriptional Regulation of Metabolism in Yeast. *PLoS computational biology*, 13(1):e1005297, Jan 2017.

[384] Gordon Chua, Quaid D Morris, Richelle Sopko, Mark D Robinson, Owen Ryan, Esther T Chan, Brendan J Frey, Brenda J Andrews, Charles Boone, and Timothy R Hughes. Identifying transcription factor functions and targets by phenotypic activation. *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):12045–50, Aug 2006.

[385] Zhanzhi Hu, Patrick J Killion, and Vishwanath R Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nature genetics*, 39(5):683–7, May 2007.

[386] Patrick Kemmeren, Katrin Sameith, Loes A L van de Pasch, Joris J Benschop, Tineke L Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W Ko, Sebastiaan van Heesch, Mehdi M Kashani, Giannis Ampatziadis-Michailidis, Mariel O Brok, Nathalie A C H Brabers, Anthony J Miles, Diane Bouwmeester, Sander R van Hooff, Harm van Bakel, Erik Sluiters, Linda V Bakker, Berend Snel, Philip Lijnzaad, Dik van Leenen, Marian J A Groot Koerkamp, and Frank C P Holstege. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–52, Apr 2014.

[387] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, Ezra G Jennings, Julia Zeitlinger, Dmitry K Pokholok, Manolis

Kellis, P Alex Rolfe, Ken T Takusagawa, Eric S Lander, David K Gifford, Ernest Fraenkel, and Richard A Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sep 2004.

[388] Ho Sung Rhee and B Franklin Pugh. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–19, Dec 2011.

[389] Christine Tachibana, Jane Y Yoo, Jean-Basco Tagne, Nataly Kacherovsky, Tong I Lee, and Elton T Young. Combined global localization analysis and transcriptome data identify genes that are directly coregulated by Adr1 and Cat8. *Molecular and cellular biology*, 25(6):2138–46, Mar 2005.

[390] Bryan J Venters, Shinichiro Wachi, Travis N Mavrich, Barbara E Andersen, Peony Jena, Andrew J Sinnamon, Priyanka Jain, Noah S Rolleri, Cizhong Jiang, Christine Hemeryck-Walsh, and B Franklin Pugh. A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces. *Molecular cell*, 41(4):480–92, Feb 2011.

[391] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia Pannier, Jaime Abraham Castro-Mondragón, Alejandra Medina-Rivera, Hilda Solano-Lira, César Bonavides-Martínez, Ernesto Pérez-Rueda, Shirley Alquicira-Hernández, Liliana Porrón-Sotelo, Alejandra López-Fuentes, Anastasia Hernández-Koutoucheva, Víctor Del Moral-Chávez, Fabio Rinaldi, and Julio Collado-Vides. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic acids research*, 44(D1):D133–43, Jan 2016.

[392] Naomi R Wray, Jian Yang, Ben J Hayes, Alkes L Price, Michael E Goddard, and Peter M Visscher. Pitfalls of predicting complex traits from SNPs. *Nature reviews. Genetics*, 14(7):507–15, 07 2013.

[393] Yali Xue, Yuan Chen, Qasim Ayub, Ni Huang, Edward V Ball, Matthew Mort, Andrew D Phillips, Katy Shaw, Peter D Stenson, David N Cooper, and Chris Tyler-Smith. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *American journal of human genetics*, 91(6):1022–32, Dec 2012.

[394] Lizhi Ian Gong, Marc A Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife*, 2:e00631, May 2013.

[395] D G MacArthur, T A Manolio, D P Dimmock, H L Rehm, J Shendure, G R Abecasis, D R Adams, R B Altman, S E Antonarakis, E A Ashley, J C Barrett, L G Biesecker, D F Conrad, G M Cooper, N J Cox, M J Daly, M B Gerstein, D B Goldstein, J N Hirschhorn, S M Leal, L A Pennacchio, J A Stamatoyannopoulos, S R Sunyaev, D Valle, B F Voight, W Winckler, and C Gunter. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–76, Apr 2014.

[396] Nadia A Chuzhanova, Emmanuel J Anassis, Edward V Ball, Michael Krawczak, and David N Cooper. Meta-analysis of indels causing human genetic disease: mechanisms

of mutagenesis and the role of local DNA sequence complexity. *Human mutation*, 21(1):28–44, Jan 2003.

[397] Rameen Beroukhim, Craig H Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S Boehm, Jennifer Dobson, Mitsuyoshi Urashima, Kevin T Mc Henry, Reid M Pinchback, Azra H Ligon, Yoon-Jae Cho, Leila Haery, Heidi Greulich, Michael Reich, Wendy Winckler, Michael S Lawrence, Barbara A Weir, Kumiko E Tanaka, Derek Y Chiang, Adam J Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld, Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic J Kaye, Hidefumi Sasaki, Joel E Tepper, Jonathan A Fletcher, Josep Tabernero, José Baselga, Ming-Sound Tsao, Francesca Demichelis, Mark A Rubin, Pasi A Janne, Mark J Daly, Carmelo Nucera, Ross L Levine, Benjamin L Ebert, Stacey Gabriel, Anil K Rustgi, Cristina R Antonescu, Marc Ladanyi, Anthony Letai, Levi A Garraway, Massimo Loda, David G Beer, Lawrence D True, Aikou Okamoto, Scott L Pomeroy, Samuel Singer, Todd R Golub, Eric S Lander, Gad Getz, William R Sellers, and Matthew Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, Feb 2010.

[398] Denghong Chen, Zhenhua Zhang, and Yuxiu Meng. Systematic Tracking of Disrupted Modules Identifies Altered Pathways Associated with Congenital Heart Defects in Down Syndrome. *Medical science monitor : international medical journal of experimental and clinical research*, 21:3334–42, Nov 2015.

[399] Fu-Jou Lai, Chia-Chun Chiu, Tzu-Hsien Yang, Yueh-Min Huang, and Wei-Sheng Wu. Identifying functional transcription factor binding sites in yeast by considering their positional preference in the promoters. *PloS one*, 8(12):e83791, 2013.

[400] Francesco Vallania, Davide Schiavone, Sarah Dewilde, Emanuela Pupo, Serge Garbay, Raffaele Calogero, Marco Pontoglio, Paolo Provero, and Valeria Poli. Genome-wide discovery of functional transcription factor binding sites by comparative genomics: the case of Stat3. *Proceedings of the National Academy of Sciences of the United States of America*, 106(13):5117–22, Mar 2009.

[401] Elizabeth T Cirulli and David B Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews. Genetics*, 11(6):415–25, Jun 2010.

[402] The International HapMap Project. *Nature*, 426(6968):789–96, Dec 2003.

[403] Michael J Wagner. Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. *Pharmacogenomics*, 14(4):413–24, Mar 2013.

[404] Christopher DeBoever, Yosuke Tanigawa, Greg McInnes, Adam Lavertu, Chris Chang, Carlos D Bustamante, Mark J Daly, and Manuel A Rivas. Medical relevance of protein-truncating variants across 337,208 individuals in the uk biobank study. *bioRxiv*, page 179762, 2017.

[405] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani

de Silva, Holger Heyn, Xianming Deng, Regina K Egan, Qingsong Liu, Tatiana Mironenko, Xeni Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S Gray, Daniel A Haber, Michael R Stratton, Cyril H Benes, Lodewyk F A Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J Garnett. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754, Jul 2016.

[406] Loes M Olde Loohuis, Jacob A S Vorstman, Anil P Ori, Kim A Staats, Tina Wang, Alexander L Richards, Ganna Leonenko, James T Walters, Joseph DeYoung, Rita M Cantor, and Roel A Ophoff. Genome-wide burden of deleterious coding variants increased in schizophrenia. *Nature communications*, 6:7501, Jul 2015.

[407] George Kritikos, Manuel Banzhaf, Lucia Herrera-Dominguez, Alexandra Koumoutsi, Morgane Wartel, Matylda Zietek, and Athanasios Typas. A tool named Iris for versatile high-throughput phenotyping in microorganisms. *Nature microbiology*, 2:17014, Feb 2017.

[408] Raphael Guerois, Jens Erik Nielsen, and Luis Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2):369–87, Jul 2002.

[409] Maureen E Hillenmeyer, Eula Fung, Jan Wildenhain, Sarah E Pierce, Shawn Hoon, William Lee, Michael Proctor, Robert P St Onge, Mike Tyers, Daphne Koller, Russ B Altman, Ronald W Davis, Corey Nislow, and Guri Giaever. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science (New York, N.Y.)*, 320(5874):362–5, Apr 2008.

[410] Leonor Miller-Fleming, Pedro Antas, Teresa Faria Pais, Joshua L Smalley, Flaviano Giorgini, and Tiago Fleming Outeiro. Yeast DJ-1 superfamily members are required for diauxic-shift reprogramming and cell survival in stationary phase. *Proceedings of the National Academy of Sciences of the United States of America*, 111(19):7012–7, May 2014.

[411] Stefanie Jarolim, Anita Ayer, Bethany Pillay, Allison C Gee, Alex Phrakaysone, Gabriel G Perrone, Michael Breitenbach, and Ian W Dawes. Saccharomyces cerevisiae genes involved in survival of heat shock. *G3 (Bethesda, Md.)*, 3(12):2321–33, Dec 2013.

[412] Himanshu Sinha, Lior David, Renata C Pascon, Sandra Clauder-Münster, Sujatha Krishnakumar, Michelle Nguyen, Getao Shi, Jed Dean, Ronald W Davis, Peter J Oefner, John H McCusker, and Lars M Steinmetz. Sequential elimination of major-effect contributors identifies additional quantitative trait loci conditioning high-temperature growth in yeast. *Genetics*, 180(3):1661–70, Nov 2008.

[413] F Osman and S McCready. Differential effects of caffeine on DNA damage and replication cell cycle checkpoints in the fission yeast Schizosaccharomyces pombe. *Molecular & general genetics : MGG*, 260(4):319–34, Nov 1998.

[414] Michael Tsabar, Vinay V Eapen, Jennifer M Mason, Gonen Memisoglu, David P Waterman, Marcus J Long, Douglas K Bishop, and James E Haber. Caffeine impairs

resection during DNA break repair by reducing the levels of nucleases Sae2 and Dna2. *Nucleic acids research*, 43(14):6889–901, Aug 2015.

[415] E Balzi, M Wang, S Leterme, L Van Dyck, and A Goffeau. PDR5, a novel yeast multidrug resistance conferring transporter controlled by the transcription regulator PDR1. *The Journal of biological chemistry*, 269(3):2206–14, Jan 1994.

[416] Jia Li, Michael Biss, Yu Fu, Xin Xu, Stanley A Moore, and Wei Xiao. Two duplicated genes DDI2 and DDI3 in budding yeast encode a cyanamide hydratase and are induced by cyanamide. *The Journal of biological chemistry*, 290(20):12664–75, May 2015.

[417] A Kaetsu, T Fukushima, S Inoue, H Lim, and M Moriyama. Role of heat shock protein 60 (HSP60) on paraquat intoxication. *Journal of applied toxicology : JAT*, 21(5):425–30, 2001.

[418] Arvind Kumar Shukla, Prakash Pragya, Hitesh Singh Chaouhan, Anand Krishna Tiwari, Devendra Kumar Patel, Malik Zainul Abdin, and Debapratim Kar Chowdhuri. Heat shock protein-70 (Hsp-70) suppresses paraquat-induced neurodegeneration by inhibiting JNK and caspase-3 activation in Drosophila model of Parkinson's disease. *PloS one*, 9(6):e98886, 2014.

[419] Shuye Pu, Jessica Wong, Brian Turner, Emerson Cho, and Shoshana J Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic acids research*, 37(3):825–31, Feb 2009.

[420] Stephen J Royle. The role of clathrin in mitotic spindle organisation. *Journal of cell science*, 125(Pt 1):19–28, Jan 2012.

[421] Sean R Collins, Maya Schuldiner, Nevan J Krogan, and Jonathan S Weissman. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome biology*, 7(7):R63, 2006.

[422] J M Cherry, C Adler, C Ball, S A Chervitz, S S Dwight, E T Hester, Y Jia, G Juvik, T Roe, M Schroeder, S Weng, and D Botstein. SGD: Saccharomyces Genome Database. *Nucleic acids research*, 26(1):73–9, Jan 1998.

[423] Andrey A Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics (Oxford, England)*, 28(10):1353–8, May 2012.

[424] Nobuhiko Tokuriki and Dan S Tawfik. Stability effects of mutations and protein evolvability. *Current opinion in structural biology*, 19(5):596–604, Oct 2009.

[425] David L Masica and Rachel Karchin. Towards Increasing the Clinical Relevance of In Silico Methods to Predict Pathogenic Missense Variants. *PLoS computational biology*, 12(5):e1004725, 05 2016.

[426] H Duzkale, J Shen, H McLaughlin, A Alfares, M A Kelly, T J Pugh, B H Funke, H L Rehm, and M S Lebo. A systematic approach to assessing the clinical significance of genetic variants. *Clinical genetics*, 84(5):453–63, Nov 2013.

[427] Simon G Coetzee, Gerhard A Coetzee, and Dennis J Hazelett. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics (Oxford, England)*, 31(23):3847–9, Dec 2015.

[428] Malin C Andersen, Pär G Engström, Stuart Lithwick, David Arenillas, Per Eriksson, Boris Lenhard, Wyeth W Wasserman, and Jacob Odeberg. In silico detection of sequence variations modifying transcriptional regulation. *PLoS computational biology*, 4(1):e5, Jan 2008.

[429] Lucas D Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, 30(11):1095–106, Nov 2012.

[430] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-variant association analysis: study designs and statistical tests. *American journal of human genetics*, 95(1):5–23, Jul 2014.

[431] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics (Oxford, England)*, 33(21):3387–3395, Nov 2017.

[432] Michael K K Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics (Oxford, England)*, 30(12):i121–9, Jun 2014.

[433] Duolin Wang, Shuai Zeng, Chunhui Xu, Wangren Qiu, Yanchun Liang, Trupti Joshi, and Dong Xu. MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics (Oxford, England)*, Aug 2017.

[434] Hana Lango, Colin N A Palmer, Andrew D Morris, Eleftheria Zeggini, Andrew T Hattersley, Mark I McCarthy, Timothy M Frayling, and Michael N Weedon. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes*, 57(11):3129–35, Nov 2008.

[435] Mingming Liu, Layne T Watson, and Liqing Zhang. Predicting the combined effect of multiple genetic variants. *Human genomics*, 9:18, Jul 2015.

[436] Amy Leung, Dustin E Schones, and Rama Natarajan. Using epigenetic mechanisms to understand the impact of common disease causing alleles. *Current opinion in immunology*, 24(5):558–63, Oct 2012.