Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Systems Biology

Extracting complementary insights from molecular phenotypes for prioritization of disease-associated mutations

Shayne D. Wierbowski[1,2], Robert Fragoza[2,3], Siqi Liang[1,2] and Haiyuan Yu[1,2]

Abstract

Rapid advances in next-generation sequencing technology have resulted in an explosion of whole-exome/genome sequencing data, providing an unprecedented opportunity to identify disease- and trait-associated variants in humans on a large scale. To date, the long-standing paradigm has leveraged fitness-based approximations to translate this ever-expanding sequencing data into causal insights in disease. However, while this approach robustly identifies variants under evolutionary constraint, it fails to provide molecular insights. Moreover, complex disease phenomena often violate standard assumptions of a direct organismal phenotype to overall fitness effect relationship. Here we discuss the potential of a molecular phenotype-oriented paradigm to uniquely identify candidate disease-causing mutations from the human genetic background. By providing a direct connection between single nucleotide mutations and observable organismal and cellular phenotypes associated with disease, we suggest that molecular phenotypes can readily incorporate alongside established fitness-based methodologies to provide complementary insights to the functional impact of human mutations. Lastly, we discuss how integrated approaches between molecular phenotypes and fitness-based perspectives facilitate new insights into the molecular mechanisms underlying disease-associated mutations while also providing a platform for improved interpretation of epistasis in human disease.

Addresses

[1] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA
[2] Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA
[3] Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA

Corresponding author: Yu, Haiyuan (haiyuan.yu@cornell.edu)

Current Opinion in Systems Biology 2018, **11**:107–116

This review comes from a themed issue on **Big data acquisition and analysis**

Edited by **Mark Gerstein** and **Haiyuan Yu**

For a complete overview see the Issue and the Editorial

Available online 17 September 2018

https://doi.org/10.1016/j.coisb.2018.09.006

2452-3100/© 2018 Elsevier Ltd. All rights reserved.

Introduction

Ever-improving next-generation sequencing technologies have led to the ongoing discovery of tens of millions of DNA variants across diverse human populations [1] and have enabled the identification of tens of thousands of disease-associated mutations [2,3]. Nonetheless, a vast majority of these variants remain uncharacterized and a corresponding understanding of how these unannotated variants may contribute to human disease and traits has yet to materialize [4]. Although numerous mutations occur in noncoding regions of genomes, missense variants are of particular interest to researchers since known disease- and trait-associated mutations have been shown to be enriched in coding regions [5]. Proper interpretation of the functional impact of missense mutations, which dominate exome sequencing datasets, remains a pivotal challenge. Overcoming this challenge will require new tools and approaches that better leverage large-scale sequencing data and that take advantage of newly emerging sources of experimentally assessed functional variant data.

Functional prediction algorithms have provided a boon towards the identification and prioritization of disease-associated mutations. Although early approaches to disease association specifically prioritized rare variants, tools such as SIFT [6–8], PolyPhen-2 [8,9], CADD [10], and PROVEAN [11–13] have provided systematic methods for predicting the impact of missense variants. Other tools, such as GWAVA [14] and LinSIGHT [15], tailor their methodology specifically to non-coding variants. These approaches share a central approach that utilizes principles of population genetics and conservation both within humans and across species as a means of approximating the fitness cost of specific variants. Cumulatively, these methods have been widely used in prior identification of disease-associated mutations [16–21]. However, while these methods continue to persist as invaluable tools for prioritizing coding and non-coding mutations in disease, annotations from these tools alone do not provide insight into the underlying molecular mechanisms of causal variants. Indeed, no method to-date can effectively identify true risk missense variants for human disease [22,23].

A guiding principle of precision medicine is to accurately measure clinical and molecular attributes of individual patients so as to tailor personalized therapies based on the outcomes of these measurements [24]. Considering millions of DNA variants segregating in human genomes, and the extraordinary level of allelic heterogeneity found in disease, success of the precision medicine effort hinges not only on the ability to detect disease-causing mutations, but also to understand and properly assess the functional consequences of these mutations. A major challenge, therefore, is to radically accelerate the pace of experimental and computational assessments of the functional impacts of millions of single nucleotide variants (SNVs) uncovered by sequencing efforts. Direct assessments of molecular phenotypes—such as impact on protein stability, enzymatic kinetics, or binding efficiencies by missense mutations or gene regulatory impacts by non-coding mutations—provide a unique and complementary perspective to current methods for detecting causal disease mutations. Integrating molecular phenotype data into fitness-based approaches for identifying deleterious mutations may also provide new insights into how causal mutations mechanistically function and provides a framework for dissecting epistatic relationships that modulate the impact of low penetrance mutations.

Caveats to fitness-based methods

Long-standing computational methods rooted in approximating fitness effects have provided considerable headway towards the identification of disease-causing mutations on genome-wide scales. However, carving out the path for future innovation in variant prioritization—and moreover mechanistic interpretation—necessitates an awareness of the limitations and caveats surrounding the current methods. Indeed, despite their widespread use, current algorithms often perform poorly in clinical settings and seldom result in measurable phenotypes. For example, Miosge and colleagues examined 33 *de novo* missense mutations occurring in essential immune system genes in mice found that only 20% of mutations predicted to be deleterious by PolyPhen-2 resulted in discernible phenotypes in mice homozygous for the *de novo* mutations tested [25]. A more recent study expanded the scope of this genotype-phenotype by inducing 116,330 random ENU mutations in mice. Their results showed that among missence mutations scored as "probably damaging" by PolyPhen-2, only 17% resulted in discernible phenotypes in mice homozygous for the tested mutation [26]. Similar limitations for variant annotation algorithms were reported for a set of 236 clinically-relevant BRCA1/2 mutations [27]. Implicit biases in the training sets used to develop variant annotation algorithms [28,29], including limited sensitivity to disease-associated common variation [30], as well as high false positive rates across classifiers [25–27,31] may contribute to the limited accuracy of these methods to predict organismal phenotypes. Moreover, variant annotation algorithms provide little to no mechanistic insight as to how a predicted deleterious variant may function. This information is critical for developing targeted hypotheses and clinical strategies to target causal mutations.

Variant annotation algorithms have limited sensitivity to disease-associated common variants

Variant annotation algorithms vary greatly in their applications as do the methodologies that drive their predictions. Briefly, algorithms specific to coding variation, including PolyPhen-2 [8,9] and Mutation Taster [102], use various protein structure- and nucleotide-based databases to generate multiple sequence alignments for evaluating conservation of examined coding sites. Ultimately though, the breadth of disease-associated mutations represented in their training sets largely determines whether a variant annotation algorithm classifies a mutation as deleterious or not [32]. Biases and errors in these training sets can therefore limit the sensitivity of these tools to accurately detect deleterious variants [28], as can limited sensitivity for variants involved in complex, non-Mendelian disease [33]. In general, the lower the allele frequency of a variant, the more likely a variant annotation algorithm is to score it as deleterious [29]. As a result, variant annotation algorithms also underperform in detecting disease- and risk-associated mutations that occur at common allele frequencies [30,33].

Given the conceptual framework of identifying causal variants through fitness effects, and the historic emphasis of previous studies on highly penetrant, Mendelian diseases, underperformance detecting these deleterious common variants is logical. Though purifying selection should limit the capacity of truly deleterious variants to achieve common allele frequencies (MAF $>$ 1.0%), the probability of such variants reaching high allele frequencies is never zero; particularly if the variant affects a trait minimally associated with reproductive fitness. Indeed, several examples of clinically-relevant, disease-associated variants at common allele frequencies follow this pattern. For example, gene dosage effects from the apolipoprotein E type 4 allele (MAF = 18.4%) increase Alzheimer's disease risk by 20–90% [34–36]. Likewise, carriers of the P12A polymorphism of PPARG (MAF = 11.0%) are significantly more likely to develop type 2 diabetes [37,38]. Similar examples of common variants (MAF $>$ 1.0%) that result in or modulate disease risk are detailed in current literature [24,39–51] and briefly summarized in Table 1. Notably, only one of these listed disease-associated mutations scores as "probably damaging" by PolyPhen-2 while only a handful of cases are scored as "deleterious" by SIFT (Table 1). Moreover, functional mutations at common allele frequencies, including R543Q and C282Y mutations in F5 [52,53] and HFE [54–57] respectively, represent disease mutations with incomplete penetrance (Table 1). Despite strong evidence linking these mutations to disease risk [52–57], a majority of carriers of these variants do not develop their associated diseases [58]. While there is evidence suggesting that many of these mutations may be annotation errors or artifacts of association studies [59,60], partially penetrant disease-associated mutations, nonetheless, still modulate disease risk. The current framework for variant annotation is evidently ill-suited to discern variants associated with subtle effects. Yet characterizing precisely these mutations will be crucial toward understanding how an individual's genetic background determines their risk for particular diseases and influences complex traits.

High discordance between variant annotation algorithms

In practice, researchers incorporate multiple variant annotation algorithms to identify putatively functional mutations from whole-exome/genome sequencing data; however, discordance between the results of these algorithms is high. Indeed, a study that applied seven different variant annotation algorithms to data from the Exome Sequencing Project found that 47% of nonsynonymous variants were predicted to be functional by at least one algorithm while only 1% of nonsynonymous variants were scored as functional by all seven annotation tools [31]. Large discrepancies were also observed between variant annotation algorithms when applied to phenotype-associated mutations and were each suggested to greatly overestimate the damaging effect of their predicted functional mutations [26]. A "majority rule" criteria in which at least four of seven variant annotation algorithms must score the variant as functional for the variant to be considered deleterious can instead be applied [3,31], but false negative rates are presumably very high when combining the results from distinct variant annotation algorithms in this manner. The distinct datasets and annotation sources used to develop each of these variant annotation algorithms can be used instead to train a single support vector machine for predicting putatively functional alleles, as developed for CADD [10]. Nonetheless, despite impressive classification accuracy, CADD achieved only a 15% success rate when applied to the aforementioned set of 33 *de novo* missense mutations in essential immune system genes studied by Miosge and colleagues [25].

Variant annotation algorithms alone provide limited mechanistic insights

Mutations can perturb cellular activity in multiple ways. In particular, disease-associated missense mutations often function by disrupting protein–protein interactions [61–63], destabilizing protein folding [61,62], or altering transcription factor activity [64,65]. Understanding the molecular mechanisms through which disease-associated mutations function is imperative for developing clinical strategies to treat their corresponding phenotypes and for drug target assessment [66,67]. In spite of this importance, only a single widely used variant annotation algorithm for coding variants, MutPred2 [68], currently evaluates the possible mechanisms by which mutations scored as deleterious may function. More precise predictions for deleterious variants and better insights to their

Table 1

A curation of the literature highlights several disease-associated variants occurring at unexpectedly common minor allele frequencies (MAF > 1.0%). These variants exhibit lower selection pressure than may be anticipated given their well-studied connections to disease phenotype, exemplifying the confounding that occurs when using fitness driven perspectives to explain and detect disease mutations. Indeed, two common variant annotation algorithms, PolyPhen-2 and SIFT, have infrequently labeled these known functional mutations with their highest functional annotations.

Gene	Mutation	ExAC MAF	rsID	PolyPhen-2 score	SIFT score	Disease	Citation
APOE	C130R	18.40%	rs429358	Benign	Tolerated	Alzheimer's disease	[34–36]
ARMS2	A69S	25.50%	rs10490924	Possibly damaging	Deleterious (low confidence)	Age-related macular degeneration	[39,40]
BTD	D444H	3.20%	rs13078881	Benign	Deleterious	Partial biotinidase deficiency	[41,42]
CFH	Y402H	32.80%	rs1061170	Benign	Tolerated	Age-related macular degeneration	[43–45]
COL4A2	E1123G	1.70%	rs117412802	Possibly damaging	Unscored	Haemorrhagic stroke	[46]
F5	R543Q	2.20%	rs6025	Benign	Tolerated	Factor V Leiden	[52,53]
HFE	C282Y	3.20%	rs1800562	Probably damaging	Deleterious	Hemochromatosis	[54–57]
INHA	A257T	2.40%	rs12720062	Benign	Tolerated	Premature ovarian failure	[47–49]
PPARG	P12A	11.00%	rs1801282	Benign	Deleterious (low confidence)	Type 2 diabetes	[37,38]
PRSS1	A16V	1.60%	rs202003805	Benign	Tolerated	Chronic pancreatitis	[50,51]
TRIM22	R321K	3.00%	rs12364019	Possibly damaging	Deleterious	Inflammatory bowel disease	[24]
TRIM22	S244L	1.40%	rs61735273	Possibly damaging	Deleterious	Inflammatory bowel disease	[24]

corresponding molecular mechanisms may be achieved through improved structural databases to detail where missense mutations physically occur with respect to protein interface residues [69,70]. Similar database improvements may also apply to variant annotation algorithms that also score noncoding mutations, for example fitCons [71] which evaluates patterns of polymorphisms and genetic divergence to estimate the "fitness consequence" of point mutations genome-wide. However, fitCons, heavily depends on the accuracy of functional elements identified by ENCODE [72]. Recently developed sequence co-variation approaches to predicting the effects of DNA variants bypass dependence on structural feature or functional noncoding annotations [73]; however, mechanistic insights as to how these epistatic dependencies emerge are not provided. As such, integrating structural and functional information from these datasets can provide improved and complementary insights to the molecular function of predicted deleterious mutations.

Molecular phenotypes: an orthogonal framework

In assessing the impact of human variants, we highlight the importance of distinguishing three related yet distinct biological concepts: overall fitness, organismal/cellular phenotype, and molecular phenotype (Figure 1). Overall fitness refers to the ability of an individual to survive and reproduce. Organismal phenotypes refer to observable features, including disease phenotypes such as diabetes, autism spectrum disorder and cancer, or traits such as height, hair color and blood type. Molecular phenotypes refer to the direct effect of a variant at the molecular level. For example, changes in gene expression, loss of protein stability, changes in enzymatic activity, or modifications to protein—protein, protein-DNA or protein-ligand interaction affinities.

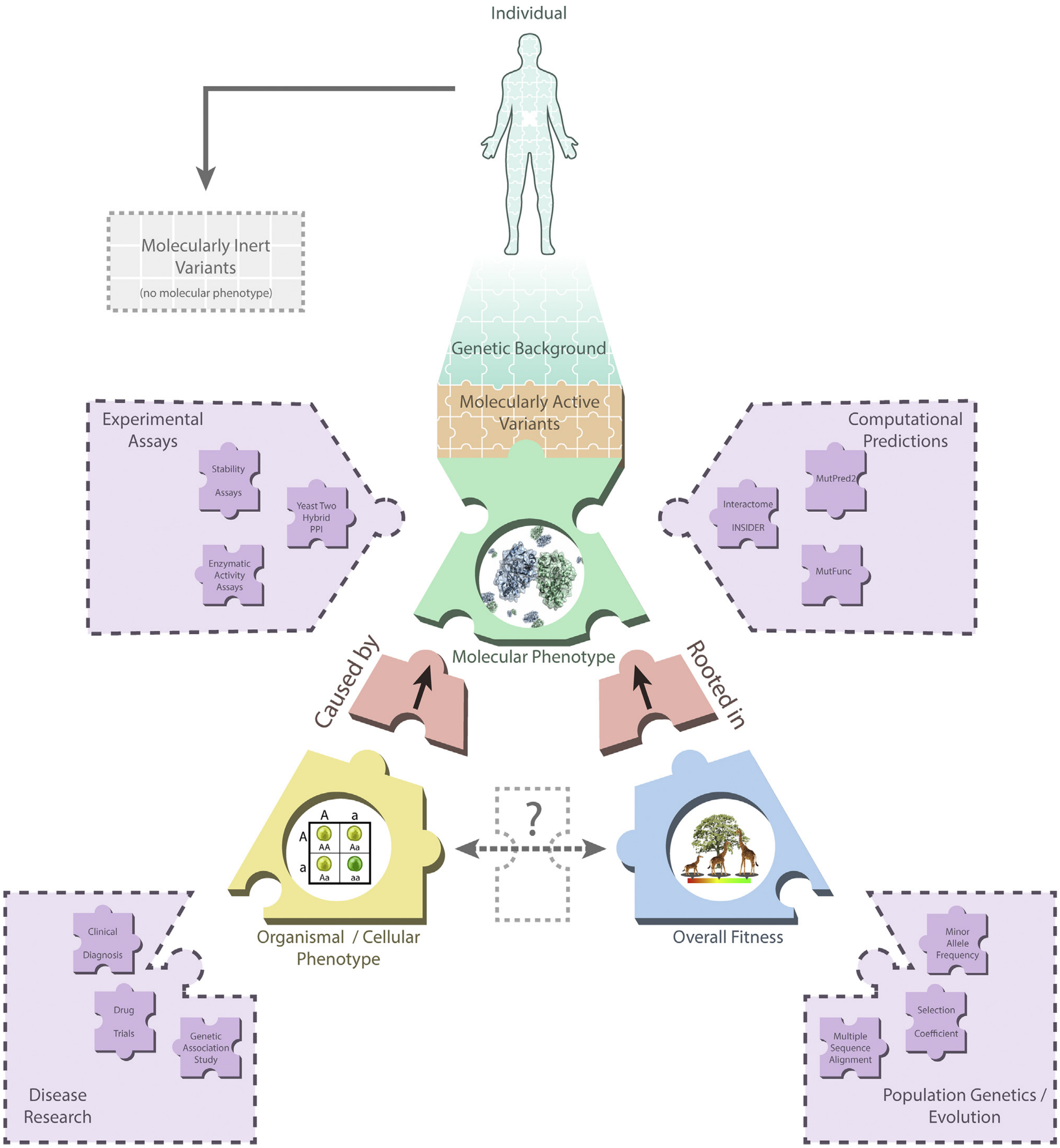
All human genetic variation separates into molecularly inert or molecularly active variants depending on whether or not each variant causes a molecular phenotype. While not all molecular phenotypes contribute directly to observable organismal phenotypes, organismal or cellular phenotypes are largely derived in molecularly active variants; and hence must be directly mediated through one or more molecular phenotypes. Likewise, overall fitness is always rooted in molecular phenotypes since molecular changes modulate the ability of the organism to perform various functions necessary for survival and reproduction. In principal, all organismal phenotypes associate with a fitness value ranging from deleterious, to neutral, to advantageous. While there is a direct relationship between organismal phenotypes and fitness, this relationship is not always clearly defined, particularly in specialized fields of disease research dealing with cancer biology, age or post-reproductive related diseases, and complex diseases with reduced penetrance [74]. In such disease studies, the one-to one correspondence between fitness score and the severity of the organismal phenotype breaks down since clinically deleterious phenotypes can have limited impact on reproduction. Molecular phenotypes can be indispensable towards characterizing these cases of ambiguous fitness-to-phenotype relationships.

Molecular phenotypes provide complementary information for identifying causal variants

Whereas most approaches leverage the link between fitness effects and organismal/cellular phenotypes, an alternative framework rooted in molecular phenotypes provides an orthogonal line of support. At least two degrees of separation lie between disease phenotypes caused by particular variants, the fitness effects of these variants, and our ability to discern these effects. By contrast methods aimed at molecular phenotypes directly address the central link. The combination of these two rationally justified, yet conceptually distinct paths connecting SNVs to disease phenotype is expected to culminate in an overall higher degree of accuracy in predicting disease associations. The availability of data and library of tools for assessing molecular phenotypes are currently leagues behind the equivalent datasets for fitness-based approaches. Therefore, it is likely that established conservation and fitness-based methods will remain a valuable step in prioritizing variants, while more direct support from the orthogonal molecular phenotype data should serve as strong confidence in the accuracy of these results.

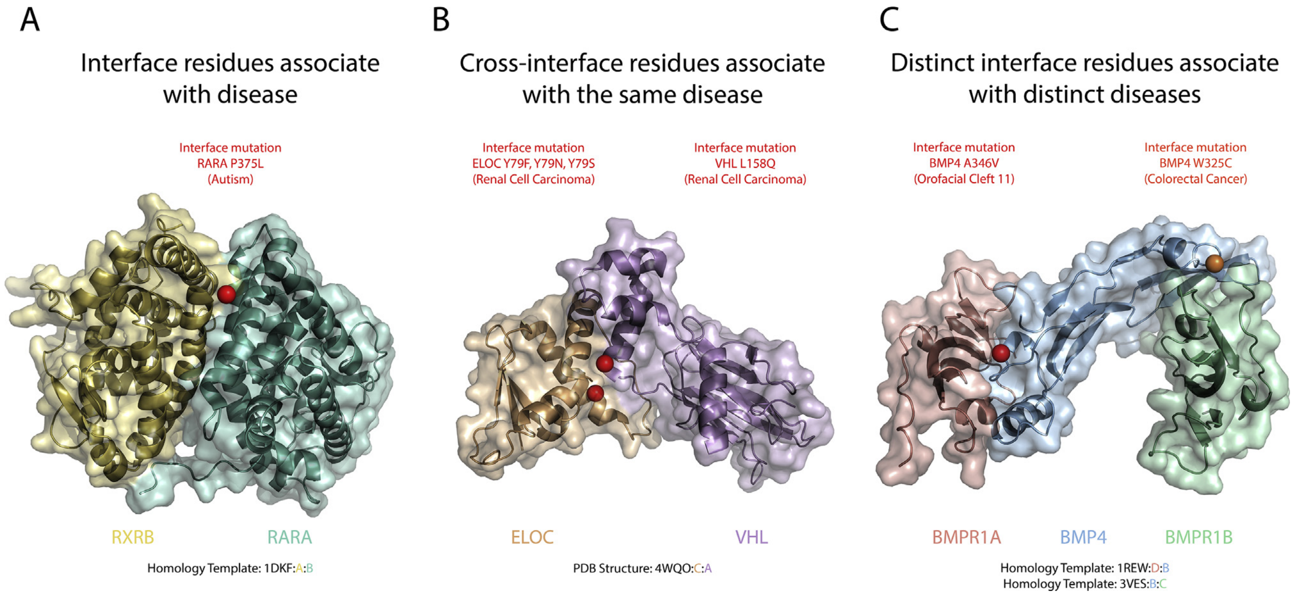
For instance, a recently developed interaction perturbation framework leveraged annotations of protein—protein interaction (PPI) interface residues [70] alongside PolyPhen-2 scores [75]. Chen and colleagues demonstrated increased accuracy in distinguishing *de novo* risk variants in autism spectrum disorder from benign mutations in unaffected siblings. Figure 2A provides a reconstructed example in which a proband PolyPhen-2 mutation scored as "probably damaging", P375L on the protein RARA, occurred on a predicted interface residue. In contrast, a second PolyPhen-2-scored "probably damaging" mutation, R83H on the same RARA protein, was reported in an unaffected individual; however, R83H did not occur on a predicted interaction interface residue. Consequently, despite matching PolyPhen-2 prediction, only the proband P375L mutation was predicted to disrupt the heterodimeric interaction between RARA and RXRB, a prediction which the authors also validated experimentally. This exemplifies the potential for molecular phenotypes to aid in pinpointing candidate causal variants that are otherwise indistinguishable from molecularly inert variants using fitness-based methods alone.

Figure 1



Graphical depiction of the relationship between three related biological concepts associated with human variations: 1) molecular phenotype, 2) organismal/cellular phenotype, and 3) overall fitness. All genetic variation is either molecularly inert or molecularly active. The cumulation of all molecularly active variants—each causing one or more molecular phenotypes—constitutes the unique genetic background of an individual. Molecular phenotypes provide the ultimate link explaining the mechanistic basis for how SNVs manifest in organismal/cellular phenotypes or come to be selected for or against through fitness effects. Although organismal phenotypes, in general, directly relate to overall fitness, weak effect diseases, late onset/post-reproductive diseases, and partially penetrant mutations often confound this relationship. Researchers have various tools to perform direct inquiries into how these three concepts relate to specific molecularly active variants. Human disease research aims to understand organismal/cellular phenotypes while population genetics provides insights into fitness, conservation, and selection. Researchers investigate molecular phenotypes either through direct experimental assays to observe underlying molecular phenotypes or through computational predictions of putative molecular phenotypes. The ultimate aim is to infer information about one point of the triangle through the other two; namely, scientists seek to infer which SNVs are causal disease variants though information about the overall fitness or molecular phenotype effects of the SNV.

Figure 2



Molecular phenotypes including the annotation of protein–protein interaction interface residues can inform the mechanism of disease-associated mutations. **A.** Homology model between RARA (template 1DKF:B) and RXRB (template 1DKF:A) used to distinguish a potentially causal mutation from a benign mutation. A *de novo* mutation, P375L, on RARA identified in an autism spectrum disorder-affected individual occurs on an interface residue with RXRB. RARA interface residue mutations were not found in an unaffected sibling. **B.** Homology model between VHL (PDB 4WQO:A) and ELOC (PDB 4WQO:C) demonstrates potential leveraging of molecular phenotypes to identify convergent mechanisms in divergent disease mutations. Variants on both of these proteins associate with the same disease and localize to the same interface. **C.** Homology model between BMP4 (template 1REW:B), BMPR1A (template 1REW:A), and BMPR1B (template 3VES:C) shows hypothesis-driven differentiation of mechanisms of different diseases based on molecular phenotype. Two variants on BMP4, A346V, and W325C, associated with divergent diseases localize to distinct interaction interfaces.

Leveraging molecular phenotype approaches towards disentangling molecular mechanisms of causal variants

The molecular phenotype framework provides clear potential to investigate the underlying mechanisms behind how variants manifest in disease phenotypes. Since the specific molecular defect associated with a variant often directly relates to the disease phenotype, identification of candidate variants based on molecular phenotype annotations should enable translational studies for disease etiology. The further development of methods to approximate and predict molecular phenotypes will facilitate the development of actional hypotheses to direct future research.

For instance, Chen et al. used experimentally derived and computationally predicted annotations of protein interaction interface residues [75] as a predictor for the molecular phenotype, loss of PPI. In addition to distinguishing a true autism risk variant, P375L, from other "probably damaging" variants, the additional knowledge that this variant intersected with the RARA-RXRB interaction interface (Figure 2A), led to the testable hypothesis that this variant would disrupt this interaction, and helped to propose a pathway for RARA's involvement in autism spectrum disorder through this interaction [75].

Extending the interface residue approximation for the loss of PPI molecular phenotype facilitates mechanistic inferences in other cases as well. This approach may be generalized to cases involving variants across both faces of an interface (Figure 2B). Corroborating cross-interface evidence may strengthen the hypothesis that disease-associated mutations function through disruption of a specific interaction and helps categorize distinct variants associated with the same disease by similarities in their molecular mechanisms. Figure 2B shows a known tumor suppressor gene-encoded protein, VHL [76,77] with a mutation, L158Q, associated with renal cell carcinoma, in complex with an elongation factor, ELOC. The localization of L158Q at the ELOC interface, suggests that the disease may function through disruption of the VHL-ELOC interaction. Moreover, ELOC contains several mutations on the same protein interaction interface, Y79F, Y79N, and Y79S, which are also associated with renal cell carcinoma, solidifying the hypothesis that these cross-interface variants drive a distinct form of renal cell carcinoma through a single shared molecular phenotype.

Understanding the molecular phenotypes caused by certain disease-associated mutations may further elucidate how several mutations on the same gene can associate with different diseases. For instance, two missense mutations found on the protein BMP4, A346V and W325C, are associated a developmental defect, orofacial cleft 11, and colorectal cancer respectively — two clinically distinct diseases. The homology models provided in Figure 2C demonstrate that these variants localize to opposites ends of the BMP4 structure and occur at distinct protein—protein interaction interfaces. These insights suggest these distinct disease phenotypes may manifest through divergent pathways related to the biological functions of their distinctly targeted interaction partners. Indeed, although BMPR1A and BMPR1B are paralogous, previous studies have linked them to unique functions and disease states [78,79].

Cumulatively, these interaction perturbation examples demonstrate how molecular phenotypes contribute to elucidation of disease etiology. We emphasize the potential to explore similar mechanistic hypotheses utilizing molecular phenotypes outside of PPI disruption. Recent studies have highlighted the value of examining other molecular phenotypes, including changes in protein stability [80,81] as well as changes in gene expression level [82,83], to unravel the pathogenic mechanisms of both coding and non-coding mutations.

Molecular phenotypes help dissect genetic epistasis and clear the path towards precision medicine

The combination of all molecularly active variants and their corresponding molecular phenotypes constitutes the genetic background that defines an individual (Figure 1). Frustratingly, some molecular phenotypes may never produce discernible organismal phenotypes, while others may do so only in the presence of specific, often unknown combinations of complementary molecular phenotypes. Indeed, recent studies in multiple organisms and human cell lines have identified complex pairwise, and even multi-way intertwinement by which deficits in individual genes affect organismal/cellular phenotypes and fitness [84—86]. The complex behavior of genetic epistasis has been a major roadblock to establishing causal relationships between genetic variants and human disease. However, there is no epistasis at the molecular level when examining molecular phenotypes of variants. Therefore, particularly compared to fitness effects which may be completely masked by epistasis, the ability to record or predict concrete molecular phenotypes associated with otherwise silent variants will prove crucial towards dissecting epistasis.

Molecular phenotype-based studies aimed at bridging this disconnect will carry immediate implications in precision medicine. On one front, leveraging molecular phenotype information to interpret the individual's genetic background is vital for deciphering variations among disease risk and drug response/toxicity among the human population. For example, Young et al. have elucidated how multiple SNVs on SORL1 affect BDNF-induced SORL1 expression in neuronal cells, contributing to risk for Alzheimer's disease [87]. More recently, Cheng-Hathaway et al. have uncovered the expression-reducing molecular phenotype of another variant, R47H on TREM2, that also increases risk of Alzheimer's disease [88]. Additionally, a study by Hauser et al. demonstrated that multiple variants on GPCR receptors impact drug response via a variety of molecular alterations, including reduced or increased onset kinetics and altered G-protein-binding specificity [89]. By providing a means to identify and evaluate functional effects at a molecular resolution, these studies help disentangle the links between human genetic variation and personalized disease risk assessment.

On another front, knowledge of molecular phenotypes of diseased tissue, especially in cancer, provides direct guidance on population-wide treatment for specialized types of disease. Tumor subtyping based on mRNA expression, protein expression, and epigenetic profiles [90—94] has already been widely used for making therapeutic decisions. A complementary effort in a recent study identified master regulators for metastatic progression of gastroenteropancreatic neuroendocrine tumors across four distinct subtypes, allowing prioritization of compounds based on patient-specific master regulator activity [95]. Harnessing molecular phenotypes that modulate both the genetic background and the disease state of an individual will significantly improve the efficacy of disease prevention, diagnosis, and treatment in a personalized manner.

Conclusion

The incorporation of direct assays for molecular phenotypes and novel computational methods that approximate molecular phenotypes in the continued efforts to identify, prioritize, and understand causal variants in human disease is positioned to provide a truly orthogonal view to the longstanding fitness-based approach. Whereas current variant annotation algorithms rooted in sequencing and fitness approximations have yielded suboptimal specificity, novel methods directed at molecular phenotypes aim to extract complementary molecular insights otherwise unavailable. Towards these ends, researchers have conducted high-throughput assays to directly measure the functional impact of thousands of disease-associated missense mutations on protein—protein interactions [61,62], protein stability [61], and DNA binding [64,65]. Literature curation efforts by the IMEx Consortium have provided protein interaction perturbation data corresponding to nearly 8,000 coding mutations in humans [96]. Continued development of high-throughput approaches—including deep-mutational scanning pipelines capable of probing nearly the entire mutational landscape of targeted proteins [97—100]—will provide an ever-larger resource of functional mutation data. This data will help elucidate the biochemical and evolutionary properties that differentiate truly damaging mutations from those that are benign.

Despite the impressive scale that high-throughput experimental pipelines have achieved [61,62,98], no experimental pipeline alone can keep pace with the rate of sequence variant discovery, highlighting the need for continued development of computational approaches and variant annotation algorithms. A comprehensive effort to integrate these sources of experimentally verified molecular phenotypes as labels to further train widely used fitness-based models will be key to improving their accuracy and clinical application, but remains as of yet unimplemented. Orthogonally, we also emphasize the continued need to develop novel algorithms distinct from the fitness paradigm that make direct predictions about putative molecular phenotypes. For instance, interaction interface residue annotations provide useful mechanistic insights, but low coverage in experimentally validated structures or homology models has limited their applicability. The recently published Interactome INSIDER resource provides a method to predict interface residues—and consequentially loss of PPI phenotypes—in the absence of structural information [70]. MutPred2 enables a combination of approaches, making predictions both for overall functional effect and prioritized potential mechanisms of action [68]. Recently, Wagih et al. have released MutFunc, a resource that aggregates and interprets several previouse datasets and algorithms to provide precomputed predictions for nearly every possible variant in *Homo sapiens*, *Saccharomyces cerevisiae*, and *Escherichia coli*. These predictions include estimates for changes to protein stability, protein interaction interfaces, post translational modifications, and transcription factor binding among other approximations for molecular phenotypes [101]. Advances in this realm of widespread predictors for specific molecular phenotypes that can prioritize targeted assays to validate the veracity of those phenotypes will prove crucial to ensure researchers can maintain up-to-date annotations of molecularly activate variants.

Conflict of interest statement

Nothing declared.

Acknowledgments

This work was supported by National Institute of General Medical Sciences grants (R01 GM104424, R01 GM124559, R01 GM125639); Eunice Kennedy Shriver National Institute of Child Health and Human Development grant (R01 HD082568); National Human Genome Research Institute grant (UM1 HG009393, R01 HG008126); National Science Foundation grant (DBI-1661380); Simons Foundation Autism Research Initiative grants (SF367561) to H.Y.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
* * of outstanding interest

1. * * Snyder M, Du J, Gerstein M: **Personal genome sequencing: current approaches and challenges**. *Genes Dev* 2010, **24**:423–431.

2. The 1000 Genomes Project Consortium: **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012, **491**:56–65.

3. * Fu W, *et al.*: **Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants**. *Nature* 2013, **493**:216–220.

4. * Stenson PD, *et al.*: **The human gene mutation database: 2008 update**. *Genome Med* 2009, **1**:13.

5. Hindorff LA, *et al.*: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *Proc Natl Acad Sci Unit States Am* 2009, **106**:9362–9367.

6. * Ng PC: **SIFT: predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**:3812–3814.

7. * Choi Y: *A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein*. ACM BCB; 2012.

8. * Adzhubei I, Jordan DM, Sunyaev SR: **Predicting functional effect of human missense mutations using PolyPhen-2**. *Curr Protoc Hum Genet* 2013. Chapter 7: p. Unit7 20.

9. Adzhubei IA, *et al.*: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**:248–249.

10. * Kircher M, *et al.*: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nat Genet* 2014, **46**:310–315.

11. Seifi M, Walter MA: **Accurate prediction of functional, structural, and stability changes in PITX2 mutations using in silico bioinformatics algorithms**. *PLoS One* 2018, **13**: e0195971.

12. * Choi Y, *et al.*: **Predicting the functional effect of amino acid substitutions and indels**. *PLoS One* 2012, **7**:e46688.

13. Choi Y, Chan AP: **PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels**. *Bioinformatics* 2015, **31**:2745–2747.

14. Ritchie GR, *et al.*: **Functional annotation of noncoding sequence variants**. *Nat Methods* 2014, **11**:294–296.

15. Huang YF, Gulko B, Siepel A: **Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data**. *Nat Genet* 2017, **49**:618–624.

16. Rosenberg S, *et al.*: **A recurrent point mutation in PRKCA is a hallmark of chordoid gliomas**. *Nat Commun* 2018, **9**:2371.

17. Graf S, *et al.*: **Identification of rare sequence variation underlying heritable pulmonary arterial hypertension**. *Nat Commun* 2018, **9**:1416.

18. Bhattacharya S, *et al.*: **Whole-genome sequencing of Atacama skeleton shows novel mutations linked with dysplasia**. *Genome Res* 2018, **28**:423–431.

19. Tubeleviciute-Aydin A, *et al.*: **Rare human Caspase-6-R65W and Caspase-6-G66R variants identify a novel regulatory region of Caspase-6 activity**. *Sci Rep* 2018, **8**:4428.

20. Bhatnager R, Dang AS: **Comprehensive in-silico prediction of damage associated SNPs in Human Prolidase gene**. *Sci Rep* 2018, **8**:9430.

21. Cunningham AD, *et al.*: **Coupling between protein stability and catalytic activity determines pathogenicity of G6PD variants**. *Cell Rep* 2017, **18**:2592–2599.

22. Iossifov I, *et al.*: **The contribution of de novo coding mutations to autism spectrum disorder**. *Nature* 2014, **515**:216–221.

23. Geisheker MR, *et al.*: **Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains**. *Nat Neurosci* 2017, **20**:1043–1051.

24. * Li Q, *et al.*: **Variants in TRIM22 that affect NOD2 signaling are associated with very-early-onset inflammatory bowel disease**. *Gastroenterology* 2016, **150**:1196–1207.

25. * Miosge LA, *et al.*: **Comparison of predicted and actual consequences of missense mutations**. *Proc Nat Acad Sci U S A* 2015, **112**:E5189–E5198.

26. * Wang T, *et al.*: **Probability of phenotypically detectable protein damage by ENU-induced mutations in the Mutagenetix database**. *Nat Commun* 2018, **9**:441.

27. * Ernst C, *et al.*: **Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics**. *BMC Med Genom* 2018, **11**:35.

28. Cooper GM, Shendure J: **Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data**. *Nat Rev Genet* 2011, **12**:628.

29. Henn BM, *et al.*: **Estimating mutation load in human genomes**. *Nat Rev Genet* 2015, **16**:333–343.

30. Ng PC, Henikoff S: **Predicting the effects of amino acid substitutions on protein function**. *Annu Rev Genom Hum Genet* 2006, **7**:61–80.

31. Tennessen JA, *et al.*: **Evolution and functional impact of rare coding variation from deep sequencing of human exomes**. *Science* 2012, **337**:64–69.

32. Care MA, *et al.*: **Deleterious SNP prediction: be mindful of your training data!** *Bioinformatics* 2007, **23**:664–672.

33. Thomas PD, Kejariwal A: **Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects**. *Proc Nat Acad Sci U S A* 2004, **101**:15398–15403.

34. * Corder E, *et al.*: **Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families**. *Science* 1993, **261**:921–923.

35. * Strittmatter WJ, *et al.*: **Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease**. *Proc Nat Acad Sci U S A* 1993, **90**:1977–1981.

36. * Deary IJ, *et al.*: **Cognitive change and the APOE ε4 allele**. *Nature* 2002, **418**:932.

37. Robitaille J, *et al.*: **The PPAR-gamma P12A polymorphism modulates the relationship between dietary fat intake and components of the metabolic syndrome: results from the Québec Family Study**. *Clin Genet* 2003, **63**:109–116.

38. Florez JC, *et al.*: **Effects of the type 2 diabetes-associated PPARG P12A polymorphism on progression to diabetes and response to troglitazone**. *J Clin Endocrinol Metab* 2007, **92**: 1502–1509.

39. * Kanda A, *et al.*: **A variant of mitochondrial protein LOC387715/ARMS2, not HTRA1, is strongly associated with age-related macular degeneration**. *Proc Natl Acad Sci Unit States Am* 2007, **104**:16227–16232.

40. * Rivera A, *et al.*: **Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk**. *Hum Mol Genet* 2005, **14**:3227–3236.

41. * Norrgard KJ, *et al.*: **Double mutation (A171T) and (D444H) is a common cause of profound biotinidase deficiency in children ascertained by newborn screening in the United States**. *Hum Mutat* 1998, **11**. 410–410.

42. * Borsatto T, *et al.*: **Biotinidase deficiency: clinical and genetic studies of 38 Brazilian patients**. *BMC Med Genet* 2014, **15**:96.

43. * Klein RJ, *et al.*: **Complement factor H polymorphism in age-related macular degeneration**. *Science* 2005, **308**:385–389.

44. * Edwards AO, *et al.*: **Complement factor H polymorphism and age-related macular degeneration**. *Science* 2005, **308**: 421–424.

45. * Haines JL, *et al.*: **Complement factor H variant increases the risk of age-related macular degeneration**. *Science* 2005, **308**: 419–421.

46. * Jeanne M, *et al.*: **COL4A2 mutations impair COL4A1 and COL4A2 secretion and cause hemorrhagic stroke**. *Am J Hum Genet* 2012, **90**:91–101.

47. * Chand AL, *et al.*: **Functional analysis of the human inhibin α subunit variant A257T and its potential role in premature ovarian failure**. *Hum Reprod* 2007, **22**:3241–3248.

48. * Chand AL, Harrison CA, Shelling AN: **Inhibin and premature ovarian failure**. *Hum Reprod Update* 2010, **16**:39–50.

49. * Shelling AN, *et al.*: **Inhibin: a candidate gene for premature ovarian failure**. *Hum Reprod* 2000, **15**:2644–2649.

50. * Witt H, Luck W, Becker M: **A signal peptide cleavage site mutation in the cationic trypsinogen gene is strongly associated with chronic pancreatitis**. *Gastroenterology* 1999, **117**: 7–10.

51. * Chen J-M, *et al.*: **The A16V signal peptide cleavage site mutation in the cationic trypsinogen gene and chronic pancreatitis**. *Gastroenterology* 1999, **117**:1508–1509.

52. * Kujovich JL: **Factor V leiden thrombophilia**. *Genet Med* 2010, **13**:1.

53. * van Mens TE, Levi M, Middeldorp S: **Evolution of factor V leiden**. *Thromb Haemostasis* 2013, **110**:23–30.

54. Beutler E: **The HFE Cys282Tyr mutation as a necessary but not sufficient cause of clinical hereditary hemochromatosis**. *Blood* 2003, **101**:3347–3350.

55. McCune CA, *et al.*: **Iron loading and morbidity among relatives of *HFE* C282Y homozygotes identified either by population genetic testing or presenting as patients**. *Gut* 2006, **55**: 554–562.

56. Whitlock EP, *et al.*: **Screening for hereditary hemochromatosis: a systematic review for the u.s. preventive services task force**. *Ann Intern Med* 2006, **145**:209–223.

57. Rossi E, Olynyk JK, Jeffrey GP: **Clinical penetrance of C282Y homozygous HFE hemochromatosis**. *Expet Rev Hematol* 2008, **1**:205–216.

58. * * Cooper DN, *et al.*: **Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease**. *Hum Genet* 2013, **132**:1077–1130.

59. Lek M, *et al.*: **Analysis of protein-coding genetic variation in 60,706 humans**. *Nature* 2016, **536**:285–291.

60. Walsh R, *et al.*: **Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples**. *Genet Med* 2016, **19**:192.

61. * Sahni N, *et al.*: **Widespread macromolecular interaction perturbations in human genetic disorders**. *Cell* 2015, **161**: 647–660.

62. * Wei X, *et al.*: **A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations**. *PLoS Genet* 2014, **10**:e1004819.

63. Zhong Q, *et al.*: **Edgetic perturbation models of human inherited disorders**. *Mol Syst Biol* 2009, **5**.

64. * Barrera LA, *et al.*: **Survey of variation in human transcription factors reveals prevalent DNA binding changes**. *Science* 2016, **351**:1450–1454.

65. * Fuxman Bass JI, *et al.*: **Human gene-centered transcription factor networks for enhancers and disease variants**. *Cell* 2015, **161**:661–673.

66. Stefl S, *et al.*: **Molecular mechanisms of disease-causing missense mutations**. *J Mol Biol* 2013, **425**:3919–3936.

67. Schenone M, *et al.*: **Target identification and mechanism of action in chemical biology and drug discovery**. *Nat Chem Biol* 2013, **9**:232.

68. * * Pejaver V, *et al.*: *MutPred2: inferring the molecular and phenotypic impact of amino acid variants.* bioRxiv; 2017.

69. Wang X, *et al.*: **Three-dimensional reconstruction of protein networks provides insight into human genetic disease**. *Nat Biotechnol* 2012, **30**:159–164.

70. * * Meyer MJ, *et al.*: **Interactome INSIDER: a structural interactome browser for genomic studies**. *Nat Methods* 2018, **15**:107.

71. * Gulko B, *et al.*: **A method for calculating probabilities of fitness consequences for point mutations across the human genome**. *Nat Genet* 2015, **47**:276.

72. * The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**: 57–74.

73. * Hopf TA, *et al.*: **Mutation effects predicted from sequence co-variation**. *Nat Biotechnol* 2017, **35**:128.

74. Wright A, *et al.*: **A polygenic basis for late-onset disease**. *Trends Genet* 2003, **19**:10.

75. * * Chen S, *et al.*: **An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders**. *Nat Genet* 2018, **50**:1032–1040.

76. Sufan RI, Jewett MAS, Ohh M: **The role of von Hippel-Lindau tumor suppressor protein and hypoxia in renal clear cell carcinoma**. *Am J Physiol Ren Physiol* 2004, **287**:F1–F6.

77. Kaelin WG: **The von Hippel-lindau tumor suppressor protein**. *Update* 2007, **435**:371–383.

78. Sahni V, *et al.*: **BMPR1a and BMPR1b signaling exert opposing effects on gliosis after spinal cord injury**. *J Neurosci* 2010, **30**:1839–1855.

79. Racacho L, *et al.*: **Two novel disease-causing variants in BMPR1B are associated with brachydactyly type A1**. *Eur J Hum Genet* 2015, **23**:1640–1645.

80. Takano K, *et al.*: **An X-linked channelopathy with cardiomegaly due to a CLIC2 mutation enhancing ryanodine receptor channel activity**. *Hum Mol Genet* 2012, **21**:4497–4507.

81. Koczok K, *et al.*: **A novel point mutation affecting Asn76 of dystrophin protein leads to dystrophinopathy**. *Neuromuscul Disord* 2018, **28**:129–136.

82. Aneichyk T, *et al.*: **Dissecting the causal mechanism of X-linked dystonia-parkinsonism by integrating genome and transcriptome assembly**. *Cell* 2018, **172**. 897–909 e21.

83. Hua JT, *et al.*: **Risk SNP-mediated promoter-enhancer switching drives prostate cancer through lncRNA PCAT19**. *Cell* 2018, **174**. 564–575 e18.

84. Costanzo M, *et al.*: **A global genetic interaction network maps a wiring diagram of cellular function**. *Science* 2016, **353**. pii: aaf1420.

85. Kuzmin E, *et al.*: **Systematic analysis of complex genetic interactions**. *Science* 2018, **360**.

86. Horlbeck MA, *et al.*: **Mapping the genetic landscape of human cells**. *Cell* 2018, **174**. 953–967 e22.

87. Young JE, *et al.*: **Elucidating molecular phenotypes caused by the SORL1 Alzheimer's disease genetic risk factor using human induced pluripotent stem cells**. *Cell Stem Cell* 2015, **16**:373–385.

88. Cheng-Hathaway PJ, *et al.*: **The Trem2 R47H variant confers loss-of-function-like phenotypes in Alzheimer's disease**. *Mol Neurodegener* 2018, **13**:29.

89. Hauser AS, *et al.*: **Pharmacogenomics of GPCR drug targets**. *Cell* 2018, **172**:41–54 e19.

90. Yersal O, Barutca S: **Biological subtypes of breast cancer: prognostic and therapeutic implications**. *World J Clin Oncol* 2014, **5**:412–424.

91. Huang KL, *et al.*: **Proteogenomic integration reveals therapeutic targets in breast cancer xenografts**. *Nat Commun* 2017, **8**:14864.

92. Zhang H, *et al.*: **Integrated proteogenomic characterization of human high-grade serous ovarian cancer**. *Cell* 2016, **166**: 755–765.

93. Chen TW, *et al.*: **APOBEC3A is an oral cancer prognostic biomarker in Taiwanese carriers of an APOBEC deletion polymorphism**. *Nat Commun* 2017, **8**:465.

94. Lomberk G, *et al.*: **Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes**. *Nat Commun* 2018, **9**:1978.

95. Alvarez MJ, *et al.*: **A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors**. *Nat Genet* 2018, **50**:979–989.

96. * * del-Toro N, *et al.*: *Capturing variation impact on molecular interactions: the IMEx Consortium mutations data set*. bioRxiv; 2018.

97. * Fowler DM, *et al.*: **High-resolution mapping of protein sequence-function relationships**. *Nat Methods* 2010, **7**:741.

98. * Fowler DM, Fields S: **Deep mutational scanning: a new style of protein science**. *Nat Methods* 2014, **11**:801.

99. * Starita LM, *et al.*: **Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis**. *Proc Natl Acad Sci Unit States Am* 2013, **110**:E1263–E1272.

100. * Starita LM, *et al.*: **Massively parallel functional analysis of BRCA1 RING domain variants**. *Genetics* 2015, **200**:413–422.

101. * * Wagih O, *et al.*: *Comprehensive variant effect predictions of single nucleotide variants in model organisms*. 2018.

102. Schwarz JM, Cooper DN, Schuelke M, Seelow D: **MutationTaster2: mutation prediction for the deep-sequencing age**. *Nat Methods* 2014 Apr, **11**:361–362.