Edgotype: a fundamental link between genotype and phenotype

Nidhi Sahni1,2, Song Yi1,2, Quan Zhong1,2, Noor Jailkhani1,2, Benoit Charloteaux1,2, Michael E Cusick1,2 and Marc Vidal1,2

Classical ‘one-gene/one-disease’ models cannot fully reconcile with the increasingly appreciated prevalence of complicated genotype-to-phenotype associations in human disease. Genes and gene products function not in isolation but as components of intricate networks of macromolecules (DNA, RNA, or proteins) and metabolites linked through biochemical or physical interactions, represented in ‘interactome’ network models as ‘nodes’ and ‘edges’, respectively. Accordingly, mechanistic understanding of human disease will require understanding of how disease-causing mutations affect systems or interactome properties. The study of ‘edgetics’ uncovers specific loss or gain of interactions (edges) to interpret genotype-to-phenotype relationships. We review how distinct genetic variants, the genotype, lead to distinct phenotypic outcomes, the phenotype, through edgetic perturbations in interactome networks altogether representing the ‘edgotype’.

Addresses
1 Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
2 Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Corresponding author: Vidal, Marc (marc_vidal@dfci.harvard.edu)

Current Opinion in Genetics & Development 2013, 23:649–657
This review comes from a themed issue on Genetics of system biology
Edited by Shamil Sunyaev and Fritz Roth
For a complete overview see the Issue and the Editorial
Available online 26th November 2013
0959-437X/$ – see front matter, © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.gde.2013.11.002

Introduction
Genotype-to-phenotype relationships are far more complicated than the ‘one-gene/one-enzyme/one-function’ paradigm introduced decades ago [1]. Different variants of the same gene may cause different functional defects of the corresponding gene product and as a consequence different diseases (allelic heterogeneity), while the same disease can be caused by mutations in different genes (genetic heterogeneity) [2,3]. The confounding phenomena of incomplete penetrance and variable expressivity are encountered far more often than expected [4]. With increasing number of genomic variants potentially associated with disease being identified by genome-wide association studies [5] and next-generation sequencing [6], it is more imperative than ever to work out underlying principles of genotype-to-phenotype relationships [7].

Genes and gene products do not act in isolation but rather interact with each other within intricate and dynamic ‘interactome’ networks, depicted as nodes and edges representing individual molecules and their mutual interactions, respectively [7,8]. Interactome networks provide an informative platform to investigate functional properties of cellular systems [9]. Comprehensive mapping of protein–protein interaction (PPI) networks has been informative for several human diseases, ataxia [10–12], autism [13], Huntington disease [14,15] and breast cancer [16–18] included.

In network representations, a genotypic variation can be modeled either as knockout or knockdown of gene function, leading to removal of a node and all of its edges, or alternatively, as interaction-specific ‘edgetic’ perturbation, leading to the removal or addition of specific interactions while other edges remain unperturbed [19**] (Figure 1). These specific perturbations of interactome networks arising from genetic variants can give rise to distinct phenotypic outcomes [20]. Edgetic network perturbation models, which emphasize the disruption of specific edges, complement classic gene-centric paradigms [21], which assess the effects of deleting or overexpressing genes, but with few exceptions [22] neglect the influence of genetic variation [23**]. Edgetics can help make sense of confounding genetic heterogeneity, and puts forth direct mechanistic connections from genotype to phenotype [19**,24**,25]. Edgetic modeling is not limited to protein-protein interactions but can be applied to any type of biomolecular interaction. Edgetic perturbation models are emerging as a powerful strategy for interpretation of genotype-to-phenotype relationships.

High-quality interactome networks
Before knowing which interactions are perturbed by particular mutations in a particular gene it is necessary to know the interactions of the wild-type non-mutated protein. Therefore, building comprehensive ‘reference’ interactome networks is clearly the first step for studying edgetic perturbations. High-throughput experimental approaches generate systematic and well-controlled data. They either test all binary combinations of possible protein pairs to determine which ones interact directly [26], or identify protein membership of protein complexes isolated
Genetic variant-induced perturbations in network properties give rise to altered phenotypes, such as disease. Distinct genetic variants of the same gene can exhibit different interaction profiles: loss of all interactions (node removal), loss of some interactions (edgetic), no loss of interactions (pseudo-wildtype), or gain-of-interaction. Nodes represent macromolecules, and edges represent biochemical or biophysical interactions between them. The stars denote a disease-associated variant or mutation. The profile of edgetic perturbations defines the edgotype, providing the explanatory connections between genotype and phenotype.

from cells, that is, co-complex associations [27]. Mapping of the binary interactome is carried out primarily by enhanced variants of yeast two-hybrid methodologies followed by orthogonal assays for validation [28,29]. Mapping of the co-complex interactome is carried out primarily by affinity purification followed by mass spectrometry [30].

Previous high-quality binary protein–protein interaction mapping efforts have identified an appreciable fraction of interactome networks for human [31,32] as well as for model organisms [33–37]. A new generation of binary interactome mapping is underway with enhanced network completeness and resolution. That enhancement comes partly from forceful implementation of an empirical framework that quantitatively assesses the quality, coverage and size of interactome maps [38–40]. Empirical determination of quality has most readily been applied to binary interactome mapping methods [36,41] but adaptations to measure co-complex interactome quality have been pursued [42].

Network perturbations underlie genotype-to-phenotype relationships

Mutations can alter interactome networks due to either node removal or edgetic perturbations [19**,43*]. Truncating mutations, including out-of-frame indels and nonsense mutations, are most likely node removal perturbations, although small in-frame indels could potentially be edgetic [19**]. In-frame missense mutations could be either edgetic, disrupting a protein interaction, or they could destabilize the protein, in effect being node removal (Figure 1). The proportion in each category is unclear as yet. Some reports argue that nearly all missense mutations would be destabilizing [21,44,45]. Other reports estimate that a considerable portion of known Mendelian missense mutations could be edgetic, especially single amino acid changes in protein-binding interfaces [19**,20*,24**,46]. An analysis of protein–protein interaction defects of 29 disease-causing missense alleles identified both node removal and edgetic perturbations in roughly equal proportion [19**].

In accounting how distinct types of network perturbation, node removal, vs. edgetic, can result in distinct phenotypic outcomes, edgetic models shed new light on perplexing genotype–phenotype correlations in human disease. For example, von Hippel-Lindau (VHL) syndrome is an autosomal dominant disorder that manifests as diverse tumors. Mutations associated with Type 1 VHL syndrome (hemangioblastomas and renal cell carcinoma) are typically truncating mutations. In contrast, mutations associated with Type 2 VHL syndrome (pheochromocytoma) are typically missense mutations. Missense mutations on the surface of the VHL protein present a
much higher risk for pheochromocytoma than those in the core of the protein [47], in agreement with a model that these surface mutations are disrupting specific edges (PPIs) which are then responsible for the development of pheochromocytoma [48].

Edgetic perturbations verified so far in human disease represent loss-of-interaction, but theoretically gain-of-interaction alleles (Figure 1), akin to gain-of-function alleles in classical genetics [49], are also possible. Consistent with an edgetic gain-of-interaction is the interaction of nardilysin (NRD1) with the p53 tumor suppressor. Only the R273H mutant of p53 associates with NRD1, not wild-type p53 or the R280K or R273C mutants of p53 [50]. The p53 R273H-NRD1 interaction is critical for promotion of cellular invasion. Systematic searches for edgetic gain-of-interactions, although technically more challenging, are feasible in the near future.

**Virus–host protein interactions as a surrogate for edgetic perturbations**

Similar to human genetic variation, pathogens induce disease states by perturbing host cellular networks [51,52]. This common observation has inspired examinations of the global landscape of host perturbations by proteins from diverse viruses [42,53**,54]. These efforts have uncovered novel disease gene associations. Systematic interaction mapping of human papillomavirus (HPV) proteins to host proteins provided evidence that supported genome-sequencing observations implicating mutations in *MAML1*, a Notch pathway component, in tumorigenesis [53**,55]. *MAML1* has since been confirmed as a gene mutated in cancer [56,57].

Systematic viral-host perturbation mapping functionally characterized FAM111A protein as a previously unidentified host range restriction factor specifically targeted by Simian Virus 40 (SV40) large T antigen [58]. In a validation of the hypothesis that viral-protein interactions can act as surrogates to identify proteins mutated in human disease (Figure 2) [53**,59], a set of *de novo* arising heterozygous point mutations in FAM111A lie behind two phenotypically related developmental disorders of previously unknown molecular etiology, Kenny-Caffey syndrome (KCS) and osteocraistostenosis (OCS) [60*]. These mutations all map to the same minimal region of FAM111A that is required for large T antigen binding [58], predicting that pathogenesis is due to these mutations disrupting the binding of FAM111A to as yet undiscovered cellular proteins. Experiments to identify these interactors, and then to uncover how the newly identified FAM111A mutations perturb these interactions, are readily envisioned.

**Edgetic approaches to study genetic variants**

Two complementary strategies can identify edgetic perturbations resulting from particular genetic variants: forward edgetics and reverse edgetics (Figure 3) [43*], analogous to the longstanding distinction between classical forward genetics (generate a phenotype then identify the responsible genetic variants) and modern reverse genetics (generate specific mutations in proteins then identify the resulting phenotype) [61].

**Forward edgetics**

Starting from known mutations associated with particular phenotypes, primarily human disease phenotypes, forward edgetics investigates interaction profile changes in proteins constructed to contain these known mutations (Figure 3). Such mutations can be readily introduced by high-throughput technologies of recombinational DNA cloning [62] and site-directed mutagenesis [63,64].

Edgetic analysis may seem to be in its infancy, the term reaching acceptance only recently [60*,65*,66], but the principle of interaction profiling for phenotypic resolution has actually been around for considerable time [67*,68,69,70,71]. These early studies introduced a series of point mutations into a single protein then tested by yeast two-hybrid (Y2H) for loss or maintenance of binary interactions against known interactors. An appealing
The example is the detailed mapping of the interaction between the mammalian retromer complex components Vps26 and Vps35. Based on structural modeling 38 distinct mutations in Vps26 were tested for functional interaction with Vps35 by yeast two-hybrid [72], the results meticulously delineating the binding site between the two proteins.

The earliest study that applied edgetic profiling systematically, that is, testing more than one interactor at a time, characterized 35 actin mutants constructed by charged-to-alanine scanning mutagenesis for binary protein interaction defects with five known actin interactors [67*]. Three classes of perturbation were identified: lethal—removed binding of all tested interactors (in current network terms ‘node removal’); differential—removed binding of some but not all interactors (by current terms edgetic); and ‘unchanged’—did not alter binding of any interactors (by current terms pseudo-wildtype) (Figure 1).

Forward edgetics has helped elucidate disease mechanisms for cancer and other diseases [73,74*]. The ATXN1 gene is frequently mutated in ataxia. ATXN1 mutants containing an increased polyglutamine repeat produce a protein that preferentially interacts with RBM17, which promotes neurotoxicity, whereas there is reduced interaction of mutant ATXN1 with the protein Capicua and reduced neuroprotection [11]. Another telling example of how edgetic modeling can clarify disease etiology is the interaction between mutant Mdm2 and ribosomal proteins (RPs) L5 and L11 in cancer. In response to ribosomal stress, these RPs interact with and inhibit Mdm2, resulting in the stabilization and activation of p53, and inhibition of c-Myc induced lymphomagenesis. The cancer associated C305F missense mutation in the acidic zinc finger domain of Mdm2 results in the loss of Mdm2 binding to L5 and L11. This edgetic perturbation causes loss of the ribosomal stress response and an increase in c-Myc induced tumorigenesis [75]. A last example concerns the Fbxw7 protein, a component of the SCF ubiquitin ligase complex that as part of the SCF complex binds to and degrades the c-Myc transcription factor. Mutations in FBXW7 have been found in multiple cancers including T-cell acute lymphoblastic leukemia, with the R465C missense mutation being the most common. Fbxw7-R465C protein is unable to bind and ubiquitylate c-Myc, resulting in increased c-Myc protein stability, acting to augment the Leukemia Initiating Cell (LIC) population [76]. The R465C mutation specifically affects cancer cells and not normal hematopoietic stem cells, which rarely express c-Myc [76].

Putative edgetic alleles of the same gene, but associated with different diseases, tend to be located in distinct interaction domains and thus likely perturb distinct interactions [24**]. A compelling example of distinct domain to distinct disease relationships comes from TP63 [77] in which two clinically distinct developmental disorders, ectodactyly ectodermal dysplasia (EED) and ankyloblepharon ectodermal dysplasia (AEC), are caused by mutations in two separate domains, one predicted to bind DNA and the other to mediate protein–protein interactions [19**].

Technology has now advanced to the point where thousands of alleles and hundreds of interactors can readily be tested for allelic perturbations [19**]. Quality of the resulting edgetic interactomic data would be assessed by implementation of the empirical framework already applied to reference interactomic data. Large-scale experimental edgetic profiling, assessing variant-associated interaction changes at large-scale with respect to the corresponding wild-type counterpart, is needed to thoroughly test edgetic models. For such efforts wild-type protein interaction partners can be retrieved from available and ongoing proteome-wide interactome mapping projects [26].

**Reverse edgetics**

The reverse edgetics approach systematically searches for specific alleles encoding mutant proteins with desired
interaction defects (Figure 3). A library of alleles is first generated by random mutagenesis, and interaction-defective alleles are then isolated from the library using reverse yeast two-hybrid selections [78,79]. Interaction profiling of the mutant proteins then identifies edgetic alleles, which are further studied functionally to investigate the phenotypic consequences [43*,80**]. The earliest application of reverse edgetic profiling targeted the E2F–DP1 interaction, a key interaction in regulation of cell proliferation, and obtained single amino acid mutations identifying a putative helix in a region conserved among E2F family members as the critical determinant for the interaction [79].

A seminal investigation of the C. elegans anti-apoptotic protein CED-9 found that systematic isolation of edgetic alleles is both practical and a fruitful strategy for investigation of gene function [80**]. CED-9 interactors were mapped by Y2H and co-affinity purification, and interaction-defective alleles were selected based on their edgetic perturbation profiles. Structural analyses confirmed that mutations in edgetic alleles specifically affect protein binding sites. In vivo characterization of CED-9 edgetic alleles demonstrated that distinct alleles were associated with distinct phenotypes.

Mapping domain–domain interactions
Analysis of resolved structures of many protein complexes shows that physical interactions can be grouped into two broad classes: domain–domain interactions and domain–motif interactions [81]. Most proteins are composed of multiple domains that bind to specific partners [82,83]. The structures of many interaction domains, including SH2, SH3, PTB, and PDZ domains, have been solved [83,84].

Fragmenting an interacting protein into domain-sized pieces and then determining which region of the protein is responsible for the interaction has been practiced almost since yeast two-hybrid technologies for binary interaction mapping were first implemented [85–87]. Fragmentation for determination of the ‘minimal region of interaction’ is commonplace because of the ease and convenience of the DNA-based technology behind Y2H. Such fragmentation can now be done systematically at large scale for hundreds if not thousands of proteins.

A first systematic experimental mapping of protein interaction domains used a high-throughput, fragment-based Y2H approach to identify interaction domains for ~200 C. elegans early embryogenesis proteins [88**]. This approach greatly increased the sensitivity of interaction detection while maintaining high specificity, leading to a fuller interactome network for embryogenesis.

Edgetics beyond protein–protein interactions
In cellular network models the nodes represent not only proteins, as discussed so far, but also DNA or RNA entities or metabolites, while the edges are physical or biochemical interactions between them (Figure 4). The edgetic strategy is accordingly not limited to protein–protein interactions. Edgetic perturbations of any type of biomolecular interaction can have relevance for improved understanding of disease etiologies.

Protein–DNA interactions
For an example of edgetic perturbation of protein–DNA interactions (PDI) possibly causative of human disease, take the S128F mutation in the T-box transcription factor Tbx19, which leads to adrenocorticotropin hormone deficiency due to virtually complete loss of DNA binding affinity [89]. Another example is the I156V mutation of the transcription factor H-Twist, leading to Baller-Gerold syndrome. This mutation resides in the highly conserved Helix II domain of this gene, and disrupts the interaction interface with DNA targets [90]. Missense mutation H275R in KLF3, encoding a Krüppel family zinc finger transcription factor, is associated with various cardiovascular defects. This point mutation, by abolishing a highly conserved His residue, impairs Klf3 binding to its canonical DNA binding sequence [91].

Protein–RNA interactions
Allelic perturbations of protein–RNA interactions may also play a role in disease etiology. The PRKRA gene, which encodes the protein PACT/RAX, is an important regulator for ear and craniofacial development. PACT mediates the assembly of the RNA Induced Silencing Complex (RISC), and is required for the biogenesis of
miRNAs which are involved in RNA silencing. A mutant protein carrying the missense mutation S130P is unable to bind dsRNA, although otherwise seems normal in tested biochemical functions. This mutant shows defective ear and craniofacial development [92,93]. The expression of a unique type of selenoproteins requires the translational recoding of the UGA stop codon to selenocysteine. This process is regulated by a selenocysteine insertion sequence (SECIS) and SECIS-binding protein 2 (SBP2). An R540Q amino acid change in SBP2 alters its RNA binding activity, resulting in abnormal hormone signaling [94].

Protein–metabolite interactions
Defects in protein–lipid interactions might also be edgetic. The E17K missense mutation in the gene AKT1 is associated with various cancers (breast, colorectal and ovarian). This mutation is localized in the lipid-binding pocket of AKT1, and therefore alters the electrostatic interactions and forms new hydrogen bonds with a phosphoinositide ligand. As a consequence, the mutant protein disrupts lipid binding, stimulating downstream signaling to induce cancer transformation [95]. Another interesting example is the association of nephrin with signaling microdomains, also known as lipid rafts. Lipid rafts spatially organize glomerular structures under physiological conditions [96]. The two mutations C265R and V822M in nephrin lead to a dysfunctional complex due to defective cell surface targeting and ineffective association with lipid rafts, predisposing to a relapsing phenotype [97].

Conclusions
While the usefulness of interactome maps to identify new candidate disease genes and modifier genes has become evident [12,16,31,32], the task of investigating the impact of diverse genomic variants on interactome networks is just underway. Edgetic profiling has already proven insightful in deciphering molecular etiologies of several Mendelian diseases. Deeper edgetic studies on disease mechanisms can now be done at large-scale and at high-throughput [43*,98].

While applied so far mostly to Mendelian disorders, as illustrated by the examples provided, there is no reason why the edgetics strategy could not be extended to complex trait disorders [23**]. Still, it has been frequently noted that even ‘simple’ Mendelian traits are highly ‘complex’ [99]. The line between Mendelian disorders and complex disorders is a gray one indeed. For either, the edgetics paradigm should guide prioritizations of disease-causing variants emerging from GWAS and next-generation sequencing projects, and provide explanations of mutation-specific disease outcome. Functional characterization of the effects of genetic variants on edgotypes will undoubtedly facilitate understanding of how network perturbations relate genotypes to phenotypes.

References and recommended reading

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

* of special interest
** of outstanding interest


This paper explicitly defines the edgetic concept for delineation of genotype-to-phenotype relationships in inherited human disease.


This paper computationally predicts human diseases that arise from mutational perturbations of protein–protein interactions.


This perspective argues forcefully that increased insight into the pathophysiology of complex traits requires a turn towards working out functional impacts of genetic variation in candidate disease genes.


This paper uses protein structural information to evaluate what proportion of human diseases may potentially arise from mutations that disrupt protein–protein interactions.


How the Rb tumor suppressor structure and function was revealed by the study of Adenovirus and SV40. *Virology* 2009, 384:274-284.


This paper presents and interprets a large-scale systematic mapping of viral protein perturbations of host interactome networks, and relates those perturbations to human disease.


60. Unger S, Gorna MW, Le Bechec A, Do Vale-Pereira S,

In this paper newly discovered mutations in FAM1111A causative of two developmental disorders are predicted to be edgetic in their action, that is, disrupt critical interactions of FAM111A with cellular proteins.


This contribution comprehensively surveys the impact of network biology for interpreting complex genotype-to-phenotype relationships in disease.


This paper represents the first systematic edgetic study, methodically investigating the effect of single amino acid changes on protein interactions of actin.


In this paper PHO82 variants associated with neuroblastoma pathogenesis are shown to lose the ability to bind to the key interacting protein HPC111D, representing a fortuitous demonstration of the path from genotype to phenotype.


This paper presents a proof-of-principle demonstration of the reverse edgetics strategy for introduction of specific edgetic alleles that disrupt critical functional interactions of a key apoptotic protein.


This paper introduces large-scale systematic fragmentation of proteins for delineation of minimal region of interaction for hundreds of proteins, providing critical information on their functional domain organization and domain-domain interactions.


