Uncovering the domain language of protein functionality and cell phenotypes using DANSy.

Adrian A. Shimpi[1,2], Kristen M. Naegle[1,2,*]

1 Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, 22903
2 Department of Genome Sciences, University of Virginia, Charlottesville, VA, 22903

*** Correspondence: kmn4mj@virginia.edu**

Abstract

Evolution has developed a set of principles that determine feasible domain combinations, analogous to grammar within natural languages. Treating domains as words and proteins as sentences, made up of domain words, we apply a linguistic approach to represent the human proteome as an n-gram network, which we call hereafter as Domain Architecture Network Syntax (DANSy). Combining DANSy with network theory, we explore the abstract rules of domain word combinations within the human proteome and identify connections that determine feasible protein functionality. We analyze the entropic information content of these domain word connections to establish a DANSy network that balances recovering most of proteome with n-gram complexity. Additionally, we explored subnetwork languages by focusing on reversible post-translational modifications (PTMs) systems that follow a reader-writer-eraser paradigm. We find that PTM systems appear to sample grammar rules near the onset of the system expansion, but then converge towards similar grammar rules, which stabilize during the post-metazoan switch. For example, reader and writer domains are typically tightly connected through shared n-grams, but eraser domains are almost always loosely or completely disconnected from readers and writers. Additionally, after grammar fixation, domains with verb-like properties, such as writers and erasers, never appear together – consistent with the idea of natural grammar that leads to clarity and limits futile enzymatic cycles. Given how some cancer fusion genes represent the possibility for the emergence of novel language, we investigate how cancer fusion genes alter the human proteome n-gram network. We find most cancer fusion genes follow existing grammar rules. Finally, we adapt our DANSy analysis for differential expression (deDANSy) analysis to determine the relationship of coordinated changes in domain language syntax to cell phenotypes. We applied deDANSy to RNA-sequencing data from SOX10-deficient melanoma cells, finding that we can use network separation and syntax enrichment to characterize the molecular basis of cell phenotypes and identify novel information distinct from gene set enrichment analysis (GSEA) approaches. Collectively, these results suggest that n-gram based analysis of proteomes is a complement to direct protein interaction approaches, is more fully described than protein-protein interaction networks, and can be used to provide unique insights for signaling pathway enrichment analysis.

Introduction

Domains are modular units of structure and function that enable protein complex formation or translate biochemical information between signaling effectors[1]. About half of the human proteome is comprised of multidomain proteins, where domains help define overall protein functionality through their independent contributions. The combination of domains within a protein, or domain architecture, arises primarily through the shuffling of preexisting domains, rather than the emergence of new domains[2–4]. Evolutionary jumps, encoded within the domain architectures of protein families, can predict the acquisition of new protein functions[3, 5]. Changes in domain architectures most commonly occur from the gain or loss of domains from the terminal ends of preexisting proteins[6]. Interestingly, only a small

fraction of possible domain combinations are observed, which cannot be described by the random shuffling of domains during genetic recombination events[3, 4, 7], but can be attributed to few domains having multiple domain partners[8, 9]. These observations have motivated the representation of proteins as vectors comprised of their domains for evaluating the evolution of protein families and the complete proteome[3, 5, 10]. Further, these representations have been used to predict the subcellular localization and gene ontology terms for individual proteins[11, 12] demonstrating the breadth of information encoded within overall protein domain architectures.

Insights from broad surveys of both protein domain architectures and amino acid sequences suggest that proteins operate with sets of rules akin to grammar within natural languages[10, 13, 14]. However, the development of protein language models often focus on the amino acid sequence to predict the structure, function, and evolution of unresolved proteins or guide novel protein design[13, 15–17]. Despite the breadth of information encoded within domain architectures few linguistic approaches have used domains as the fundamental unit of the protein language. Applications that utilize domains as "words" within a protein have shown linguistics can recover protein functionality and evolution across the tree of life independent of known protein interaction networks or signaling pathways[5, 10]. However, these past methods rarely consider the word ordering and word repeats, which likely reflect the evolutionary pressures related to the non-random shuffling of domains and the observation of the predominant modification of domain architectures at the terminal ends[3, 4, 6, 7]. The sequential order can have important consequences on protein functions like modifying the avidity or specificity of catalytic domains like the tyrosine kinase domain[18, 19]. Like natural languages, there are likely constraints on the relative location and combination of certain word types that make up human proteins. For example, catalytic domains might represent verbs and the binding domains they appear with are adverbs, or possibly as the nouns or subject of the sentence that help define the action the verb will have. Certain types of adverbs or the combination of too many verbs without modifying nouns may be prohibited from a language that ultimately needs to be interpreted by the biochemical networks of the cell, while also maintaining sentence clarity (akin to minimizing energy usage).

N-gram analysis is a linguistic approach which maintains both the composition and sequential order of words within natural languages and can be adapted to protein domain architectures. By treating individual domains as words and the complete domain architecture as a sentence, n-grams with n domains can be extracted from either single or multiple domain architectures. A 2-gram model has shown that pairwise domain combinations can recapitulate the evolution of proteome complexity[10]. The smaller fraction of multidomain proteins in prokaryotes than eukaryotes[20] has limited n-gram analysis to mostly 2-gram models for comparative genomic studies. However, certain 2-3 domain combinations, or supra-domains, are rearranged together as a unit across protein families[21]. If an n-gram model does not extract n-grams longer than these supra-domains, the diversity of feasible domain combinations in a proteome may not be sufficiently recovered.

Here, we evaluate n-gram models of various lengths to describe the human proteome. We integrate these models with network analysis techniques to identify protein domains and multidomain n-grams that act as hubs to connect obligate domain families, enabling a simplification of the overall network. An advantage of representing the human proteome as a language-based network is that all proteins have defined domains, unlike networks that rely on protein-protein interactions that have yet to be fully annotated[22, 23]. However, the connections within the network reflect more about the related functional connections amongst keywords in the human proteome than it does direct protein-protein interactions. Here, we develop and test an n-gram network framework termed Domain Architecture Network Syntax (DANSy) at the level of the entire proteome and characterized the entropic information needed to recover most of the proteome. Next, we applied n-gram analysis to measure the emergent properties of the specific words and languages within reader-write-eraser systems that coordinate post-translational modifications (PTMs). Surprisingly, we find that despite vastly different evolutionary timescales, most reversible PTM-systems convergently evolved to have tight connections between the readers and writers and very loosely, if at all connected, erasers. Looking across evolution, it appears that how PTM systems sampled different word ordering gives a possible measure of the relative time since the appearance of the system and the time at which the language becomes fixed in terms of its grammar rules and composition.

We then ask if DANSy analysis can delineate how cancer gene fusions alter, or not, the functional connections within the proteome. Interestingly, we find that predominantly somatic fusion genes arise with the same characteristics that appear to guide the overall evolution of most multidomain protein architectures[3, 4]. Finally, we adapted our DANSy analysis for differential expression (deDANSy) to study the subnetwork languages associated with cell phenotypes. Based on our analysis of multiple RNA-sequencing datasets, we verify that deDANSy like gene set enrichment analysis (GSEA) captures the coordination of molecular functions to establish phenotypes, but can be used to further identify unique molecular effectors overlooked by gene set definitions. Hence, we find that DANSy provides unique insights about the functional connections between proteins and systems within cells at both the whole proteome or within individual signaling subnetworks like the PTM systems or cellular phenotypes.

Results

Generating domain n-gram networks to describe the domain architecture landscape

In order to describe the entire human (or other species) proteome as n-grams, we used the InterPro domain annotations[24] to have a controlled vocabulary coupled with UniProt[25], which incorporates information from the Pfam[26], SMART[27], and CATH-Gene3D[28] databases. We used our Python-based package CoDIAC[29] to fetch InterPro domains and the domain architectures, which we then convert to n-grams to represent a protein. For example, the SRC and ABL kinase families have the domain architecture SH3|SH2|Kinase, which becomes three 1-grams (SH3, SH2, Kinase), two 2-grams (SH3|SH2, SH2|Kinase), and one 3-gram (SH3|SH2|Kinase). Applying n-gram extraction to the entire human proteome, we found some key characteristics to help understand the principles of the human proteome language. Most of the proteome (95%) contain 10 or fewer domains (Fig. 1B, S1A), and a significant outlier is Titin (TTN, UniProt ID: Q8WZ42) at 303 domains, which has five times as many domains as the next largest, Obscurin (OBSCN, UniProt ID: Q5VST9) at 66 domains (Fig S1A). We therefore set 66 domains as the maximum n-gram length, which produced 44,425 n-grams from >18000 proteins. The majority of n-grams were found in only one protein (Fig. 1C), suggesting a specialized syntax that is not shared between protein sentences. On the other end of the distribution, some n-grams reuse the same domain word. In the top ten n-grams, we found the protein kinase domain, the seven transmembrane region of rhodopsin-like G-protein coupled receptors (GPCR_Rhodpsn_7TM), and the Zinc Finger C2H2 type (Znf-C2H2) domain (found in eight of the top ten most recurring n-grams) (Fig. S1B). Since n-grams with Znf-C2H2 domain repeats dominated the top n-grams, we further explored the top 1-gram or heterotypic (containing more than 1 type of domain) n-grams. We found almost exclusively 1-grams, such as the EGF-like and pleckstrin homology (PH) domains, in the top 10 with 200-350 proteins per 1-gram. Meanwhile, the KRAB|Znf-C2H2|Znf-C2H2 was the only heterotypic n-gram identified in the top 10 (Fig S1C). These results highlight that certain protein families, such as the Znf-C2H2 containing proteins, dominate n-gram counting metrics and may obscure the importance of other critical domain families, but there is also a richness of different domain combinations.

Domain n-grams represent feasible domain combinations, which encode how compatible the functionality of individual domain words. When domain n-grams are shared within protein sentences, a functional property of the n-gram related to its compatibility with other domains is captured. We use these abstract connections to build a network of the proteome where nodes represent individual n-grams (e.g. SH2, SH3|SH2), and edges designate parent-child relationships where shorter n-grams are found within the longer n-grams (Fig. 1D). For example, for the SH3|SH2 2-gram found within the SRC and ABL kinase family architecture, it is connected directly to the SH2 and SH3 1-grams, and the SH3|SH2|Kinase 3-gram. We extended this to the 44,245 n-grams we extracted from the proteome to construct a complete n-gram network (Fig. 1D). We used gross topographical network features, like the number of connected components, to identify n-gram families that share a common set of words. The n-gram network of the proteome contained 1345 connected components and 700 isolates (Fig. S2A). Each isolate represent a singular domain found only within a single family of proteins. Most non-isolate

connected components, which contain 2 or more domain words, represented 5 or fewer proteins except the largest connected component, which represented 9937 proteins (Fig. 1E). Since relatively few domains have a diverse number of domain partners[8], and whole protein families may be represented within individual connected components of the n-gram network, we integrated a node collapsing step during n-gram network construction to reduce redundant n-grams representing the same set of proteins (Fig. 1F). The collapsed network (Fig. 1G) increased the number of isolates to 1106 nodes with most representing n-grams of length 5 or less (Fig. S2). Altogether, this suggests that the collapsed n-gram network maintains connections that represent non-redundant grammar structures.

Next, we used node-specific metrics to identify critical n-grams of the network. For this, we calculate two complementary centrality metrics: 1) the degree centrality – to identify the domains or domain combinations with a diverse set of domain partners, and 2) the betweenness centrality – to identify domain n-grams that connect n-gram families by measuring the shortest path between individual nodes. Focusing on the largest connected component, since it represents the most protein families and potential functions, only a small fraction of n-grams have high values (>0.01) for both centrality measurements (Fig. 1H). We used the relative relationship between the centrality measurements (Fig. 1I) to define general observations that relate the "grammatical" contexts that capture the functional properties of different n-grams. Specifically, n-grams with high degree and betweenness centrality values, such as the protein kinase, EGF-like, and Znf-C2H2 domains, have a diverse set of domain partners that demonstrate their flexibility in generating word structures that maintain clear functionality. However, n-grams with only high degree centrality values represent domain combinations that have several partners but only connect a distinct n-gram family (e.g. highly repetitive domain n-grams that only link to other repetitive n-grams). If an n-gram has a high betweenness centrality but low degree centrality, like the GPCR_Rhodpsn_7TM domain, then it likely acts as a connection between a small fraction of other n-grams to the rest of the network. In this group (high between, low centrality), even if the individual n-gram is highly prevalent across the proteome, it is highly constrained in what are considered feasible domain partners. This category of n-grams represents grammar structures that rarely require modifiers to define their action. Altogether, our results demonstrate how the analysis of domain architectures represented by a network of n-grams can be used to explore the rules of the protein language and can find abstract connections of protein functionality. We refer to this analysis collectively as Domain Architecture Network Syntax (DANSy) analysis hereafter.

We set out to evaluate how different n-gram lengths alter the grammar structures of the protein language as captured within the DANSy analysis. Importantly, about 25% of the proteome has 3 or more domains and whole protein families can be represented by up to 14-grams (Fig. 1, S2). We constructed n-gram networks with different maximum n-gram lengths up to 15-grams, and compared this to the full n-gram network. We found a 10-gram network retains 85% of the entropic information content of the full n-gram network (Supplementary Note 1). Further, the 10-gram model retained all isolates in the network and only lost nodes within the largest connected component representing longer n-grams. Meanwhile, the 2-gram model recreated about 40% of the information content, but introduced 771 new isolates and retained only half of the other non-isolate connected components, suggesting that much of this information is lost in a 2-gram only network, which were the focus of prior studies[3, 8, 10]. Further, we observed relatively minimal gains in information content and changes in the network topology between 10- and 15-gram models (See Supplementary Note 1 for further details). Thus, for representing the human proteome hereafter, we selected a 10-gram model, which balances maximizing the information encoded within protein domain architectures and complexity.

Domain modules within reversible PTM systems infrequently share n-grams.

We next wanted to evaluate what DANSy analysis can provide on individual signaling subnetworks, like the phosphorylation system, in addition to whole proteome-level insights. We focused on phosphorylation because it is found across many biological processes[30], fall under two broad families (pTyr and pSer/Thr), and has been suggested to have evolved because the biochemical properties of phosphate groups allow it to be rapidly and readily reversible[31]. Signal transduction pathways mediated by phosphorylation are tightly controlled by a three module system which operates under a

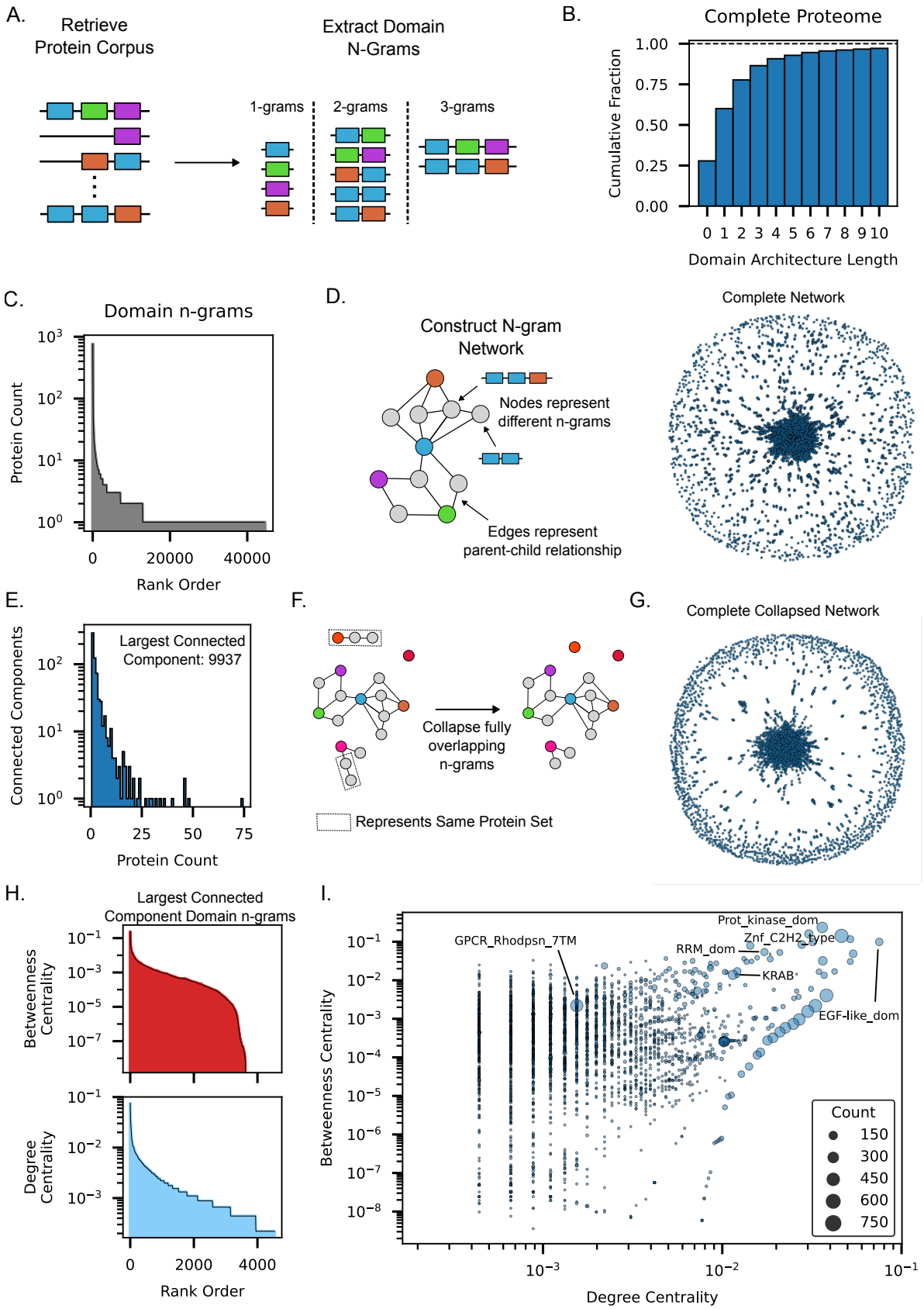

Figure 1. Legend on the following page

Figure 1. Human proteome n-gram network characterization.: A) Overview of extracting n-grams of length n from a collection of protein domain architectures. B) The cumulative distribution of domain architecture lengths across all proteins in the human proteome. C) The number of proteins for each domain n-grams. D) The schematic (left) of constructing an n-gram network where individual n-grams are nodes and edges represent parent-child relationships where a shorter n-gram is found in a longer n-gram. The actual n-gram network of the complete proteome (right). E) The number of proteins represented for each connected component except the largest connected component which has 9937 proteins. F) Schematic of collapsing n-grams that which fully represent the same set of proteins. G) The collapsed n-gram network of the human proteome. H) Network centrality measurements for each node in the largest connected component of the collapsed n-gram network in rank order. I) The relationship of each centrality measurement with each other and marker size representing the protein count of each n-gram.

reader-writer-eraser paradigm. For example, the pTyr machinery consists of the Tyr kinase (writer), Tyr phosphatase (PTP, eraser), and SH2 and PTB (readers) domains. The pTyr system is considered to be evolutionarily newer than the pSer/Thr system, having evolved near the origins of metazoan species[32, 33]. Unfortunately, the transient nature of phosphorylation and the relatively low affinity of reader domains have made protein-protein interaction network definition especially challenging. To evaluate if DANSy can identify features that differ between the two systems, we generated n-gram networks on the components of each phosphorylation system separately (domains and classification in Table S1). Specifically, we generated two different n-gram networks for each system: a Complete model, which contains all n-grams in the system; and, a PTM System Domain Focused model, where n-grams must contain at least one of the domains of the PTM system. By analyzing the PTM System Domain Focused networks, we can study broad differences in the grammar of system, and how it constrains the combinations of different word types represented by each domain to modulate the overall system. The pTyr system generated a complete connected graph (i.e. a single connected component), while the pSer/Thr system had multiple connected components with n-grams containing the phosphatase (eraser), 14-3-3 (reader), or MH2 (reader) domains as individual, disconnected nodes from the rest of pSer/Thr machinery (Fig. 2A, S4A). Comparing the Complete versus PTM System Domain Focused networks of the pTyr system highlighted that only the erasers PTPN6 and PTPN11 with the domain architecture SH2|SH2|PTP use other pTyr regulatory domains. Meanwhile, the SH2 and Tyr Kinase domains share multiple n-grams supporting findings that SH2 domains modulate kinase processivity[19, 34]. Meanwhile, the pSer/Thr system only had the FHA domain (a pSer/Thr reader domain) directly connected to Ser/Thr-kinase domain subnetwork (Fig. 2A,S4A). Collectively, these results suggest a common set of rules exists between the two systems, which includes that eraser domains rarely if at all use the other system components to modify their function and that reader domains can be used as "adverbs" to modulate writer domain targets.

To determine if the network topologies of the pTyr or pSer/Thr systems were characteristic of other reversible PTM systems that share the reader-writer-eraser paradigm, we generated networks for the acetylation, methylation, and ubiquitin systems. From these networks, we found they more closely resembled the pSer/Thr system with multiple connected components (Fig. S4B). However, the methylation system had n-grams with JmjC eraser domains connected to the rest of the network within the PTM domain focused network, and the DOT writer domains were isolated within both networks (Fig. S4A). Notably, the PTM System Domain Focused n-gram networks showed within each PTM system, the eraser domains are rarely found within the n-grams that contain other system components.

We next asked if the PTM System Domain Focused networks were sufficient to recapture the information within the Complete n-gram networks. We calculated the relative entropy between the two models, and found that the domain focused networks encode similar n-gram distributions across each system (Fig. S4C). Collectively, these results suggest these PTM systems have developed a common set of grammatical rules that determine the feasible domain combinations to maintain biochemical functions. Interestingly, these rules suggest that eraser domains do not combine with other domains of the system to carry out their function. Thus, reader domains that act as "adverbs" to modify activity of the writer

domains, rarely are used to modulate erasers. Additionally, these networks suggest a constraint that the catalytic domains ("verbs") - writers and erasers - must be on separate proteins that likely prevents inefficient and futile processing of the PTMs.

Characterizing the evolution of the phosphorylation domain modules.

Since the pTyr n-gram network was the only PTM system that could be represented as a complete graph, and it is considered one of the most recently evolved PTM systems[32, 33], we wanted to determine if the n-gram network topology had evolved over time and could reflect the evolutionary age of the system. We retrieved both the pTyr and pSer/Thr systems from 20 species starting from *Saccharomyces cerevisiae*, which contains a single proto-SH2 domain and three PTP domains[32] (Fig. S5). Assembling the n-gram networks for each species, we observe a rapid expansion of both the n-grams and edges between n-grams for the pTyr system but not the pSer/Thr system (Fig. 2A, S8, S9). The emergence of metazoans led to the stabilization of the pTyr n-gram network, and is reflected in the number of connected components and the relative distributions of individual domains (Fig. 2A,B, S5). For the pSer/Thr system network topology, the separated Ser/Thr phosphatase subnetwork was established early and few changes in the distribution of individual domains (Fig. 2A,B, S6). These observations were further supported by using relative entropy to compare the n-gram distributions for each species to a network generated from all species (Fig. 2C). The relative entropy for the pSer/Thr networks were lower to begin with in most pre-metazoan species, which reflects the increased similarity index of the n-grams found within the pSer/Thr network for these species (Fig. S7). Interestingly, these results suggest a potential convergence of n-grams in both systems during the evolution of vertebrates within our queried species. However, the individual species networks show an interesting evolution of the PTM system, which is not readily observed when using the number of individual domains to study the system evolution alone. The pTyr network of the pre-metazoan species *Capsaspora owczarzaki* and *Monosiga brevicollis* had several n-grams that connect the PTP domain to the rest of the network that are lost within metazoans. However, once processes advantageous to metazoans evolved, only the SH2|SH2|PTP architecture was retained. Interestingly, the lost architectures involved connections between both catalytic domains, but the only metazoan species where a connection remained was in *Nematostella vectensis*, which is one of the earliest metazoans included in our analysis. This suggests during this evolutionary period representing the transition to metazoans, species were sampling several configurations of the network during the rapid pTyr system expansion. However, the pTyr system converged to the similar set of grammatical rules as the other PTM systems, where eraser domains are loosely connected to the rest of the network, and domains with opposing verb actions are not found within the same protein.

Using n-gram networks to study gene fusion domain architectures.

While advantageous domain architectures have been selected for throughout evolution, selective pressures during the development and progression of cancer present new opportunities for the exploration of new grammatical structures. In particular, the development of cancer fusion genes represents one avenue for generating novel architectures by forming chimeric proteins from parent genes with disparate functions. Relative to the complete n-gram network, gene fusions can generate domain architecture that relate to the gross topology of the n-gram network in three different fashions: 1) there is no change as the fusion n-gram already exists within the network as seen with STK11-TYK2 fusion gene, which retains only the two kinase domains from TYK2, 2) there is a reinforcement of existing connections where a novel domain architecture is generated and uses domains that share a common domain or n-gram partner to shorten the network path and reinforce an existing connected component, as seen with the BCR-ABL1 fusion gene, or 3) a bridge is formed between two connected components as the novel domain architecture uses domains that do not share any common domain partners like the FGFR3-TACC3 fusion (Fig 3A).

To investigate the nature and extent of how cancer gene fusions alter n-gram networks, we retrieved gene fusions identified in patients from 20 study cohorts within the Cancer Genome Atlas (TCGA) from ChimerDB[35]. We limited fusions to those mapped to the current human genome build (hg38) and result in in-frame fusions. Predicted domain architectures were generated by mapping genomic

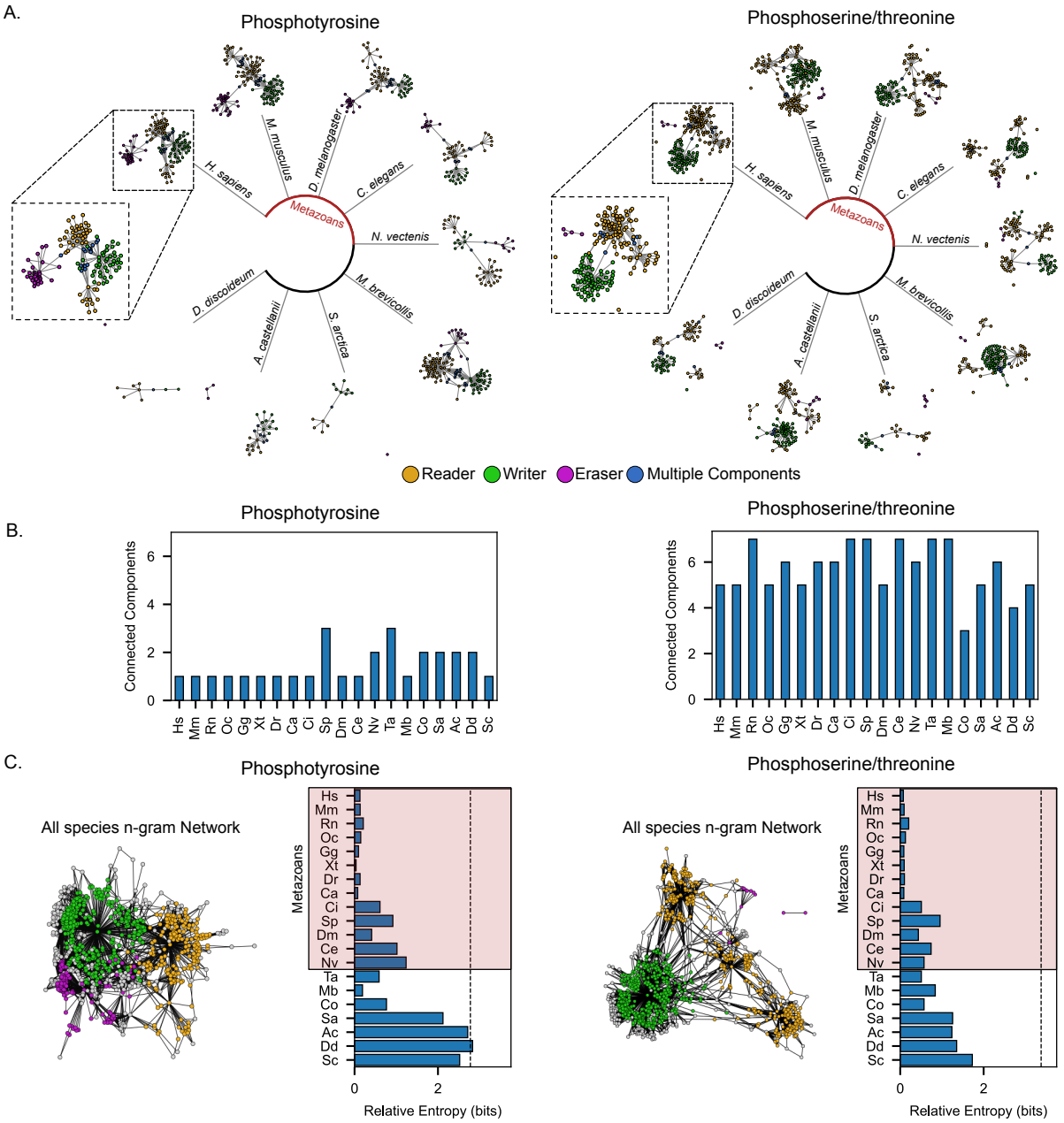

Figure 2. The evolution of the protein domain architectures of the pTyr and pSer/Thr systems. A) Complete n-gram networks of the phosphotyrosine (left) or phosphoserine/threonine (right) systems for species ranging from *Dictyostelium discoideum* to humans. B) The number of connected components for each species phosphorylation system. C) The complete n-gram network containing all n-grams across all queried species and the relative entropy for each species compared to the all species network. Hs: *Homo sapiens*, Ms: *Mus musculus*, Rn: *Rattus norvegicus*, Oc: *Oryctolagus cuniculus*, GG: *Gallus gallus*, Xt: *Xenopus tropicalis*, Dr: *Danio rerio*, Ca: *Carassius auratus*, Sp: *Strongylocentrotus purpuratus*, Dm: *Drosophila melanogaster*, Ce: *Caenorhabditis elegans*, Nv: *Nematostella vectensis*, Ta: *Trichoplax adhaerens*, Mb: *Monosiga brevicollis*, Co: *Capsaspora owczarzaki*, Sa: *Sphaeroforma arctica*, Ac: *Acanthamoeba castellanii*, Dd: *Dictyostelium discoideum*, Sc: *Saccharomyces cerevisiae*

breakpoint information to the protein coding sequence, and identifying the domains donated from each parent gene (Fig. S10). Across 20 cancer types, at least 40% of all unique fusion gene domain architectures do not change the n-gram network with the exceptions of Acute Myeloid Leukemia (LAML) and Testicular Germ Cell Tumors (TGCT). Both of these cohorts had fewer than 50 fusion genes identified and less than 20 different domain architectures (Fig. 3B). Most novel domain architectures reinforced existing domain connections (20-30%), and only 10% of all fusions bridge connected components within the n-gram network (Fig. 3B). Since novel domain architectures present potentially highly advantageous grammatical structures, we next evaluated whether they represent a recurrent fusion across multiple cancer types or within a single cancer. For example, the FGFR3-TACC3 fusion is highly recurrent across multiple cancers[36] and represents a fusion that bridge connected components within the n-gram network. Meanwhile, the TMPRSS2-ERG fusion is highly recurrent within prostate cancer[37], and does not alter the n-gram network. Further, we included fusions whose novel domain architectures represented singleton gene fusions, which have only been identified in one patient. We found that irrespective of whether a fusion was recurrent or a singleton fusion, most fusions still reinforce existing domain combinations (Fig. 3C). Collectively, these results suggest that most novel domain architectures of cancer-related gene fusions adhere to the existing grammar of the human proteome.

Since few gene fusions novelly connected disparate functions, we reasoned these fusions might generate highly aberrant, but advantageous signaling pathways that might result in worse cancer patient survival. Surprisingly, we found in our pan-cancer analysis the fusions that reinforce existing domain combinations were more deleterious (Fig. S11). This suggests fusions that co-opt existing functions increase tumor cell fitness to promote disease progression. However, gene fusions are relatively rare compared to other somatic mutations[38–40], and for individual cancers survival trends were difficult to conclude as most cohorts had less than 25 patients (5-10% of the patient population) with a single fusion gene (Fig S13, S12.

Since we showed that DANSy analysis can identify general rules guiding domain architecture development, we wanted to explore the rules that define the relationship between parent gene domain contributions to the final gene fusion domain architecture. To achieve this, we constructed a modified n-gram network where nodes are the domain architectures contributed by each parent gene and the final fusion gene domain architecture with edges connecting the parent architectures to the fusion. Thus, if a kinase domain was donated from the 3' fusion gene (e.g. with TYK2) it would be a separate node in the network than a kinase domain that represents the fusion gene (e.g. the complete STK11-TYK2 fusion). We found 106 fusion n-grams with unique parent gene contributions and fusion domain architecture (Fig. S14B). Interestingly, we found most fusion genes did not gain a domain within the coding sequence from one of the parent genes (Fig. S14), which explains why most fusions recreate or reinforce established domain architectures (Fig. 3B). Next, we explored the subnetworks of specific domain n-grams to determine whether an n-gram is 1) donated primarily by either the 5' or 3' parent genes or 2) generate novel domain architectures. For example, the KRAB and HD domains are almost exclusively donated by the 5' or 3' parent gene respectively. Most fusions with either of these domains do not have domains contributed from the other parent gene, but when a domain is contributed, it generates a novel domain combination. Meanwhile, the Paired domain, which only comes from the 5' parent gene, always resulted in novel domain architectures (Fig. 3D). We next asked whether the highly abundant Znf-C2H2, protein kinase, and the GPCR-Rhodpsn-7TM domains were differentially involved in gene fusions, which reflects the broad grammatical categories we determined from centrality measurements of the whole proteome. Interestingly, the ability to generate gene fusions reflected this property, where the Znf-C2H2 and protein kinase domains, which represent flexible n-grams that generate multiple word structures, generated more fusion genes with 50% being novel domain architectures (Fig. 1I). Meanwhile, the GPCR_Rhodpsn_7TM domain, which rarely combines with other domain words to define its function, was rarely found in gene fusions and often was the only domain donated to the fusion. Collectively, these results suggest the intrinsic properties of parent gene domains determine their role in generating fusions that are effectively established in observed cancer samples.

We next wished to understand if the existing proteome syntax ultimately describes the diversity of novel domain combinations and what guides them in gene fusions. Using the Gini-Simpson Diversity

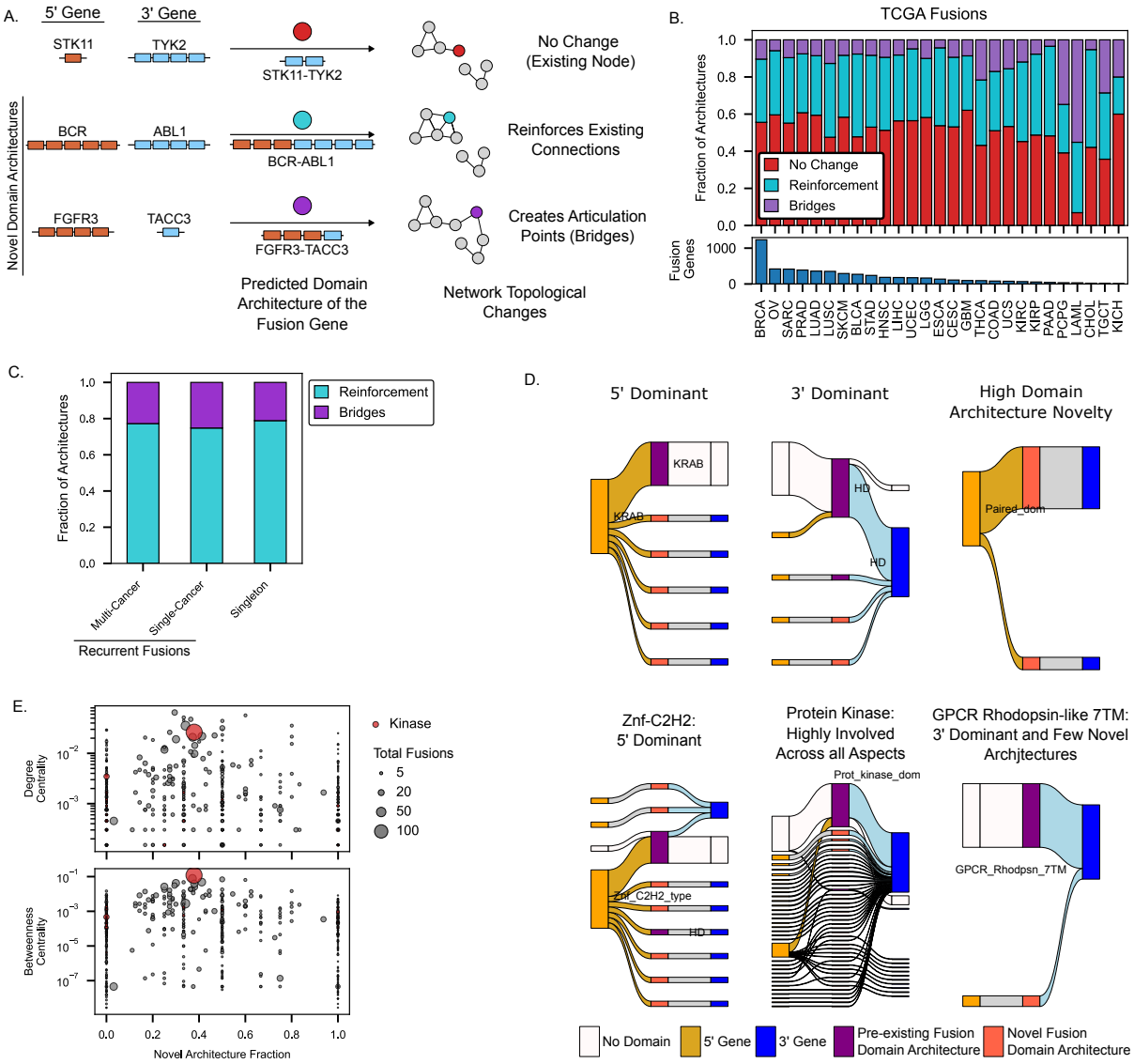

Figure 3. Characterizing predicted protein domain architectures of cancer gene fusions. A) Schematic of the possible changes to the n-gram network caused by individual fusion genes with example fusion genes are predicted domain architectures. For novel domain architectures these can either reinforce existing connected components or reduce the number of connected components. B) The fractional distribution of fusion gene impacts on the n-gram network (top) and the total number of fusion genes (bottom) retrieved for each TCGA study cohort. C) The distribution of architecture n-gram network impacts for recurrent or singleton fusions within either multiple or single cancer. D) Representative Sankey diagrams of select protein domains that exhibit 5'/3' donation propensity and/or generation of novel domain architectures, and diagrams for the top 3 most prevalent domains across the human proteome. E) Network centrality measurements relative to the fraction of novel domain architectures individual domain n-grams generate. N-grams which contain the protein kinase domain are highlighted in red.

Index, we calculate the diversity of domains that precede or follow an individual in the complete proteome to then relate to that n-gram's propensity of being donated by the 5' or 3' gene. In other words, if an n-gram is donated primarily by the 5' parent gene is this because it has a high diversity index for n-grams that follow it, and for 3' dominant n-grams the preceding n-gram diversity. The diversity indices of n-grams within fusions genes skewed slightly higher than the rest of the proteome, but this was not a defining characteristic of the top overall fusion n-grams (Fig. S15). Further, the diversity of both following and preceding n-grams were similar for the most prevalent n-grams in gene fusions irrespective of whether they were primarily donated by the 5' or 3' parent gene. Since the centrality measurements in the n-gram network describe the flexibility and capacity of n-grams to connect with other n-gram families, we asked if these alone explain if an n-gram frequently creates novel domain architectures, but found no correlation (Fig. 3E). Next, we explored if these centrality measurements, combined with the diversity indices, explain whether a domain frequently generates novel architectures in fusion genes. We found no dominant explanation amongst the most prevalent n-grams in fusion genes, but rather individual explanations. For example, the KRAB and Znf-C2H2 domains have an intrinsic property where they are flexible enough to create several n-grams – as reflected by a high degree centrality. From the diversity index, we find that these existing combinations are dominated by a specific n-gram, which suggests that a separate property limits potential partners (Fig. 1I, S15B). Thus, when the KRAB or Znf-C2H2 domains are in gene fusions, they have flexible properties to generate novel domain architectures frequently, but have preferred domain partners, which modify their functionality in specific ways. Conversely, the GPCR_Rhodpsn_7TM domain has both a low centrality measurement and diversity indices almost equal to 0, which indicates it is not found in other n-grams frequently, but has 1 domain word partner it almost exclusively prefers to make longer n-grams. Thus, the GPCR_Rhodpsn_7TM domain was not expected to be involved in many novel domain architectures or gene fusions. Altogether, these results showcase how the grammatical rules uncovered through DANSy analysis, in combination with properties like the diversity of n-gram partners, can contextualize the broad properties determining if a domain word will generate novel domain architectures.

Kinase fusion have been frequently identified across most cancers[41] with the tyrosine kinase domain specifically being overrepresented within gene fusions[39, 42, 43]. The protein kinase domain was found in the most fusion genes, but not limited to being donated by either the 5' or 3' parent gene (Fig. 3D). About 40% of all protein kinase containing fusion genes were novel domain architectures. The grammatical properties from our DANSy analysis and the diversity indices suggest the protein kinase domain creates several word structures that are not dominated by a single n-gram. We have also shown that the pTyr system was constantly testing new connections during its evolution before converging to a similar set of rules seen in the pSer/Thr system. Thus, we next wanted to ask the kinase domains in kinase fusions were using inherent properties of all kinase domains, or represented a continued evolution process seen in the pTyr system. By stratifying kinase fusions by which kinase family donated the kinase domain, we found a similar number of fusion genes and fraction of novel domain architectures from both the pSer/Thr and pTyr kinase domains (Fig. S16A,C). An n-gram network from these kinase fusions revealed only 20 n-grams were common to both kinase fusion types and the most highly connected nodes within the network were either specific to pSer/Thr kinases (AGC-kinase, C-terminal domain) or common to both (FN3, Ig subtype 2, or the PH domains) (Fig. S16B-D). Despite new domain n-grams being generated in both pTyr and pSer/Thr kinase fusions, similar fractions of novel to pre-existing n-grams were observed. The lack of differences between pTyr and pSer/Thr kinase domains suggest the phosphorylation system the kinase domain originates does not determine if it will form novel domain n-grams. Rather, these results suggest that kinase domains generally have intrinsic properties that enable neofunctionalization through diverse domain combinations to reinforce that DANSy analysis uncovers abstract connections of protein and domain functionality.

Applying DANSy to differential expression to uncover the domain language separating phenotypes.

Thus far, we have demonstrated the application of DANSy to describe how proteome-level changes arise or develop through the introduction of new domains or domain combinations. We next wanted to explore the ability of DANSy to uncover what domain n-grams are utilized in describing cell phenotypes. Here, we focus on transcriptomic phenotypes generated from RNA-sequencing, but note that proteomics-defined phenotypes can be similarly analyzed. To achieve this we project the differentially expressed gene (DEG) status, which delineate the genes of different conditions, to n-grams in a dataset-specific network. To determine whether the n-gram subnetworks from the up-regulated DEGs are distinct or overlap with those from down-regulated DEGs, we calculate the network separation, which has been previously used to describe the molecular pathological similarities of diseases[44]. Network separation compares the mean shortest distance of nodes in a subnetwork A to nodes in subnetwork B relative to the distance between nodes within each subnetwork (See Methods and Supplementary Note 2). To avoid situations where an n-gram represents one or few genes, which may weakly bridge the network to drive down separation, we also calculate the enrichment of each n-gram. Additionally, we evaluate the statistical significance of the separation between up and down DEGs n-gram subnetworks by using a combination of thresholding and random sampling. First, we subsample DEGs, which should retain highly enriched and distinct n-gram subnetworks, to create a distribution of network separation values that describe the experimental condition. Then, we construct a null model from randomly selected genes, chosen to match the distribution of domain architecture lengths (capturing highly connected subnetworks generated from long domain architectures, see Supplementary Note 3). Thus, by combining n-gram enrichment, DEG subsampling and network separation we developed a DANSy-based analysis to determine how separated and robustly enriched n-gram neighborhoods are for up and down regulated DEGs. This differential expression DANSy (deDANSy) analysis is implemented through 5 major steps (Fig. 4A):

1. We subsample 70% of the DEGs, while maintaining the relative ratio of up to downregulated DEGs. We then generate a null distribution of the same size with randomly chosen genes that maintains the distribution of domain architecture lengths. The random distribution of genes are then split into either "up" or "down" regulated genes at the same relative fraction as the subsampled genes. The network separation for both the subsampled and random genes is then calculated as a baseline value (ns_i).

2. For both the subsampled and random genes, the statistical enrichment (p_E) of each n-gram is calculated using a hypergeometric distribution.

3. We prune the DEG status of individual n-grams in the network using a p_E threshold, and calculate the network separation. The pruning and network separation are then repeated with increasingly restrictive p_E thresholds.

4. The dispersion of network separation throughout the pruning process is calculated using the interquartile range (IQR_{ns}). The IQR_{ns} provides a measure of the robustness of the statistical enrichment with large values relating to significantly enriched n-grams that are well separated, while small values relate to either highly overlapping or lowly enriched n-gram neighborhoods.

5. Steps 1-4 are repeated 100 times to generate distributions of both ns_i and IQR_{ns} values. Both these distributions are statistically analyzed to generate scores based on the effect size for either the separation or stable enrichment of n-gram neighborhoods (See Methods for further details).

We focused on datasets related to either cytoskeletal/extracellular matrix remodeling, cell culture methods, or cancer biology to capture a variety of applications with variable differences in gene expression[45–50]. Defining DEGs for meaningful comparisons in each dataset, we found that most comparisons yield DANSy networks that are less separated than randomly chosen genes suggesting our analysis captures the broad coordinated nature of phenotypic changes. Interestingly, we found the

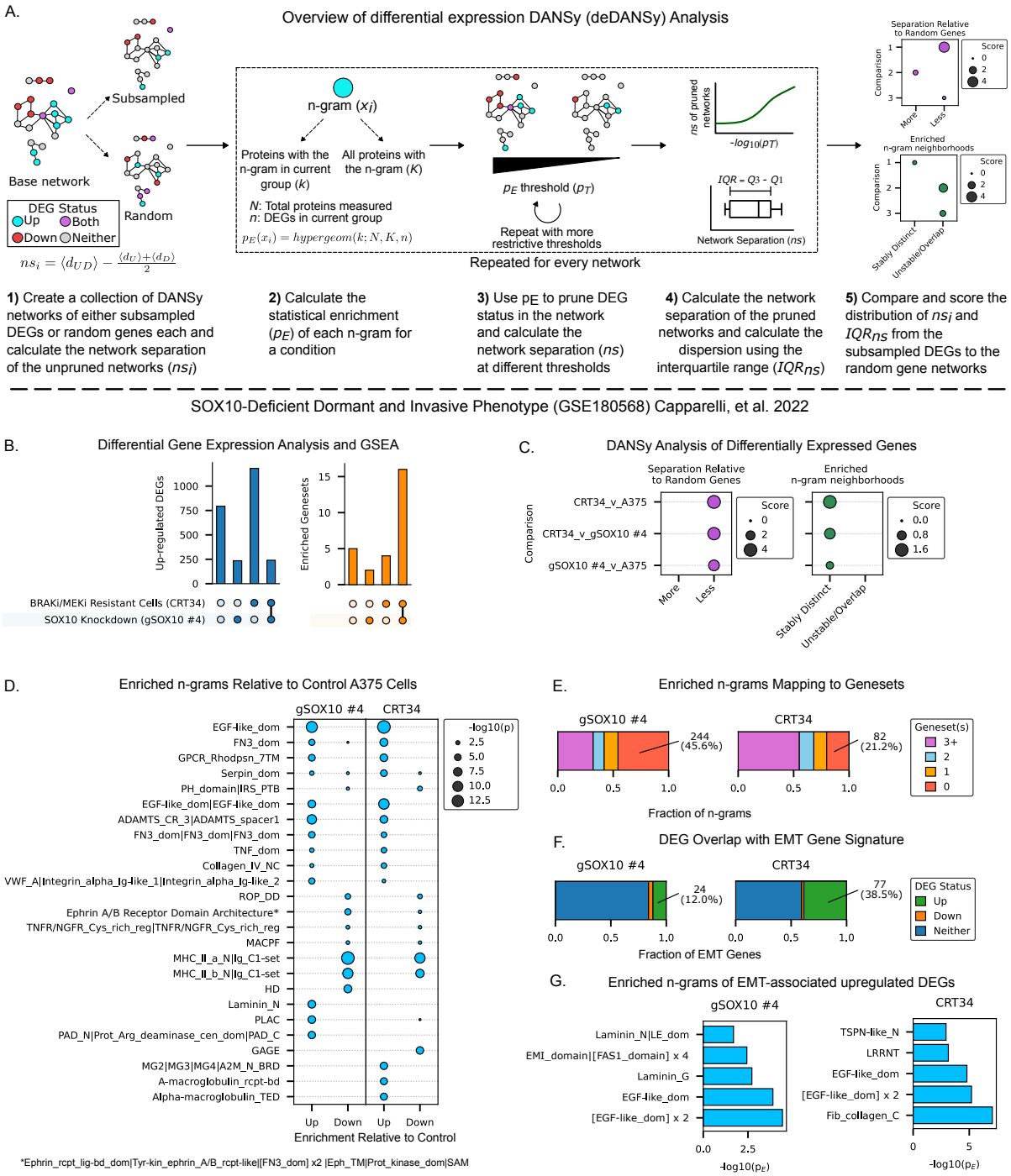

Figure 4. Legend continued on the next page.

Figure 4. Application of DANSy to differential expression analysis. The application of DANSy analysis to differential expression performed through a retrospective analysis of RNA-seq results of SOX10-deficient melanoma cells that exhibit dormant and invasive phenotypes [45] (GEO accession: GSE180568). A) Schematic overview of the application of DANSy to differential expression (deDANSy) which consists of 5 major steps to find how robustly separated up and down regulated genes are on a DANSy network. Subsampled DEGs and random genes are used to create a collection of DANSy networks and the separation between up or down DEGs calculated in each unpruned network (ns_i). Then, the statistical enrichment (p_E) of n-grams and the differentially expressed gene (DEG) status of individual proteins are used to find robust enriched subnetworks by calculating the dispersion of network separation values (IQR_{ns}) when DEG status is pruned using p_E. The distributions of the ns_i and IQR_ns generated from subsampled DEGs and randomly chosen genes are compared using the Mann-Whitney U test and Cohen's d effect size to generate final scores and determine statistical significance. See Methods for a complete description. B) UpSet plots showing the number of unique or shared up-regulated DEGs (left) or enriched genesets from GSEA(right) of cell lines either generated from CRISPR-mediated SOX10 knockdown (gRNA) or in vivo establishment of therapeutic resistant cells (CRT). The first bar represents DEGs/genesets unique to additional gRNA/CRT cell lines of the study. C) The deDANSy analysis which shows how separated and distinct the n-grams from DEGs of gSOX10 #4 and CRT34 cells are relative to their parental A375 cells. D) Example individual n-grams from each of the cell lines from the top 10% significantly enriched n-grams. E) The fraction of enriched n-grams that can be mapped to the Hallmark genesets. The number and percentage of all the enriched n-gram that map to no gene set are given to the left of each plot. F) The relative fraction of genes found in the Hallmarks epithelial to mesenchymal transition (EMT) gene set that are also DEGs for each cell line. G) The enrichment of individual n-grams for the up-regulated DEGs that are within the EMT gene set.

enriched language subnetworks for the up and down regulated genes were significantly distinct between most comparisons. Unlike the network separation scores, where ≥90% of scores fell under less separation than random expectation, the stable enrichment scores were more spread across both the stably distinct and the unstable/overlapping categories (Fig. S17). Overall, these suggest that our deDANSy analysis describes the coordinated functions underlying phenotypes, but further identifies whether these differences rely on the enrichment of distinct subnetworks of functionality.

We next sought to compare how our analysis relates to gene set enrichment analysis (GSEA), given its prevalence to identifying the coordinated enrichment of function in RNA-seq datasets. We focused our analysis on the study that described how melanoma cells with low SOX10 expression exhibit a dormant and invasive phenotype, which were tolerant to BRAF and MEK inhibitors (BRAFi+MEKi)[45]. Importantly, this study generated SOX10 deficient cells through either a CRISPR SOX10 knockout (gSOX10) or with an in vivo selection of BRAFi+MEKi combination resistant/tolerant cells (CRT). We defined up-regulated DEGs (log2 fold-change ≥2 and p-adjusted value ≤0.001) for each gSOX10 and CRT cell lines relative to the parental A375 line (2 for each method), which showed about 200 genes shared between all cell lines and 200-1000 unique to each cell line (Fig. 4B). In agreement with the original study, we found 16 statistically enriched Hallmark genesets[51] shared between all the cell lines (Fig. 4B, S18)[45]. Seeing this discrepancy in the fraction of up-regulated genes and enriched genesets shared between cell lines, we asked whether deDANSy can expand GSEA findings by identifying whether the cell lines are utilizing similar functional domain language. Our deDANSy analysis showed the gSOX10 and CRT cell lines had consistently significant and enriched distinct language subnetworks relative to the control line. Interestingly, deDANSy found that the gSOX10 #4 and the CRT34 lines have different distinctly stable enriched subnetworks (Fig. 4C). Meanwhile, the second CRT and SOX10 cell lines did not consistently show significant separation and distinct subnetworks (Fig. S17A). Thus, we focused on comparing the gSOX10 #4 and CRT34 cell lines to determine if both cell lines are using similar domain language to establish the shared gene set phenotypes.

To understand what enriched domains comprise these language subnetworks, we focused on the enriched n-grams of each condition relative to the A375 control cell line. Many of the significantly enriched domains such as the EGF-like or the GPCR_Rhodpsn_7TM domains were shared between

both conditions and had similar enrichment p-values. However, out of the were 300-500 significantly enriched n-grams in the up and down-regulated DEGs, only a small fraction (<15%) were shared between the cell lines (Fig. S19). Domains such as the Laminin N-terminal (Laminin_N), GAGE, PLAC, and macroglobulin (e.g. MG2, MG3) domains, which are uniquely enriched in individual cell lines, were amongst the top 10% enriched n-grams (Fig. 4D). We next sought to understand if using n-grams could reduce unintentional study bias when using the Hallmark genesets by analyzing the enriched n-grams of these genesets. We find most genesets have key groups of enriched n-grams that are frequently shared amongst related pathways, but most n-grams map to multiple genesets (Fig S20). Additionally, 45.6% of the gSOX10 #4 and 21.2% of the CRT34 enriched n-grams, such as the GAGE and PLAC domains, were not found in any of the enriched Hallmark genesets (Fig. 4E,S21,S22). However, this analysis revealed not all enriched genesets have a reciprocal enrichment of n-grams (Fig. S21,S22) suggesting that GSEA and deDANSy together can describe the full scale of coordinated changes in signaling driving phenotypic changes.

Since deDANSy identifies coordinated functions, like GSEA, but with complementary information, we evaluated whether n-gram enrichment can determine if the gSOX10 and CRT cells use similar functionality to enrich for a shared phenotype. Here, we focused on the epithelial to mesenchymal transition (EMT) gene set as it was the most significantly up-regulated gene set with a normalized enrichment score >2.1 and FDR q-value ≤0.0001 for both cell lines. Further, EMT is an important process in the metastatic cascade and the invasion of cancer cells into the surrounding tissue[52]. In both cell lines <40% of the genes that make up the EMT gene set were considered a DEG with only 24 genes being an up-regulated DEG in the gSOX10 cells (Fig. 4F). We analyzed the enriched n-grams of these EMT-associated DEGs and found that only the n-grams that contain the EGF-like domain were commonly enriched in both cell lines (Fig. 4G). Interestingly, in both cell lines the enriched n-grams are those associated with extracellular matrix (ECM) proteins (Fig. 4G). Specifically, the CRISPR-mediate gSOX10 cells were enriched for domains associated with the laminin or fasciclin family of proteins. Meanwhile, the drug-induced CRT34 cells were enriched for the C terminal domain of fibrillar collagens (Fib_collagen_C) and the thrombospondin-like, N-terminal domain (TPSN-like_N). Thus, despite both cell lines secreting ECM, deDANSy highlights each are using different ECM components to achieve similar EMT phenotypes. Collectively, these results demonstrate how deDANSy analysis can identify the coordinated changes associated with differential expression, and it can determine whether common phenotypes are generated with similar or different molecular effectors, which are difficult to parse with GSEA.

Discussion

Since protein domains are modular units of structure and function that follow grammar-like rules, we set out to leverage these properties to abstract the biochemical connections underlying protein functionality. We established DANSy where a network of n-grams from protein domain architectures is analyzed for either the whole proteome or within individual signaling subnetworks. Certain highly abundant domains, like the protein kinase and Znf-C2H2 domains, act as hubs in the n-gram network to bridge n-gram families of distinct functions reflecting the diversity of biological processes that involve those domain functions. Importantly, DANSy only relies on protein domain architectures to capture the language of a subset of the proteome, not protein-protein interactions or known involvement in a biological process. Therefore, DANSy can identify domains without clearly defined functions, like the GAGE domain in our application of DANSy to differential gene expression as discussed below, but critical to the process or signaling pathway being studied.

When we applied DANSy to cancer fusion genes, we found most predicted fusion gene domain architectures follow the established grammatical rules of the proteome. While our analysis of patient survival was limited by the rarity of gene fusions relative to somatic mutations[38–40], we can use DANSy to study how individual domain n-grams contribute and use inherent properties of the domains that enable new functionality to develop through gene fusions. We showcased this with kinase fusions, which are one of the largest classes of gene fusions[39, 41], to suggest the kinase domain has an intrinsic

property that allow neofunctionalization. Since the n-gram network properties describe the rules governing the flexibility of an n-gram to gain new domain combinations, we can use DANSy to understand additional types of fusions genes. For example, ETS fusions are highly prevalent in prostate cancer[35, 53], and the Pointed|ETS domain architecture was one of the most common domain n-grams in our analyzed fusions. However, the Pointed|ETS fusion rarely generated novel domain architectures (Fig. S14C) nor impacted n-gram network topology. Since ETS fusions poorly predict clinical outcomes[37, 54], this suggests that the ETS fusions can use existing functions and pathways to create aberrant signaling, but these changes are not advantageous to promote disease progression. This would agree with recent reports suggesting some gene fusions are not necessarily putative drivers of disease[55], but potentially passenger aberrations[56]. Thus, DANSy can contextualize the potential of a gene fusion to promote disease progression, but with recent reports of chimeric mRNA species that generate fusion genes in disease-free tissues[57, 58], the grammar to determine how deleterious a fusion domain architecture is can be further expanded.

Our DANSy analysis uncovered a set of grammatical rules that each reversible PTM system converged toward through evolution. As the pTyr system rapidly expanded, species were sampling several configurations of the rules to determine feasible domain ordering. During this sampling, domains that act as "verbs" separated from one another, and the eraser domains loosely connected, if at all, with the rest of the system. This suggests that the n-gram network properties can potentially measure if a system is converging to a set of grammatical rules or continuing to rapidly test new domain combinations. Since domains can encode the evolutionary jumps of protein families[5], we can use them to find other potential systems that could be testing new configurations. For example, the KRAB domain recently evolved during the emergence of vertebrates[59] to counter the expansion of transposable elements within mammalian genomes[60]. In the n-gram network, the KRAB domain had a grammatical context where it could bridge multiple n-gram families, but had a preferred domain n-gram in the natural proteome. However, the KRAB domain can generate several novel domain architectures within fusion genes. Whether this is an intrinsic property of the KRAB domain for neofunctionalization or part of the evolutionary process of exploring the grammatical constraints of the transposable element regulatory system remains unclear, but can be explored with DANSy.

Our analysis has generated a computational framework for describing domain architectures using both linguistics and network approaches. However, our analysis still is limited in describing the complete domain landscape present within cells. We retrieved and analyzed only the canonical protein isoforms, which can omit proteoforms of individual genes that arise due to alternative splicing that impact protein-protein and domain-domain interactions[61]. However, alternative splicing infrequently impacts domain architectures but when reported involves repetitive domains such as the Znf-C2H2 and Ig-like domains[62]. Additionally, the continued improvement of protein structure algorithms like AlphaFold[16] have led to an expansion of the predicted structural folds without known functions but still represent the evolution of protein families[28, 63]. Altogether, these suggest the proteome continues to evolve through various mechanisms, but DANSy can integrate these domain architectures to characterize an expanded proteome.

Through our analysis of differential expression via our deDANSy analysis, we demonstrate how a DANSy-based analysis can expand and complement the insights generated from standard GSEA workflows. The base deDANSy analysis provides snapshots of how expansive and separated the functional domain subnetworks are between pairs of conditions. Further, it identifies whether highly coordinated and distinct functional subnetworks exist that delineate two conditions analogous to GSEA. However, GSEA relies on using predefined gene sets, which can miss important but enriched functional units. For example, the GAGE and PLAC domains were found to be significantly enriched in our analysis, but proteins which contain these domains did not map to any of the Hallmark genesets. Further, the gene set definitions may encompass a wide variety of potential means, but can easily collapse the subtle mechanistic differences underlying the separate methodologies (e.g. gene knockout or therapeutic-induced) that generate the phenotype of interest. Thus, GSEA alone may miss possible critical or novel effectors of the signaling pathways due to incomplete coverage of the proteome from the chosen genesets. Meanwhile, our deDANSy relies only on whether a protein contains domains, which are

found in >75% of the human proteome and represent a more complete definition of the protein interaction networks. However, the domain-less proteins are one of the largest limitations of our DANSy analysis. If a protein does not contain a domain, the (de)DANSy analysis will miss insights related to its function which may be covered by methods such as GSEA. Thus, we caution on utilizing only deDANSy to identify coordinated changes in signaling pathways, but rather to integrate deDANSy results with methods such as GSEA to characterize cell phenotypes. Nevertheless, we have developed the DANSy framework and demonstrated the types of insights DANSy analysis provides by abstracting protein interaction networks while more fully describing the proteome without relying on the challenges in annotating all possible protein/gene interactions. We envision that DANSy will provide new perspectives on the coordination of signaling pathways that regulate various biological processes.

Materials and Methods

Retrieving UniProt IDs for the human proteome and post-translational modification systems for additional species. For the human proteome, UniProtKB IDs were retrieved using pybiomart to map between Ensembl gene IDs and reviewed UniProtKB IDs. For individual post-translational modification (PTM) systems, InterPro IDs for the different domains that make up the PTM system were retrieved using the IDs listed in Table S1 using the InterPro module of CoDIAC[29]. For analyzing the phosphorylation systems in all species except humans, mice (*Mus musculus*) and rats (*Rattus norvegicus*) both reviewed and unreviewed UniProt records were fetched. To ensure unreviewed UniProtKB records are not mapping to same gene within individual species, UniProt entries on the evidence level, cross references to Ensembl[64], the NCBI Gene[65], and RefSeq[66] databases, protein length, and gene symbol were fetched to determine which record represented identified genes and full protein coding sequences. Reviewed records were given top priority followed by records with cross references to other databases. Records which mapped to the same cross reference IDs were then compared by protein existence levels and finally amino acid length sequence. For UniProtKB records which required comparing amino acid length sequences, the record with the longest length was retained. The complete list of species and strain names used during fetching is provided in Table S2.

Fetching InterPro protein domain architectures and generating domain n-grams. To fetch protein domain architectures from InterPro, we utilized the UniProt module from our recently developed python package CoDIAC[29] by inputting fetched UniProtKB IDs into the UniProt module from CoDIAC. The resulting reference file contains all InterPro and UniProt domain architectures and reference sequences for each queried protein and was used for downstream analysis. All results were fetched using the 2025 January 7th build of the InterPro and UniProt databases. The retrieved protein domain architectures were then separated into n-grams of the length of interest for each n-gram model. For building the PTM System Domain Focused n-gram models, only n-grams that contained domains of interest were fetched from the complete protein domain architecture.

Measuring n-gram model information gain and relative entropy. N-gram models rely on the Markov assumption where the probability of a specific n-gram depends on the conditional probability of the next domain, d_n, given the preceding sequence of domains, $d_{n-(N-1):n-1}$. This can be estimated for each n-gram using the maximum likelihood estimate (MLE):

$$p_{MLE}(d_n|d_{n-N+1:n-1}) = \frac{C(d_{n+N-1:n-1}, d_n)}{C(d_{n+N-1:n-1})}$$

Where N represents the maximum length of n-grams being evaluated (i.e. N=2 for bigrams or N=5 for 5-grams). The counts of a specific n-gram $(d_{n+N-1:n-1}, d_n)$ is represented by $C(d_{n+N-1:n-1}, d_n)$. These probabilities are then used to calculate the entropy of an n-gram model $(H_n(x))$ using Shannon's entropy:

$$H_n(x) = -\sum p(x) \log_2 p(x)$$

For individual domains the probabilities are given by the relative frequencies for each domain within the corpus of domain architectures. For longer n-grams, the entropy represents the sum of weighted probabilities which can be estimated by:

$$\begin{aligned} H_n(x) &= -\frac{1}{N_{ng}} \sum C(d_{n+N-1:n-1}) \log_2 \frac{C(d_{n+N-1:n-1}, d_n)}{C(d_{n+N-1:n-1})} \\ &= -\frac{1}{N_{ng}} \sum C(d_{n+N-1:n-1}) \log_2 p_{MLE}(d_n | d_{n+N-1:n-1}) \end{aligned}$$

Where N_{ng} represents the total number of n-grams with length N. The entropy of an n-gram model is then used to determine the relative information gain $I(x)$ from the unigram (only single domain frequency distributions) model ($H_1(x)$):

$$I(x) = H_n(x) - H_1(x)$$

To compare how the distributions of n-grams change using different n-gram models we use the relative entropy also known as the Kullback-Leibler divergence defined as:

$$D_{KL}(P||Q) = \sum P(X) \log_2 \frac{P(X)}{Q(X)}$$

Where $P(X)$ and $Q(X)$ represent the probability distributions within both n-gram models and $Q(X)$ is the baseline model that contains all n-grams within $P(X)$.

Calculating network centrality measurements. For nodes within the largest connected component, the betweenness and degree centrality measurements were calculated as implemented in the networkx python package[67]. Degree centrality is defined as the total fraction of all nodes connected to node v. The betweenness centrality measurement of node v is the sum of fractions of pairwise shortest paths that pass through node v, which is defined as:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

Where V is the set of nodes, $\sigma(s,t)$ represents the number of shortest paths between nodes s and t, while $\sigma(s,t|v)$ represents the number of shortest paths that go through node v.

Predicting domain architectures for gene fusions. Genomic breakpoints for gene fusions were retrieved from the ChimerDB[35] for the TCGA cohort. Fusions that were mapped to the current human genome build (hg38), predicted to be in-frame, and were designated as found within the ChimerSeq+ dataset representing high confidence fusions were selected for further analysis (5579 total fusions). Genomic breakpoints were then retrieved for each parent gene and mapped to exon and the protein coding sequence positions. The base pair position was then translated to an amino acid position and used to determine which domains were donated towards the fusion gene (Fig. S10). Domains which were truncated by the breakpoint location were not included in the final predicted domain architecture.

Calculating Gini-Simpson Diversity Index for domain n-grams. For n-grams with a length of 10 or less, for each domain n-gram of interest d_n all n+1 n-grams containing the n-gram were retrieved. The set of domain n-grams were then split into n-grams where d_n started or ended the n-gram. For the n-grams starting with d_n were used to calculate the diversity index for following n-grams, while those ended were used for the diversity index of preceding n-grams. The diversity index was calculated using the Gini-Simpson Diversity Index as defined for small datasets:

$$D = 1 - l = 1 - \frac{\sum_{i=1}^{R} n_i(n_i - 1)}{N(N-1)}$$

Where R is the collection of n+1 domain n-grams, N is the total count of the n-grams in the set, and n_i is the count for each individual domain n-gram.

Species N-gram Similarity Index For comparing the n-grams of the pTyr and pSer/Thr systems between individual species, the Jaccard similarity index, J, was calculated using:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where A and B represent the sets of domain n-grams found within individual species.

Network Separation Metrics To determine how separated pairs of gene sets or datasets are along a reference network, the network separation metric was adapted from Menche et al. [44], which compares the mean shortest distance between nodes within their respective networks, to the mean shortest distance of nodes between the networks.

$$ns_{AB} = \langle d_{AB} \rangle - \frac{\langle d_A \rangle + \langle d_B \rangle}{2}$$

Where $\langle d_A \rangle$ refers to the mean shortest distance of nodes associated within a single gene set A and $\langle d_{AB} \rangle$ refers to the mean shortest distance of nodes associated with gene set A to nodes associated with gene set B.

Within the domain n-gram networks, domain families may be represented by isolates or individual connected components which do not connect to the rest of the network. For distances within networks representing single gene sets (e.g. $\langle d_A \rangle$), nodes which do not connect to any other node have a distance of 0. However, for distances between gene sets these nodes are given a penalty term based on the diameter of the connected component (CC) they represent. So, the distance for single node between networks A and B is computed as:

$$d_{node} = \begin{cases} 0 & \text{shared node} \\ \text{distance to closest node in network B} & \text{node is in a shared CC} \\ 1 & \text{isolate within only 1 network} \\ \text{diameter of CC} + 1 & \text{node is in a CC not shared between networks} \end{cases}$$

TCGA Patient Survival Analysis Clinical attributes for the TCGA study cohorts were retrieved from cBioPortal[68]. For univariate patient survival analysis, patients were stratified on if a tumor harbored one or multiple of the fusion genes within our analyzed fusion gene dataset. Patients with single fusions were then further stratified based on the n-gram network impacts caused by the fusion gene (Fig. 3A). Statistically significant changes in survival probabilities were determined by using pairwise log-rank tests as implemented in the lifelines python package.

Differential Gene Expression For all datasets except for the HAS overexpression dataset[47], the raw reads of each dataset were retrieved from the Gene Expression Omnibus (GEO) using the NCBI-generated RNA-seq raw counts matrix tool for each gene expression series dataset. The raw counts were then analyzed in R using the DESeq2 package[69]. Genes with fewer than 10 raw counts across all conditions were filtered out. Each pairwise comparison was analyzed using the Wald test to calculate statistical significance as implemented in DESeq2. For the HAS overexpression dataset, the differential gene expression results were directly retrieved from the publication, which had undergone the same analysis pipeline. For each dataset the log2 fold-change ($\geq$0.585) and the adjusted p-value ($\leq$0.05) were used to designate DEGs with exact cutoffs provided in Table S3 to ensure 100-1700 DEGs per comparison for deDANSy analysis (see Supplementary Note 3 for details on DEG number limits).

Applying GSEA to the SOX10-deficient melanoma cell dataset For GSEA, genes were ranked by the Wald test statistic in each comparison. The pre-ranked list was then analyzed for enrichment of the Hallmark genesets downloaded from the Molecular Signature Database (MSigDB)[51] using GSEApy[70] using 1000 permutations and gene set size cutoffs of 15 and 500 genes.

The deDANSy algorithm and implementation All steps of the deDANSy algorithm were implemented in Python3. Key mathematical descriptions and considerations are described in Supplementary Notes 2 and 3. The algorithm can be customized through various parameters including the definition of DEGs, number of random networks, and number of trial runs to calculate the false positive rate (FPR) of the deDANSy scores. For each dataset the DEG cutoffs are provided in Table S3. There were 100 random and subsampled networks compared and 50 trials for calculating the FPR.

For each dataset, a DANSy network is established for all measured genes. The DEGs are determined and the distribution of domain architecture lengths calculated. The relative ratio of each length is then subsampled to 70%, and used to select both the subsampled DEGs and random genes. For both the ratio of up to down DEGs are retained (i.e. if it is a 30:70 split in the full DEG list, then both random and subsampled genes retain the 30:70 split) with random genes randomly designated as up or down-regulated. Within these networks each n-gram was then labeled with the status of which DEG set (i.e. up, down, both, or neither) it was associated with. The network separation (ns_i) was then calculated between subnetworks associated with either the up or down-regulated genes.

After the unpruned network separation is calculated, the enrichment of each n-gram (p_E) for each condition (i.e. up or down) is calculated using a hypergeometric distribution. The p_E is then used to prune the DEG status of an n-gram based on a threshold p-value (p_T) and a new network separation (ns_T) value calculated. This pruning and ns_T calculation is repeated with progressively smaller thresholds down to $p_T = 10^{-10}$. The dispersion of ns_T is then calculated using the interquartile range (IQR_{ns}).

Generating random and subsampled networks and calculating both ns_i and IQR_{ns} are then repeated 100 times to create distributions which are then compared for statistical significance using the Mann-Whitney U-test (p_{score}). The p_{score} then undergoes FPR correction by repeating all steps above with randomly chosen genes in the dataset to correct for false discoveries.

To score the separation and how distinct the subnetworks are we calculate Cohen's d effect size between the subsampled (x_1) and random (x_2) distribution.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Where,

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

The sign of d designates which category for either the separation (ns_i) or distinct subnetwork enrichment (IQR_ns) as outlined below:

$$sign(d_{ns}) = \begin{cases} +1 & \text{More separated} \\ -1 & \text{Less separated} \end{cases}$$

$$sign(d_{IQR}) = \begin{cases} +1 & \text{Stably distinct subnetworks} \\ -1 & \text{Unstable or overlapping subnetworks} \end{cases}$$

Standalone n-gram enrichment analysis For gene set or condition-specific analysis of n-gram enrichment, the enrichment was calculated similar to deDANSy using a hypergeometric distribution. However, the statistical enrichment of an n-gram was corrected for false discoveries using the FPR using 100 trials of randomly chosen genes without accounting for n-gram length distribution (See Supplementary Note 3).

Data availability

For RNA-sequencing source data please see corresponding GEO numbers for each dataset: SOX10-deficiency for dormant and invasive phenotype (GSE180568)[45], iBMEC tissue chips (GSE129290) [49], mechanical condition of breast cancer cells (GSE127887)[48], type III collagen and breast cancer cell dormancy (GSE182890) [50], adipocyte culturing methods (GSE115020)[46]. The HAS overexpression data was retrieved from Supplementary Table 2 in the original publication Shimpi & Tan, et al[47]. All source data for plots are provided as a Source Data File.

Code availability

All code for DANSy and deDANSy is available at our Git repository: https://github.com/NaegleLab/DANSy and all the code for the specific applications in this paper are found at https://github.com/NaegleLab/DANSy_Applications.

Acknowledgments

We would like to acknowledge the constructive and critical assessment of the presented work from the entire Naegle lab. The authors thank Bruce Mayer for helpful discussions about the phosphotyrosine system of *Monosiga brevicollis*. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM138127. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Mayer BJ (2015) The discovery of modular binding domains: building blocks of cell signalling. *Nat Rev Mol Cell Biol* 16(11):691–8.

2. Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A (2008) Arrangements in the modular evolution of proteins. *Trends Biochem Sci* 33(9):444–51.

3. Ekman D, Björklund AK, Elofsson A (2007) Quantification of the elevated rate of domain rearrangements in metazoa. *J Mol Biol* 372(5):1337–48.

4. Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J (2007) The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol* 8(4):319–30.

5. Jin J, et al. (2009) Eukaryotic protein domains as functional units of cellular evolution. *Science Signaling* 2(98).

6. Buljan M, Bateman A (2009) The evolution of protein domain families. *Biochem Soc Trans* 37(Pt 4):751–5.

7. Apic G, Huber W, Teichmann SA (2003) Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *Journal of Structural and Functional Genomics* 4(2/3):67–78.

8. Basu MK, Carmel L, Rogozin IB, Koonin EV (2008) Evolution of protein domain promiscuity in eukaryotes. *Genome Research* 18(3):449–461.

9. Basu MK, Poliakov E, Rogozin IB (2008) Domain mobility in proteins: functional and evolutionary implications. *Briefings in Bioinformatics* 10(3):205–216.

10. Yu L, et al. (2019) Grammar of protein domain architectures. *Proceedings of the National Academy of Sciences* 116(9):3636–3645.

11. Hayete B, Bienkowska JR (2005) Gotrees: predicting go associations from protein domain composition using decision trees. *Pac Symp Biocomput* pp. 127–38.

12. Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J Biol Chem* 277(48):45765–9.

13. Ofer D, Brandes N, Linial M (2021) The language of proteins: Nlp, machine learning & protein sequences. *Comput Struct Biotechnol J* 19:1750–1758.

14. Bepler T, Berger B (2021) Learning the protein language: Evolution, structure, and function. *Cell Systems* 12(6):654–669.e3.

15. Lin Z, et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379(6637):1123–1130.

16. Jumper J, et al. (2021) Highly accurate protein structure prediction with alphafold. *Nature* 596(7873):583–589.

17. Scaiewicz A, Levitt M (2015) The language of the protein universe. *Curr Opin Genet Dev* 35:50–6.

18. Errington WJ, Bruncsics B, Sarkar CA (2019) Mechanisms of noncanonical binding dynamics in multivalent protein-protein interactions. *Proc Natl Acad Sci U S A* 116(51):25659–25667.

19. Mayer BJ, Hirai H, Sakai R (1995) Evidence that sh2 domains promote processive phosphorylation by protein-tyrosine kinases. *Curr Biol* 5(3):296–305.

20. Wang M, Kurland CG, Caetano-Anollés G (2011) Reductive evolution of proteomes and protein structures. *Proc Natl Acad Sci U S A* 108(29):11954–8.

21. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA (2004) Supra-domains: evolutionary units larger than single protein domains. *J Mol Biol* 336(3):809–23.

22. Luck K, et al. (2020) A reference map of the human binary protein interactome. *Nature* 580(7803):402–408.

23. Peng X, Wang J, Peng W, Wu FX, Pan Y (2017) Protein-protein interactions: detection, reliability assessment and applications. *Brief Bioinform* 18(5):798–819.

24. Paysan-Lafosse T, et al. (2023) Interpro in 2022. *Nucleic Acids Res* 51(D1):D418–D427.

25. UniProt Consortium (2025) Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Res* 53(D1):D609–D617.

26. Mistry J, et al. (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res* 49(D1):D412–D419.

27. Letunic I, Khedkar S, Bork P (2021) Smart: recent updates, new developments and status in 2020. *Nucleic Acids Res* 49(D1):D458–D460.

28. Waman VP, et al. (2024) Cath 2024: Cath-alphaflow doubles the number of structures in cath and reveals nearly 200 new folds. *J Mol Biol* 436(17):168551.

29. Kandoor A, et al. (2024) Codiac: A comprehensive approach for interaction analysis reveals novel insights into sh2 domain function and regulation. *bioRxiv* p. 2024.07.18.604100.

30. Needham EJ, Parker BL, Burykin T, James DE, Humphrey SJ (2019) Illuminating the dark phosphoproteome. *Science Signaling* 12(565).

31. Hunter T (2012) Why nature chose phosphate to modify proteins. *Philos Trans R Soc Lond B Biol Sci* 367(1602):2513–6.

32. Pincus D, Letunic I, Bork P, Lim WA (2008) Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proc Natl Acad Sci U S A* 105(28):9680–4.

33. Lim WA, Pawson T (2010) Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* 142(5):661–7.

34. Machida K, et al. (2007) High-throughput phosphotyrosine profiling using sh2 domains. *Mol Cell* 26(6):899–915.

35. Jang YE, et al. (2020) Chimerdb 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Res* 48(D1):D817–D824.

36. Costa R, et al. (2016) Fgfr3-tacc3 fusion in solid tumors: mini review. *Oncotarget* 7(34):55924–55938.

37. Pettersson A, et al. (2012) The tmprss2:erg rearrangement, erg expression, and prostate cancer outcomes: a cohort study and meta-analysis. *Cancer Epidemiol Biomarkers Prev* 21(9):1497–509.

38. Mitelman F, Johansson B, Mertens F (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat Genet* 36(4):331–4.

39. Latysheva NS, Babu MM (2016) Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res* 44(10):4487–503.

40. Forbes SA, et al. (2017) Cosmic: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 45(D1):D777–D783.

41. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C (2014) The landscape of kinase fusions in cancer. *Nat Commun* 5:4846.

42. Frenkel-Morgenstern M, Valencia A (2012) Novel domain combinations in proteins encoded by chimeric transcripts. *Bioinformatics* 28(12):i67–74.

43. Ortiz de Mendíbil I, Vizmanos JL, Novo FJ (2009) Signatures of selection in fusion transcripts resulting from chromosomal translocations in human cancer. *PLoS One* 4(3):e4805.

44. Menche J, et al. (2015) Disease networks. uncovering disease-disease relationships through the incomplete interactome. *Science* 347(6224):1257601.

45. Capparelli C, et al. (2022) Targeting sox10-deficient cells to reduce the dormant-invasive phenotype state in melanoma. *Nat Commun* 13(1):1381.

46. Harms MJ, et al. (2019) Mature human white adipocytes cultured under membranes maintain identity, function, and can transdifferentiate into brown-like adipocytes. *Cell Rep* 27(1):213–225.e5.

47. Shimpi AA, et al. (2023) Convergent approaches to delineate the metabolic regulation of tumor invasion by hyaluronic acid biosynthesis. *Adv Healthc Mater* 12(14):e2202224.

48. Watson AW, et al. (2021) Breast tumor stiffness instructs bone metastasis via maintenance of mechanical conditioning. *Cell Rep* 35(13):109293.

49. Vatine GD, et al. (2019) Human ipsc-derived blood-brain barrier chips enable disease modeling and personalized medicine applications. *Cell Stem Cell* 24(6):995–1005.e6.

50. Di Martino JS, et al. (2022) A tumor-derived type iii collagen-rich ecm niche regulates tumor cell dormancy. *Nat Cancer* 3(1):90–107.

51. Liberzon A, et al. (2015) The molecular signatures database (msigdb) hallmark gene set collection. *Cell Syst* 1(6):417–425.

52. Nieto MA, Huang RYJ, Jackson RA, Thiery JP (2016) Emt: 2016. *Cell* 166(1):21–45.

53. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM (2008) Recurrent gene fusions in prostate cancer. *Nat Rev Cancer* 8(7):497–511.

54. Esgueva R, et al. (2010) Prevalence of tmprss2-erg and slc45a3-erg gene fusions in a large prostatectomy cohort. *Mod Pathol* 23(4):539–46.

55. Mertens F, Johansson B, Fioretos T, Mitelman F (2015) The emerging complexity of gene fusions in cancer. *Nat Rev Cancer* 15(6):371–81.

56. Kalyana-Sundaram S, et al. (2012) Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer. *Neoplasia* 14(8):702–8.

57. Singh S, et al. (2020) The landscape of chimeric rnas in non-diseased tissues and cells. *Nucleic Acids Res* 48(4):1764–1778.

58. Babiceanu M, et al. (2016) Recurrent chimeric fusion rnas in non-cancer tissues and cells. *Nucleic Acids Res* 44(6):2859–72.

59. Imbeault M, Helleboid PY, Trono D (2017) Krab zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 543(7646):550–554.

60. Rosspopoff O, Trono D (2024) Take a walk on the krab side: (trends in genetics, 39:11 p:844-857, 2023). *Trends Genet* 40(2):203–205.

61. Yang X, et al. (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164(4):805–17.

62. Light S, Elofsson A (2013) The impact of splicing on protein domain architecture. *Curr Opin Struct Biol* 23(3):451–8.

63. Schaeffer RD, et al. (2023) Classification of domains in predicted structures of the human proteome. *Proc Natl Acad Sci U S A* 120(12):e2214069120.

64. Harrison PW, et al. (2024) Ensembl 2024. *Nucleic Acids Res* 52(D1):D891–D899.

65. Brown GR, et al. (2015) Gene: a gene-centered information resource at ncbi. *Nucleic Acids Res* 43(Database issue):D36–42.

66. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic Acids Res* 40(Database issue):D130–5.

67. Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using networkx in *Proceedings of the 7th Python in Science Conference*, eds. Varoquaux G, Vaught T, Millman J. (Pasadena, CA USA), pp. 11 – 15.

68. Cerami E, et al. (2012) The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2(5):401–4.

69. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol* 15(12):550.

70. Fang Z, Liu X, Peltz G (2023) Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics* 39(1):btac757.