mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome

Michael J. Meyer[1,2,3,†], Ryan Lapcevic[1,2,†], Alfonso E. Romero[4,†], Mark Yoon[1,2], Jishnu Das[1,2], Juan Felipe Beltrán[1,2], Matthew Mort[5], Peter D. Stenson[5], David N. Cooper[5], Alberto Paccanaro[4], and Haiyuan Yu[1,2,*]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, 14853, USA

[2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York, 14853, USA

[3]Tri-Institutional Training Program in Computational Biology and Medicine, New York, New York, 10065, USA

[4]Department of Computer Science and Centre for Systems and Synthetic Biology, Royal Holloway, University of London, Egham TW20 0EX, UK

[5]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN UK

[†]The authors wish it to be known that, in their opinion, the first 3 authors should be regarded as joint First Authors

[*]To whom correspondence should be addressed. Tel: 607-255-0259; Fax: 607-255-5961; Email: haiyuan.yu@cornell.edu

Abstract

A new algorithm and web server, mutation3D (http://mutation3d.org), proposes driver genes in cancer by identifying clusters of amino acid substitutions within tertiary protein structures. We demonstrate the feasibility of using a 3D clustering approach to implicate proteins in cancer based on explorations of single proteins using the mutation3D web interface. On a large scale, we show that clustering with mutation3D is able to separate functional from non-functional mutations by analyzing a combination of 8,869 known inherited disease mutations and 2,004 SNPs overlaid together upon the same sets of crystal structures and homology models. Further, we present a systematic analysis of whole-genome and whole-exome cancer datasets to demonstrate that mutation3D identifies many known cancer genes as well as previously underexplored target genes. The mutation3D web interface allows users to analyze their own mutation data in a variety of popular formats and provides seamless access to explore mutation clusters derived from over 975,000 somatic mutations reported by 6,811 cancer sequencing studies. The mutation3D web interface is freely available with all major browsers supported.

Keywords: cancer, clustering, protein structures, somatic mutations

Grant Sponsors: TODO

MM, PDS and DNC acknowledgement the financial support of Qiagen Inc through a License Agreement with Cardiff University.

Introduction

A hallmark of the genomic era has been the application of whole-genome and whole-exome sequencing to the study of genetic disease, especially cancer. This effort has led to the development of new statistical methods (Hodis, et al., 2012; Lawrence, et al., 2013; Sjöblom, et al., 2006), which have identified many potential genomic targets of interest by combing the deluge of data produced by large cohort studies. While these methods have been largely successful in identifying genes with previously unknown roles in tumorigenesis, we have yet to fully realize the promised boon to development of therapies—although the list of potential disease-causing and driver mutations has grown, the list of approved therapeutics has remained static (Das, et al., 2014).

Although the underlying causes of this time lag are complex, they can at least be partially attributed to the level of resolution of current methods, which typically identify potentially functional genes based on mutation frequencies at the level of whole genes (Cancer Genome Atlas, 2012; Lawrence, et al., 2014; Vucic, et al., 2012; Wood, et al., 2007). However, many genes carry out a diverse set of functions (pleiotropy), the derangement of any one of which may be sufficient to cause cancer. Further, disruption of different functions of the same gene often lead to clinically distinct types of cancer (Hanahan and Weinberg, 2011; Muller and Vousden, 2013). Finally, even when a specific gene has been identified as being potentially involved in tumorigenesis, researchers may have little idea as to which of its functions has been disrupted.

All of these challenges facing current methodologies make it difficult to develop targeted therapeutic strategies.

Here we present mutation3D, an algorithm and web server (http://mutation3d.org) designed to identify somatic cancer-causing genes by leveraging the structure-function relationships inherent in their protein products. In tumorigenesis, mutations are selected that confer a competitive advantage to pre-cancerous cells. Since many mechanisms of tumorigenesis involve alterations to protein function, and protein function is determined by protein structure, tumorigenically selected driver mutations may localize to positions that will affect protein structures. Therefore, mutations causing the same cancer type in a cohort of patients may form clusters (or hotspots) in regions of protein structures wherein alterations confer a competitive advantage to tumor cells by disrupting specific protein functions. For instance, mutations localized at interaction interfaces may disrupt protein complexes or transient interactions, and mutations localized in the hydrophobic core may destabilize the protein entirely (Kucukkal, et al., 2015; Nishi, et al., 2013; Petukh, et al., 2015).

Recent studies have begun to leverage structure-function relationships in proteins to predict cancer gene targets by searching for nonrandom distributions of mutations in protein crystal structures (Kamburov, et al., 2015) and enrichment across protein domains (Miller, et al., 2015). We present the first tool to identify and visualize individual clusters within protein structures. Furthermore, we also provide an option to search for clusters in homology models, expanding our coverage of the human proteome more than three-fold (Supp. Note S1). Through an intuitive, freely available web interface, researchers can use mutation3D to inspect clusters of amino acid substitutions in an interactive molecular viewer to determine whether to follow up with the target based on its structural features. Furthermore, mutation3D can analyze data from whole-genome

sequencing (WGS; abbrev. also including whole-exome) studies to perform cluster analysis of variants at the level of the structural proteome.

Methods

mutation3D clustering algorithm

The algorithm underlying the mutation3D web interface is complete-linkage (CL) clustering (Sørensen, 1948), a hierarchical clustering method in which clusters first comprise single elements and are then merged with nearest neighboring clusters or unassigned elements until a single cluster comprises all elements. Notably, the clusters found by complete-linkage clustering, as opposed to single-linkage clustering (Sneath, 1957), are assured to have a diameter less than or equal to a specified linkage distance, which results in tight well-defined clusters. Because of this property, this method can also be referred to as furthest-neighbor clustering, since the dissimilarity of elements within a cluster is determined by the distance between the two elements furthest from each other in n-dimensional space.

In our implementation of this classic machine learning algorithm, we cluster the three-dimensional locations of the α-carbons of those amino acids whose codons contain missense mutations. The coordinates of all atoms within proteins were derived from both PDB structures and structural models (Pieper, et al., 2011) based on PDB entries covering proteins either in part or in full. For any given protein, many overlapping models may be available from either or both sources. mutation3D will invariably use entries from the PDB when they are available, as these experimentally determined crystal structures are considered to be a 'gold standard' in structural biology. To increase structural coverage of the proteome, the user may also select a subset of homology-based models to include, based upon several quality metrics available via the Advanced Query page (Supp. Note S2). Once a set of PDB structures and structural models has been established for a single protein, mutation3D attempts to cluster amino acid substitutions on all models separately, and reports any model or experimentally determined structure in which a

cluster has been found. In our analyses, we consider it sufficient to implicate a protein in cancer if any of its models are found to contain a cluster.

Some whole proteins or regions of proteins may not have been crystallized or modeled to-date. Owing to the lack of structural coordinates in these regions, we would be unable to identify clusters of mutations. There are some cases in which a single genomic mutation may give rise to defects to distinct proteins, in which case mutation3D will attempt to find clusters across all proteins and models for which this mutation has an effect on protein products.

Users may elect to set the CL-distance, or the maximum allowable distance between α-carbons in a cluster of substituted amino acids. We refer to this as the *maximum cluster diameter* as this is equivalent to the maximum allowable diameter in Angstroms of a sphere encapsulating all α-carbons in a cluster. With regard to the complete linkage clustering algorithm, the CL-distance is the maximal dissimilarity between elements, after which no new merging of elements and groups of elements occurs. In mutation3D, we call this parameter the *Maximum Clustering Diameter*, which is measured in Angstroms, and represents the maximum distance between amino acid substitutions after which no further merging of single mutations with clusters occurs and clusters are assigned based on current hierarchical groupings of mutations. For more information on all algorithm parameters and their default values, see Supp. Notes S2 and S3.

Statistical significance of clusters

In order to calculate the statistical significance of clusters found by complete-linkage clustering, mutation3D performs an iterative bootstrapping method to calculate a background distribution of cluster sizes arising from a random placement of an equivalent number of substitutions in a given protein structure. By default, mutation3D will randomly rearrange all amino acid substitutions

15,000 times in a given structure and calculate the minimum CL-distance at which a cluster of size *n* (where *n* is all cluster sizes found in the original data) is observed in the randomized data. For each cluster in the original data, *P*-values are computed empirically as the percentile rank of its CL-distance among all CL-distances for randomized clusters containing the same number of amino acid substitutions. The clustering algorithm/statistical significance calculator is implemented in C++ and is available for download as a command-line tool.

There is precedent, even within cancer gene detection, for the use of iterative bootstrapping methods when the background distributions are unclear or complicated (Hodis, et al., 2012; Lawrence, et al., 2014). Here we use bootstrapping to account for vastly different configurations of the protein backbone in different protein structures.

Compiling a protein structure and model set

In order to build a repository of protein structures and models, we curated experimentally-determined crystal structures from the PDB and homology models from ModBase by searching for canonical isoforms of Swiss-Prot structures or chains in both. Since many PDB structures provide too little coverage of their target proteins to be useful for clustering, we retained only those structures that cover at least 250 amino acids or 40% of their target protein. We only retained ModBase models that have an MPQS score ≥ 0.5, and maintain a default cutoff of MPQS ≥ 1.1 in the mutation3D interface and in our analyses. All structures and models were compared against each other to remove redundancies (i.e. a ModBase model that is of higher quality than, and whose range of amino acids is entirely contained within, a second ModBase model derived from the same PDB structure was considered not to add any novel structural information to our repository). Furthermore, the amino acid indices of all models and structures

were realigned using SIFTS (Velankar, et al., 2013) to match the amino acid indices of the Swiss-Prot protein they represent.

mutation3D web interface

To build the mutation3D web interface, we leveraged the power and flexibility of several well known JavaScript packages, such as JQuery and Bootstrap, in addition to a package designed to draw static two-dimensional figures (KineticJS). The cornerstone of our display system is an entirely JavaScript-based molecular viewer, GLmol, which allows users to view interactive 3D protein structures natively in modern web browsers supporting the new WebGL standard, without downloading any additional software. We have made modifications to these software packages to allow triggering of events by the user, such as highlighting mutations and mutation clusters simultaneously in the 3D and 2D representations of proteins.

To speed up web accession for both single and batch queries, mutation3D runs on a multi-core web server and the calculation of clusters is distributed among available computing cores using multithreaded CGI programs.

Compiling mutations and variants affecting aromatase

We compiled a list of all inherited missense mutations from the Human Gene Mutation Database (Stenson, et al., 2014) (HGMD) that (i) occurred within the exons of the *CYP19A1* gene [MIM# 107910] encoding the protein aromatase and (ii) have been shown in the primary literature to cause aromatase deficiency [MIM# 613546] (Supp. Table S1). We also compiled a set of all missense SNPs with total minor allele frequency (MAF) ≥ 1% (combined African and European ancestry) from the Exome Sequencing Project (Fu, et al., 2013) (ESP) that give rise to amino

acid substitutions in aromatase (Supp. Table S2). Please note that nucleotides are indexed in coding sequences, using the A of the ATG translation initiation start site as nucleotide 1. Visual inspection was performed by highlighting C_α positions in aromatase (PDB: 3S79) using PyMol (Schrodinger, 2010).

Segregating disease mutations from SNPs

For each Swiss-Prot protein from UniProt, a set of pathogenic inherited mutations from HGMD (Stenson, et al., 2014) was assembled for the catalogued disease with the greatest number of associated mutations in that protein. Proteins with fewer than three pathogenic mutations (two of which were required to occur at unique amino acid positions) associated with any one disease were not considered, as this is the minimum requirement for identifying a cluster with default mutation3D parameters (Supp. Notes S2 and S3). Separately, we assembled non-synonymous SNPs (nsSNPs) with MAF $\geq$ 1% from the ESP 6500 set, only retaining proteins if there were at least three SNPs in the protein, two of which caused amino acid substitutions at unique amino acid positions. We intersected these two sets and only retained proteins that occurred in both sets as meeting the individual criteria of three mutations from each set, two of which must have been at unique amino acid positions, for a total of 6 or more variants per protein. In total, we retained 8,869 inherited disease-associated mutations from HGMD and 2,004 nsSNPs from ESP 6500 in 336 proteins.

We used mutation3D to identify clusters in the resulting proteins, employing a fairly strict definition of a cluster whereby a cluster was identified if three or more substitutions were found within the complete linkage clustering distance of 15 Å, with at least two substitutions occurring at unique amino acid locations. 3D model sets were derived from PDB structures and ModBase

models indicated to be of high quality by an MPQS ≥ 1.1 (full details on default parameters for mutation3D are available in Supp. Notes S2 and S3). We report the average per-protein clustering rates across all proteins for which models from the correct set were available. *P*-values were calculated using a *U* test.

Measuring the overlap between mutation3D-implicated genes and the Cancer Gene Census

To assess how efficiently mutation3D is able to capture validated cancer genes, we ran mutation3D with default parameters (Supp. Notes S2 and S3) on all WGS screens in COSMIC v75 (285 studies). We varied the maximum cluster diameter from 5 Å to 25 Å and identified the fraction of proteins implicated (as having one or more clusters of amino acid substitutions) that are known cancer genes. We define known cancer genes to be the union of genes included in the Cancer Gene Census (Futreal, et al., 2004) and MutSig drivers list (Lawrence, et al., 2014). These overlaps were computed as the number of gene overlaps with the known cancer genes divided by the total number of genes implicated by mutation3D in each tissue category and overall (this is also known as the precision or positive predictive value (PPV)):

$$PPV = TP / (TP + FP)$$

where TP is the number of true positives and FP is the number of false positives predicted by mutation3D. It should be noted that since the our set of known cancer genes is far from complete, this estimation is likely to represent the lower bound of the true precision of our method. Furthermore, we acknowledge that even genes in the set of known cancer genes may not be drivers in all cancer types. However, the overlap between our results and the known cancer

genes is likely to correlate with the underlying precision of our method and there is no reason to believe that the overlap will be biased in certain cancer types. Therefore, this measurement can be used to estimate the lower bound of the precision of our method in comparing its performance across different cancer types. Calculation of sensitivity and specificity is inappropriate in this instance because no method could re-capitulate all known cancer genes as no data set (single WGS study or a group of WGS) can be assumed to harbor all mechanisms underlying tumorigenesis. We also computed the overlap of *all* genes in these 285 COSMIC studies with known cancer genes for each tissue category and across all tissues, to show that performing 3D clustering at any maximum cluster diameter increases precision over random expectation for this data set. *P*-values were calculated using a *Z* test to compare each fraction of identified genes by clustering at different diameter thresholds to the fraction of identified genes without clustering.

Assessing the likelihood of mutations clustered with mutation3D to be causal

In addition to predicting driver genes based on those found to contain clusters, mutation3D has the ability to predict those mutations likely to drive cancer phenotypes by their inclusion in clusters. Here, we used two proxies for causal driver mutations: that they should be more likely to be damaging and they should be more frequently observed in WGS studies.

We determined PolyPhen-2 scores (using the HumVar-trained model for assigning categories) of those mutations likely to be most deleterious biochemically based on a Grantham score (Grantham, 1974) in the top 25%. This shows how a combined biochemical and evolutionary genetics approach could lead to the discovery of new driver mutations. PolyPhen-2 scores were accessed using the Ensembl Variant Effect Predictor, assembly GRCh38.p5 (http://www.ensembl.org/Tools/VEP) (McLaren, et al., 2010).

We further determined the fraction of mutations from WGS studies found in clusters that are observed at high frequencies (in the top 2%) throughout COSMIC WGS studies.

Results

Single-protein spatial mutation case studies

The specific relationship between 3D regions of protein structure and their functions can be illustrated by the proximity of amino acid substitutions arising from known disease-causing and cancer-associated mutations in tertiary protein structures. We searched the Human Gene Mutation Database (Stenson, et al., 2014) (HGMD), a large-scale disease database of gene mutations causing human inherited disease, and the Catalogue of Somatic Mutations in Cancer (Forbes, et al., 2011) (COSMIC), a somatic cancer mutation database, for examples of spatially specific disruptions that might explain disease phenotypes. This is intended as a proof-of-principle, showing that there is a plausible connection between the spatial arrangement of mutations and disruptions of function, and that this relationship can be quickly captured through visual inspection.

Disease mutations and nsSNPs segregate in aromatase

According to HGMD, aromatase deficiency is known to be caused by at least 9 unique missense mutations in the cytochrome P450, family 19, subfamily A, polypeptide 1 (*CYP19A1*) gene leading to amino acid substitutions at 8 positions along the aromatase protein backbone (Supp. Table S1). The Exome Sequencing Project (ESP) 6500 data set (Fu, et al., 2013) contains two common non-synonymous SNPs (nsSNPs) with MAF $\geq$ 1% in this gene, which we consider

likely to be benign given their high frequency of occurrence (Supp. Table S2). Based on the primary sequence alone, no clear pattern or separation can be detected between the disease mutations and nsSNPs (Figure 1a). However, when we inspect the locations of these two classes of mutation on an experimentally-determined crystal structure of aromatase (PDB: 3S79 in Figure 1a), it is evident that the verified disease mutations and common nsSNPs are localized in quite different regions of the protein, suggesting somewhat different functional consequences depending upon the location of a mutation within the tertiary structure of the protein.

Commonly observed cancer mutations form a tight cluster in GTPase KRas

Cancer mutations may also aggregate within clusters in protein structures, and this aggregation is likely to have profound implications for our ability to differentiate functional driver mutations from neutral passenger mutations. Consider the canonical oncogenic protein GTPase KRas: the tight clustering of commonly mutated amino acid substitutions in codons 12, 13 and 61 suggests that these mutations cause similar structural perturbations that may lead to many types of cancer (Figure 1b). In fact, it has long been known that substitutions in these codons confer tumorigenesis, and several mechanisms have been proposed (Pylayeva-Gupta, et al., 2011) (Supp. Note S4, Supp. Table S3). Interestingly, another amino acid substitution E49K has only been reported once in a single patient (Guedes, et al., 2013) and is predicted to be benign by PolyPhen-2 (Adzhubei, et al., 2010). The clear spatial separation of the known driver mutations from the putatively benign mutation indicates a highly specific correlation between protein structure and function in cancer. Owing to its very high mutation frequency in many different types of cancer, *KRAS* [MIM# 190070] is readily identifiable as tumorigenic by many methods; however, mutation3D is uniquely positioned to be able to detect similar cases of spatially

specific disruption in proteins currently unknown for their roles in tumorigenesis by relating cancer sequencing data to aberrations in the structural proteome.

Coordinating mutations and structural data into a tool for whole-genome inference

mutation3D identifies mutations that group together to form statistically significant clusters on the folded protein backbone based on atomic coordinates derived from experimentally determined crystal structures and homology models. Cluster significance is measured by an iterative bootstrapping model, in which observed mutations are randomly rearranged on a protein structure, and the size of the observed cluster is ranked compared to all randomly derived clusters to compute an empirical *P*-value (see Methods for details). The accompanying web interface provides visualization of these clusters as well as the ability to rapidly switch views between all available structures. Figure 2 describes the curation of structural and mutation data, and user accession and download procedures.

Structural data underlying mutation3D

In assembling a set of protein structures and models for use with mutation3D, we relied on the huge advances made in structural proteomics over the past decade. Alongside the explosion of genomic sequencing data, the availability of structural proteomic data, including crystal structures and homology models, has increased dramatically. In 2003, there were 25,864 crystal structures in the Protein Data Bank (Berman, 2000) (PDB), covering 6.7% of the human proteome. Now, with the number of entries in the PDB exceeding 100,000, we can visualize nearly 90% (with reasonable accuracy and coverage—see Supp. Figure S1) of the human proteome through a combination of experimentally-determined crystal structures and structural

models based on shared structural elements among homologous proteins. mutation3D curates both crystal structures from the PDB and high-quality homology models from ModBase (Pieper, et al., 2011) to populate its repository of over 135,000 protein structures (Figure 2a). This significant underpinning of structural proteomic data ensures that mutation3D is useful for large-scale sequencing projects, as nearly all DNA mutations of interest within coding regions will be mappable to 3D locations in protein structures.

Seamless access to large-scale somatic cancer mutation sets

Perhaps the richest large-scale source of missense mutation data derives from WGS studies of cancer patient cohorts. According to COSMIC, in the year 2003, 187 peer-reviewed articles were published reporting on average a single gene with protein-altering somatic mutations in tumor-normal sequencing studies. In 2012, 572 studies reporting an average of 144 mutated genes were published. With the growing ease of sequencing, the scientific community has largely embraced the wholesale sequencing of tumor samples, and an accompanying class of statistical methods to identify genes characterized by elevated mutation rates across large patient cohorts (Cancer Genome Atlas, 2012; Hodis, et al., 2012; Lawrence, et al., 2014; Lawrence, et al., 2013; Sjöblom, et al., 2006; Wood, et al., 2007). These methods have been largely successful, and have led to the discovery of many genes previously not known to be involved in tumorigenesis. However, studying cancer at the level of whole genes ignores the fact that many genes and their protein products perform multiple cellular functions (pleiotropy). By incorporating available protein structures and models into cancer gene detection, we can harness the inherent structure-function relationship in proteins to identify more specific tumorigenic etiologies based on specific spatial disruptions that could become therapeutic targets.

The mutation3D web interface allows users to rapidly analyze pre-processed missense mutation data from the most recent build of COSMIC through intuitive web forms on the *Advanced* query page (http://mutation3d.org/advanced_form.shtml, click the *COSMIC* tab under *Data Source*). Currently, we have catalogued over 975,000 missense mutations in 6,811 primary cancer sequencing studies that users can search for by author, journal, PMID and size of dataset (Figure 2b). Additionally, users may choose to tune the default clustering parameters (Supp. Note S3) and protein structural model set (Supp. Note S2) based on the types of evidence needed to support clusters for their specific application. A list of candidates, with links to 3D views of the mutations overlaid onto structural models (Figure 2c), are retrieved within seconds, even for the largest WGS studies in COSMIC.

mutation3D identifies well-validated gene candidates and plausible new targets

We ran mutation3D on large sets of known inherited disease and cancer mutations to demonstrate the power of clustering to reveal shared etiologies in the structural proteome. Here, and in all following large-scale analyses, mutations associated with each distinct disease phenotype are considered separately from mutations associated with unrelated phenotypes so that a correspondence can be made between clusters in functionally relevant parts of protein structures and potential defects in molecular function that may cause one specific disease or type of cancer. We demonstrate the ability of mutation3D to distinguish functional from non-functional mutations in disease and to re-discover many known cancer-causing genes as well as discovering several new putative targets. Parameters for all tests performed are available in Methods and in Supp. Table S4.

mutation3D distinguishes disease mutations from common variants

To illustrate the efficacy of mutation3D in distinguishing functional from non-functional variants, we considered all proteins harboring at least 3 mutations associated with a single disease (according to HGMD) and all missense population variants (SNPs) from the ESP 6500 data set for this same set of proteins (see Methods for details). We were able to show that the resulting set of 8,869 disease-causing amino acid substitutions are more likely to be clustered by mutation3D than are 2,004 putatively benign substitutions arising from missense SNPs when considering only those mutations associated with a single disease at a time mixed together with SNPs in the same proteins (Figure 3a-b). This trend is apparent irrespective of whether the protein structure set is confined to known PDB structures, homology models from ModBase, or a combination of the two.

This analysis illustrates mutation3D's ability to distinguish functional from non-functional variants when all functional variants share an associated phenotypic consequence. Because it is often difficult to determine which cancer mutations are drivers and which are passengers, mutation3D's ability to distinguish functional disease mutations from non-functional SNPs serves as a proxy measure of its ability to separate functional driver mutations from a background of largely non-functional passenger mutations.

mutation3D identifies both new and well-known cancer genes

To confirm that mutation3D identifies plausible driver gene candidates in cancer (as judged by the existence of one or more clusters of substitutions in structures of their protein products), we computed statistically significant clusters from mutations in all WGS studies cataloged by COSMIC. First, we calculated the proportion of the identified cancer candidates that have been

previously proposed as cancer drivers based on a combination of the Cancer Gene Census database (Futreal, et al., 2004) and the MutSig driver list (Lawrence, et al., 2014). This is likely to be correlated with the lower bound of precision, or positive predictive value, of our method (see Methods). Figure 3c illustrates the calculated proportion values for all publications analyzed and for specific tissues within these studies, plotted over several cluster sizes. The results concur with our expectation that tighter mutation clusters should exhibit high precision for known cancer genes since substitutions in close physical proximity will be more likely than distant substitutions to be contained within the same interface domain or within the hydrophobic protein core. As expected, we also observe lower precision in the identification of genes involved in cancers of the skin, which are characterized by very high mutation rates (Alexandrov, et al., 2013). By contrast, cancers of the breast are known to harbor driver mutations in a relatively small number of genes and contain a relatively low proportion of passenger mutations (Kan, et al., 2010), thereby allowing mutation3D to precisely identify known cancer genes irrespective of cluster size.

To confirm that our statistical model yields plausible measures of cluster significance, we computed the statistical significance of clusters found in COSMIC WGS data. We find that our iterative bootstrapping model (See Methods) produces *P*-values that are highly correlated with the likelihood of a gene to be a known cancer genes (Figure 3d). We repeated both this and the study in Figure 3c using the Cancer Gene Census and MutSig cancer gene list separately to define a list of known cancer genes. We find the relative observed trends remain the same, confirming the robustness of our analyses (Supp. Figure S2).

We also find that the somatic mutations within these clusters are predicted to be more deleterious by PolyPhen-2 when found in smaller, more specific clusters (Figure 3e).

Furthermore, mutations within clusters are observed at much higher frequencies within WGS studies, suggesting they are likely to be driver mutations (Figure 3f). Overall, these analyses suggest a tendency for functionally important mutations to form clusters in cancer patient cohorts, whereas less important passenger mutations are more likely to fall outside these clusters.

We next investigated whether mutation3D preferentially reports potential oncogenes or tumor suppressors. We find that of genes annotated in either class based on the Cancer Gene Census, there is not a significant difference in the likelihood mutation3D will find clusters within their protein products (Supp. Note S5, Supp. Figure S3). This suggests that mutation3D is equally robust in its ability to detect oncogenes and tumor suppressors.

Finally, we produced a list of the genes whose protein products most commonly exhibit clusters of mutations within the same set of COSMIC WGS publications. We find that mutation3D implicates many well-known cancer genes (*TP53, KRAS, EGFR, BRAF*, etc.) as well as some genes that are missing from the Cancer Gene Census (Figure 4a). Visual inspection of the most significant clusters for each of these proposed genes demonstrates the power of 3D clustering (Figure 4b). A list of all genes found in at least 4 studies across COSMIC is available in Supp. Table S5.

Discussion

Researchers have already begun to acknowledge the added benefit of linear clustering approaches to the detection of driver mutations in two recently proposed methods (Lawrence, et al., 2014; Tamborero, et al., 2013). However, these methods do not take into account the 3D positions of mutations within protein products, disregarding information available due to structure-function relationships in proteins. Two other recent methods (Ryslik, et al., 2012;

Ryslik, et al., 2014) perform 1D clustering of mutations after a projection of 3D structural coordinates into 1D, potentially resulting in loss of information (Supp. Note S6, Supp. Figure S4). Clustering methods have also been used to detect signatures of positive selection (Tusche, et al., 2012; Wagner, 2007; Zhou, et al., 2008); however, the goals and assumptions of these methods and mutation3D are quite different (Supp. Note S7, Supp. Figure S5). Another recent method detects non-random distributions of mutations in protein crystal structures (Kamburov, et al., 2015). Although this method has shown in principle that 3D structural information is valuable for identifying target genes, it does not distinguish individual clusters and its analysis is limited to PDB structures.

Compared to the standard class of methods that do not search for clusters of amino acid substitutions, but instead employ measures of mutation frequency at the gene level to detect drivers of cancer, the added value of mutation3D lies in its orthogonal use of protein structures to make a more direct connection between alterations of structure and disruptions of function. We do not intend that mutation3D should replace these methods (Hodis, et al., 2012; Lawrence, et al., 2013; Sjöblom, et al., 2006). Instead, mutation3D gives scientists the ability to inspect their data through an additional lens—to visualize and form hypotheses about functional gene and protein candidates proposed by any method of cancer gene detection, and to find cases in which directly searching for structural disruptions may provide insights not available by other means. Even beyond its potential to improve candidate gene identification, mutation3D is valuable simply in terms of its ability to display mutations on all available high-quality structures and models, a task that requires significant effort on any scale without mutation3D, and can be accomplished on massive scales with mutation3D.

Throughout this study, we have evaluated the ability of mutation3D to identify *whether or not* a gene is involved in cancer because this is a standard for the cancer gene detection methods of today. However, such a metric may underrate the true ability of mutation3D, which can propose specific tumorigenic etiologies based on the structural localization of mutations. Even in cases where mutation3D identifies the same gene as another method, analyzing and viewing the mutations using mutation3D may present a specific hypothesis supported by both statistical and structural evidence, which may be more likely to inspire follow-up studies.

In addition to providing structural evidence for single proteins, the mutation3D web interface (http://mutation3d.org) allows users to rapidly search for clusters of mutations in the proteome (by inputting their data in a variety of popular genomic and proteomic formats), view and download clustering reports. Through the Advanced Query interface, users may adjust the clustering parameters and build structure and model sets for custom analysis of their own data or to seamlessly access pre-analysis of over 975,000 missense mutations in 6,811 primary cancer studies catalogued by COSMIC. Owing to the amount of data already available via the mutation3D web interface and the continual accumulation of cancer sequencing and protein structural data, mutation3D is likely to produce future insights based on structural localization of mutations in the human proteome.

References

Adzhubei I, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, Bork P, Kondrashov A, Sunyaev S. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7(4):248-249.

Alexandrov L, Nik-Zainal S, Wedge D, Aparicio S, Behjati S, Biankin A, Bignell G, Bolli N, Borg A, Børresen-Dale A-L, Boyault S, Burkhardt B, et al. 2013. Signatures of mutational processes in human cancer. Nature 500(7463):415-421.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28(1):235-242..

Cancer Genome Atlas N. 2012. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487(7407):330-337.

Das J, Fragoza R, Lee HR, Cordero NA, Guo Y, Meyer MJ, Vo TV, Wang X, Yu H. 2014. Exploring mechanisms of human disease through structurally resolved protein interactome networks. Mol BioSys 10(1):9-17. .

Forbes S, Bindal N, Bamford S, Cole C, Kok C, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague J, Campbell P, et al. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39(Database issue): D945-950.

Fu W, O'Connor T, Jun G, Kang H, Abecasis G, Leal S, Gabriel S, Rieder M, Altshuler D, Shendure J, Nickerson D, Bamshad M, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493(7431):216-220.

Futreal P, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton M. 2004. A census of human cancer genes. Nature Reviews Cancer 4(3):177-183.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185(4154):862-864.

Guedes J, Veiga I, Rocha P, Pinto P, Pinto C, Pinheiro M, Peixoto A, Fragoso M, Raimundo A, Ferreira P, Machado M, Sousa N, et al. 2013. High resolution melting analysis of *KRAS, BRAF* and *PIK3CA* in *KRAS* exon 2 wild-type metastatic colorectal cancer. BMC Cancer 13:169.

Hanahan D, Weinberg R. 2011. Hallmarks of cancer: the next generation. Cell 144(5):646-674.

Hodis E, Watson I, Kryukov G, Arold S, Imielinski M, Theurillat J-P, Nickerson E, Auclair D, Li L, Place C, Dicara D, Ramos A, et al. 2012. A landscape of driver mutations in melanoma. Cell 150(2):251-263.

Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, Lander ES, Getz G. 2015. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A 112(40):E5486-5495.

Kan Z, Jaiswal B, Stinson J, Janakiraman V, Bhatt D, Stern H, Yue P, Haverty P, Bourgon R, Zheng J, Moorhead M, Chaudhuri S, et al. 2010. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466(7308):869-873.

Kucukkal TG, Petukh M, Li L, Alexov E. 2015. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. Curr Opin Struct Biol 32:18-24.

Lawrence M, Stojanov P, Mermel C, Robinson J, Garraway L, Golub T, Meyerson M, Gabriel S, Lander E, Getz G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 505(7484):495-501.

Lawrence M, Stojanov P, Polak P, Kryukov G, Cibulskis K, Sivachenko A, Carter S, Stewart C, Mermel C, Roberts S, Kiezun A, Hammerman P, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499(7457):214-218.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26(16):2069-2070.

Miller Martin L, Reznik E, Gauthier Nicholas P, Aksoy Bülent A, Korkut A, Gao J, Ciriello G, Schultz N, Sander C. 2015. Pan-cancer analysis of mutation hotspots in protein domains. Cell Systems 1(3):197-209.

Muller P, Vousden K. 2013. p53 mutations in cancer. Nature Cell Biol 15(1):2-8.

Nishi H, Tyagi M, Teng S, Shoemaker BA, Hashimoto K, Alexov E, Wuchty S, Panchenko AR. 2013. Cancer missense mutations alter binding properties of proteins and their interaction networks. PLoS One 8(6):e66273.

Petukh M, Kucukkal TG, Alexov E. 2015. On human disease-causing amino acid variants: statistical study of sequence and structural patterns. Hum Mutat 36(5):524-34.

Pieper U, Webb B, Barkan D, Schneidman-Duhovny D, Schlessinger A, Braberg H, Yang Z, Meng E, Pettersen E, Huang C, Datta R, Sampathkumar P, et al. 2011. ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 39(Database issue): D465-474.

Pylayeva-Gupta Y, Grabocka E, Bar-Sagi D. 2011. RAS oncogenes: weaving a tumorigenic web. Nat Rev Cancer 11(11):761-774.

Ryslik GA, Cheng Y, Cheung K-HH, Modis Y, Zhao H. 2012. Utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics 14:190.

Ryslik GA, Cheng Y, Cheung K-HH, Modis Y, Zhao H. 2014. A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics 15(1):86.

Schrodinger, LLC. 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

Sjöblom T, Jones S, Wood L, Parsons D, Lin J, Barber T, Mandelker D, Leary R, Ptak J, Silliman N, Szabo S, Buckhaults P, et al. 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314(5797):268-274.

Sneath P. 1957. The application of computers to taxonomy. J Gen Microbiol 17(1):201-226.

Sørensen T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. 5:1-34.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133(1):1-9.

Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. 2013. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 29(18):2238-2244.

Tusche C, Steinbruck L, McHardy AC. 2012. Detecting patches of protein sites of influenza A viruses under positive selection. Mol Biol Evol 29(8):2063-71.

Velankar S, Dana J, Jacobsen J, van Ginkel G, Gane P, Luo J, Oldfield T, O'Donovan C, Martin M-J, Kleywegt G. 2013. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res 41(Database issue):9.

Vucic E, Thu K, Robison K, Rybaczyk L, Chari R, Alvarez C, Lam W. 2012. Translating cancer 'omics' to improved outcomes. Genome Res 22(2):188-195.

Wagner A. 2007. Rapid detection of positive selection in genes and genomes through variation clusters. Genetics 176(4):2451-2463.

Wood L, Parsons D, Jones S, Lin J, Sjöblom T, Leary R, Shen D, Boca S, Barber T, Ptak J, Silliman N, Szabo S, et al. 2007. The genomic landscapes of human breast and colorectal cancers. Science 318(5853):1108-1113.

Zhou T, Enyeart PJ, Wilke CO. 2008. Detecting clusters of mutations. PLoS One 3(11):e3765.

Figure Legends

Figure 1. Missense mutations in both Mendelian disorders and cancer form clusters in tertiary protein structures. Linear protein models are given below each structure to illustrate the importance of studying 3D crystal structures. (a) Protein substitutions arising from mutations known to cause aromatase deficiency (in red) are shown overlaid upon an experimentally determined crystal structure of aromatase. Protein substitutions arising from common missense SNPs with MAF ≥ 1% (in blue) are shown to aggregate within regions of the protein structure that are distinct and spatially separated from those harboring the pathogenic substitutions, suggesting a strong relationship between the position of a substitution in the protein and its functional consequence(s). (b) Mutations causing amino acid substitutions in codons 12, 13 and 61 account for over 99% of the mutations in GTPase KRas reported by COSMIC. The three most common amino acid substitutions at these positions (shown in red) form a tight cluster in a crystal structure of GTPase KRas, whereas a substitution (E49K) only observed once in COSMIC (shown in blue), is likely to be a passenger mutation and falls outside the 3D mutation cluster even though it appears to be in close proximity in the linear model.

Figure 2. An overview of the mutation3D clustering and web accession procedures. (a) Sources of 3D protein structures and models and missense mutations in cancer. Pre-computation of clusters of amino acid substitutions for large data sets occurs with each COSMIC update. (b) There are three options for users to determine clustering: by inputting their own data as substitutions in single proteins (or nucleotide mutations in genes), by uploading a file of mutations, or by analyzing missense mutations from one of the 6,811 publications curated by

COSMIC. (c) The mutation3D web interface shows clusters on both linear models and interactive 3D models. Users may select among available models and structures. Individual queries will lead directly to this page, while batch queries will first lead to a table of proteins and clusters (shown below).

Figure 3. (a-b) Known inherited disease-associated missense mutations from HGMD and missense SNPs from ESP 6500 with MAF ≥ 1% were clustered using mutation3D, with the percentage of variants within proteins containing clusters reported. (a) The combined set of resulting amino acid substitutions was plotted onto 3D protein models derived from the PDB alone, ModBase alone, and a combination of the two. (b) Fractions of clustered mutations were recalculated only for those mutations that reside within protein regions for which a 3D structure or model exists. (c-e) mutation3D was run on 285 WGS somatic tissue screens in COSMIC. (c) A higher fraction of protein candidates identified are known cancer genes at smaller values of cluster size (maximum cluster diameter). (d) A higher fraction of protein candidates identified are known cancer genes at smaller clustering *P*-values. (e) Mutations in tighter clusters are predicted by PolyPhen-2 to be more damaging than those in sparser clusters and in all WGS studies. (f) Mutations in tighter clusters are more likely to be observed at high frequency across COSMIC WGS studies. For all panels, * indicates $P < 0.01$.

Figure 4. (a) The top 20 genes implicated in WGS studies by mutation3D ranked by the number of publications in which clusters were observed for each. (b) The most significant cluster for each of the 20 implicated genes are shown in 3D.