Supporting Information for the article:

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[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

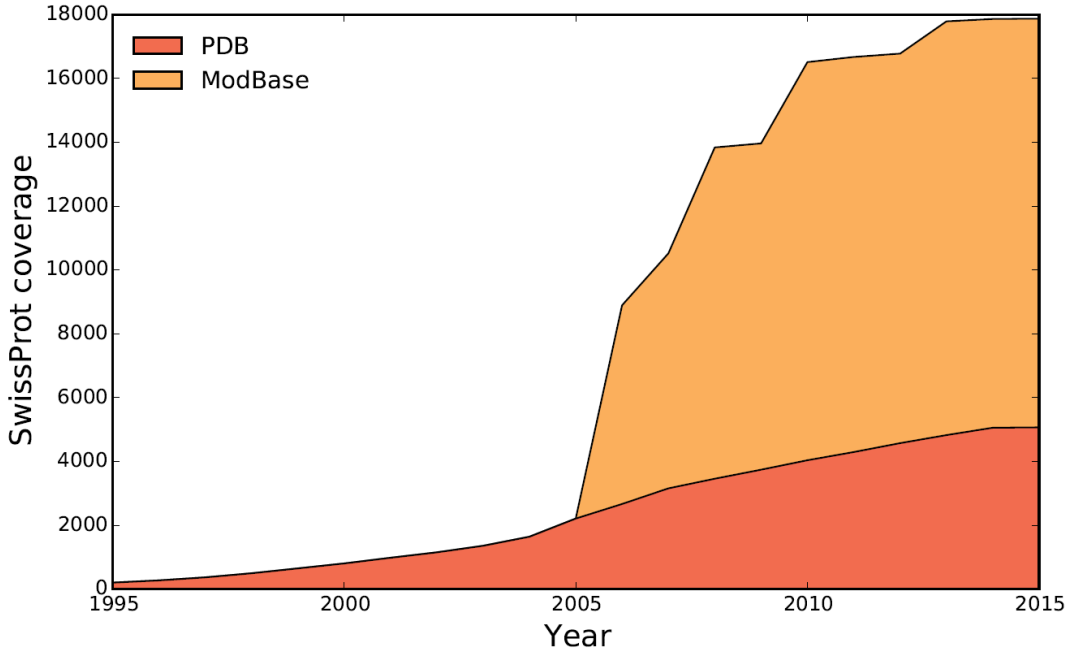
[†]The authors wish it to be known that, in their opinion, the first 3 authors should be regarded as joint First Authors

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

Contents

Supplementary Note S1: Sources of the Structural Proteome

In order to build our repository of protein structures and models, we curated experimentally-determined crystal structures from the PDB (Berman, et al., 2000) and homology models from ModBase (Pieper, et al., 2011) by searching for Swiss-Prot (UniProt-Consortium, 2015) tagged structures or chains in both (see Methods). Over the past 20 years, the coverage of these databases has increased dramatically, enabling clustering on the scale of the whole proteome (Supp. Fig. S1).



Supp. Figure S1. The growth of the two sources of protein structures for mutation3D. By April 2015, the PDB and ModBase together accounted for ~ 88% of the verified human proteome (SwissProt) as measured by the number of unique entries in ModBase and PDB matching unique SwissProt IDs (for verified, canonical isoforms) divided by the total number of SwissProt proteins in UniProt. At the time of this calculation, there were 5,068 PDB entries representing unique SwissProt proteins (≥ 20 residues), 12,809 ModBase entries matching unique SwissProt proteins (≥ 20 residues and MPQS ≥ 0.5) not in the PDB, and 20,204 total SwissProt proteins catalogued by UniProt.

Supplementary Note S2: Model Filter Categories

mutation3D supplements its crystal structure coverage of the proteome with homology models derived from ModBase (Pieper, et al., 2011). ModBase provides many quality metrics to determine the accuracy of each of their models, several of which we have included for filtering purposes. The user may select a subset of models by setting these parameters on the advanced page. All parameters are set to their most relaxed levels, except for *MPQS*, as this parameter encompasses all other parameters, thereby allowing some leniency in the others (i.e. one low quality parameter can be compensated for by several high quality parameters). ModBase states that a model should be considered to have a reliable fold assignment if any one of parameters 3-5 (below) fall within an acceptable range. Thus, setting all parameters to their suggested thresholds may inappropriately and unnecessarily remove certain models from further consideration. However, individual users may still choose to filter on different or additional parameters by setting any combination of the parameters below.

(1) *Protein Coverage*: The fraction of the full length UniProt (UniProt-Consortium, 2015) protein covered by the model, irrespective of the identity or accuracy of the 3D amino acid positions in the model. There is no suggested cutoff for *Protein Coverage* as it can vary based upon the user's specific requirements.

(2) *Sequence Identity*: The fraction of a model's amino acid sequence that is identical to the UniProt amino acid sequence, on a scale of 0 to 1. Note that this measure is independent of the *Protein Coverage* (i.e. *Sequence Identity* can still equal 1 if all of the amino acids included in the model are identical to amino acids in the covered region of the protein irrespective of how large this region is). There is no suggested cutoff for *Sequence Identity* as it can vary based upon the user's specific requirements.

(3) *e-value*: The significance of the alignment between the template PDB (Berman, et al., 2000) structure sequence and the target UniProt sequence as reported by NCBI's PSI-BLAST program (Altschul, et al., 1997) or similar alignment score calculated by ModPipe (Pieper, et al., 2011). The ModBase-suggested quality cutoff using *e-value* alone is $e\text{-}value < 10^{-4}$.

(4) *Discrete Optimized Protein Energy (DOPE) score*: Also known as *zDOPE*, it is a derived atomic distance-dependent statistical potential calculated by ModBase from a sample of native structures. The ModBase publication describes this measure in great detail (Pieper, et al., 2011). Lower *zDOPE* scores indicate a more accurate model. The ModBase-suggested quality cutoff using *zDOPE* score alone is $zDope < 0$.

(5) *ModPipe Quality Score (MPQS)*: A composite score calculated by ModBase comprising several measures including measures 1-4. Since this score incorporates all previous scores, mutation3D uses the ModBase-suggested threshold for a high quality model of $MPQS \geq 1.1$ by default.

Supplementary Note S3: Clustering Parameters

These parameters define the properties of an acceptable amino acid substitution cluster in mutation3D. Suggested values are pre-set in both the standard query interface and the advanced page, but the user may choose to change the parameters from the advanced page.

(1) *Minimum Number of Substitutions*: The minimum number of amino acid substitutions required to form a cluster. This refers to the absolute number of substitutions, irrespective of whether they exist at the same amino acid position or arise from the same underlying nucleotide mutation across multiple clinical samples. By default, the *Minimum Number of Substitutions* = 3.

(2) *Minimum Number of Unique Substitutions*: The minimum number of amino acid substitutions existing at unique amino acid indices required to form a cluster. In this sense, multiple mutations within one codon are counted as a single unique substitution, irrespective of whether or not they give rise to different mutant amino acid residues. By default, the *Minimum Number of Unique Substitutions* = 2.

(3) *Maximum Cluster Diameter*: The maximum allowable distance in Angstroms between any two amino acid substitutions in a cluster. This is the same as the CL-distance in Complete Linkage Clustering and can be thought of as the maximum allowable diameter of a sphere containing all points in a cluster. By default, the *Maximum Cluster Diameter* = 15 Å.

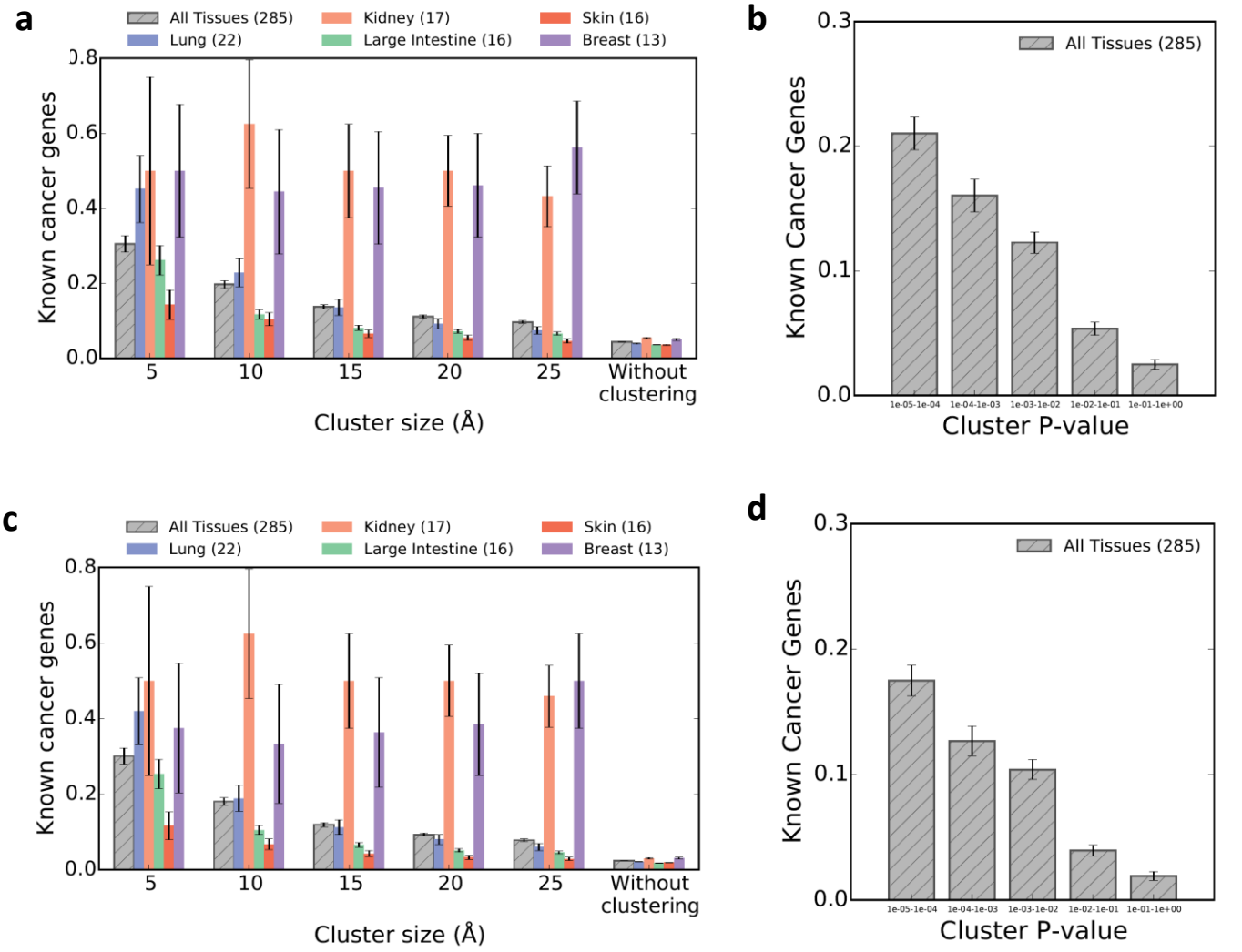
(4) *Minimum Linear Separation*: A post-clustering filter parameter to remove clusters that do not span a specified distance in amino acid index space (i.e. the number of amino acids separating two residues in a linear chain). Users interested in identifying examples of clusters that would only be observable in 3D space should set this parameter higher. By default, *Minimum Linear Separation* = 0.

Supplementary Note S4: Amino acid substitution patterns in the Ras GTPase protein family

To assess the plausibility of the postulate that driver mutations, because of their functional similarity, could occur as clustered amino acid substitutions in the same protein, we considered the canonical Ras family of cancer genes (*KRAS, NRAS, HRAS*) and their corresponding protein products, each 189 amino acids in length. According to COSMIC (Forbes, et al., 2011), 99% of all somatic missense mutations in these genes occur in codons 12, 13 and 61 (Supp. Table 1). That these mutations have been noted at such high frequencies in tumor tissues strongly supports the view that they are drivers of tumorigenesis rather than passengers (Stratton, et al., 2009).

It is highly likely that the functional mechanisms by which missense mutations in codons 12 and 13 confer their tumorigenic phenotypes are closely related because of the assured proximity of the juxtaposed amino acids within the tertiary structures of their respective proteins. However, the mechanism by which missense mutations in codon 61 exert their tumorigenic effects is less clear, given its position on the protein backbone. In cases such as these, and systematically across genome-wide cancer mutation datasets, efforts to discern functionality shared by linearly remote but spatially clustered missense mutations must be made at a protein structural level using crystal structures and models.

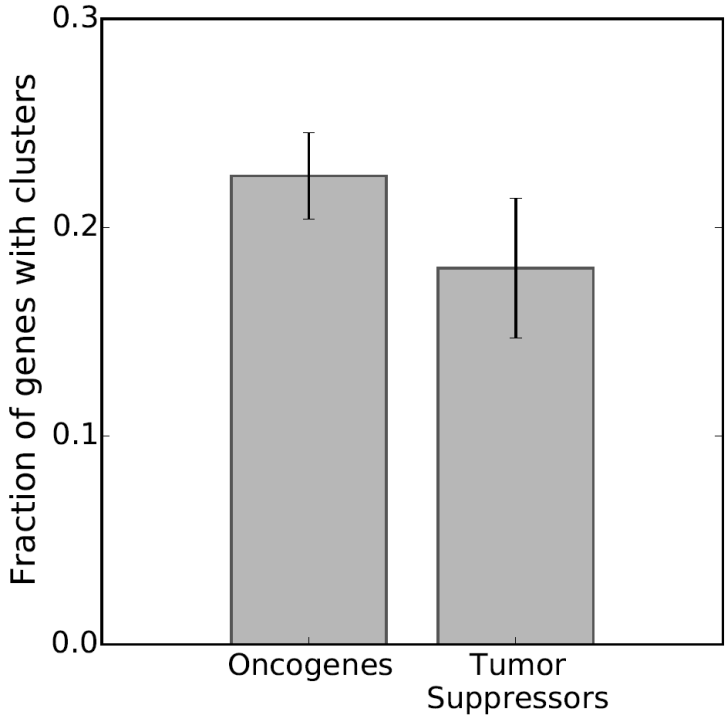
In the case of the RAS-family protein products, there are many available crystal structures and models. In Figure 1b, we highlight the locations of the three known driver mutations in KRAS and show that they form a tight cluster within the crystal structure. This spatial relationship was revealed by mutation3D through clustering in many primary sequencing studies available through COSMIC that sequenced Ras-family genes, and suggests that the three substitutions in Ras-family proteins act in a mechanistically similar fashion so as to confer the tumorigenic phenotype. This conclusion is not of course novel since it has been shown many times before that the mechanisms by which these three mutations exert their effects in the RAS proteins are closely related and that the cancer phenotype is likely to be attributable to the constitutive binding of GTP (Pylayeva-Gupta, et al., 2011). Such an example does however illustrate how mutation3D could in principle be used to explore previously unknown disease etiologies and potentially inform treatment regimens.



Supp. Figure S2: mutation3D was run on 285 WGS somatic tissue screens in COSMIC. (a-b) Using the Cancer Gene Census as the set of known cancer genes and (c-d) using the MutSig cancer gene list as the set of known cancer genes. (a & c) A higher fraction of protein candidates identified are known cancer genes at smaller values of cluster size (maximum cluster diameter). (b & d) A higher fraction of protein candidates identified are known cancer genes at smaller clustering P-values.

Supplementary Note S5: Oncogenes vs. Tumor Suppressors

We explored the ability of mutation3D to detect oncogenes and tumor suppressors. Since oncogenes tend to act via gain-of-function and tumor suppressors tend to act via loss-of-function, we considered it likely that clusters would be found more frequently in oncogenes than in tumor suppressors, as gain of function may require more specific mutations than loss of function (Hanahan and Weinberg, 2000). However, we also considered that tumor suppressors may be more likely to contain missense mutations in the first place, while oncogenes are known to often act through somatic copy number alterations (Beroukhim, et al., 2010). Overall we find that mutation3D does not preferentially detect either class of cancer gene, demonstrating robustness to detect both oncogenes and tumor suppressors (Supp. Figure S3).

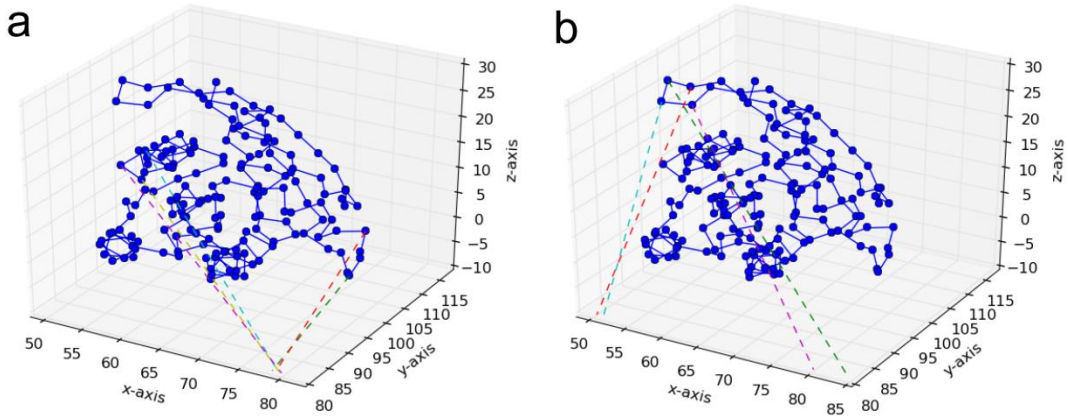


Supp. Figure S3: The Cancer Gene Census annotates many genes as oncogenes or tumor suppressors. Plotted are the fractions genes in each set that have been detected by mutation3D with at least one cluster in COSMIC WGS studies.

Supplementary Note S6: Reduction-to-1D Clustering Methods

Two recent reduction-to-1D clustering methods, iPAC (Ryslik, et al., 2012) and GraphPAC (Ryslik, et al., 2014), take into account the 3D structure of proteins in order to identify non-random somatic mutations in cancer. The key difference between mutation3D and these methods is that mutation3D performs 3D clustering by using the coordinates of α-carbons directly in protein models. Any cluster found by mutation3D therefore exists by definition in 3D space, and visualization of clusters using mutation3D's web interface demonstrates this. On the other hand, while reduction-to-1D methods make use of the 3D coordinate information in protein models, they first reduce the number of dimensions from 3 to 1 in order to use algorithms designed for 1D clustering, such as Non-Random Mutational Clustering (Ye, et al., 2010) (NMC).

While clustering performed by NMC in 1D may be accurate given the projected 1D coordinates, the projected coordinates themselves will not retain all of the information of the original 3D coordinates. Although such reduction-to-1D methods attempt to minimize loss of information due to dimensionality reduction, they cannot eliminate it. Here, this loss of information can lead to both false positives and false negatives for the identification of clusters in the original 3D space. To illustrate this, we have shown in Supp. Figure S4 how dimensionality reduction using Multidimensional scaling (Borg, 1997) (MDS) to transform 3D coordinates into 1D can lead to both false positive and false negative 3D clusters in KRAS.



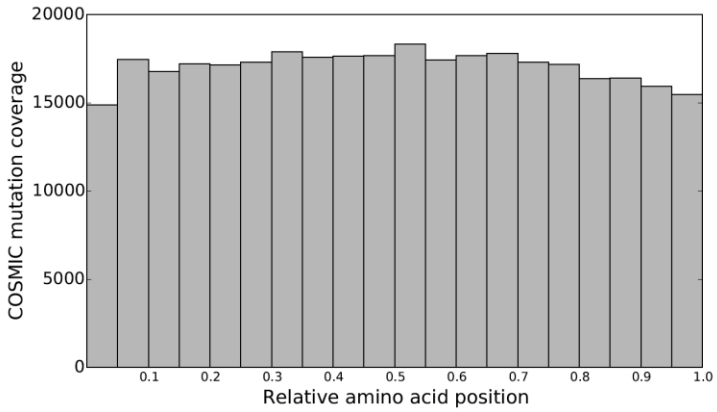
Supp. Figure S4: α-carbons in KRAS are plotted in 3D space according to the coordinates provided by the PDB (3GFT). In each figure panel, selected projections from 3D to 1D coordinates using MDS are shown scaled between 50 and 85 on the x-axis. (a) α-carbons in two distinct areas of the 3D structure are projected very near in 1D space. While any of these could potentially be clustered based on the 1D projections, not all of the initial α-carbon positions are close in 3D space, resulting in false positive identification of clusters. (b) Four α-carbons close in 3D space are very far apart when projected into 1D using MDS, resulting in false negatives.

Supplementary Note S7: Statistical Bootstrapping Model

Clusters in mutation3D are found in protein structures and models using an implementation of complete-linkage clustering. These clusters exist by virtue of the spatial arrangement of amino acid substitutions. However, in order to assess whether these clusters represent significant findings or simply arise by chance we must employ a model of statistical significance.

There are many methods (Kan, et al., 2010; Lawrence, et al., 2013; Sjöblom, et al., 2006) already available to address the hypermutation hypothesis—that genes with a number of mutations far above expectation given background mutations rates are likely to be involved in cancer. We make a distinction from these methods in order to test whether the spatial arrangement of mutations on the protein backbone alone is significant. In other words, do amino acid substitutions occur in non-random proximity given the contours of the protein backbone?

In order to test this hypothesis, we must perform randomized iterations of amino acid substitution placement to produce a background distribution (see Materials and Methods), as each protein model and structure will have a different null expectation. For instance, a highly bunched structure will be very likely to produce clusters if all residues are with a short distance from each other. On the other hand, a nearly linear structure, with residues maximally far apart from each other, will be less likely to produce clusters given the same number of randomly chosen substitution positions.



Supp. Figure S5. The relative positions of all amino acid substitutions in 175 whole genome studies in COSMIC. Relative position = (amino acid index)/(protein length).

There is a substantial body of literature (Ramsey, et al., 2011; Toth-Petroczy and Tawfik, 2011; Tusche, et al., 2012; Zhou, et al., 2008) describing analyses of positive selection rates suggesting that amino acids on the surface of a protein are more likely to be mutated than those that are buried. When this is the case, a statistical model designed to identify clusters should adjust for the difference in these substitution rates to avoid finding spurious clusters on protein surfaces. However, we note that amino acids substitutions in WGS cancer screens do not preferentially occur near the C or N termini of proteins (Supp. Fig. S5), which should be more likely to be exposed in protein structures (this litmus test has been used in these evolutionary analyses to propose the need to adjust for the differences in observed rates of surface and buried residues). Therefore, since we do not observe a difference in substitution rates between buried and surface residues in cancer, and because mutation3D has been designed for a more general use case to detect clusters in any dataset, mutation3D does not treat buried and surface residues differently in its iterative bootstrapping model.

Supp. References

Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17):3389-3402.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Research 28(1):235-42.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, et al. 2010. The landscape of somatic copy-number alteration across human cancers. Nature 463(7283):899-905.

Borg I, Groenen, P. J. F. 1997. Modern multidimensional scaling : theory and applications.

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 Research 39(Database issue):50.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. Cell 100(1):57-70.

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.

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.

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 Research 39(Database issue):74.

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

Ramsey DC, Scherrer MP, Zhou T, Wilke CO. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188(2):479-88.

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.

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.

Stratton M, Campbell P, Futreal P. 2009. The cancer genome. Nature 458(7239):719-724.

Toth-Petroczy A, Tawfik DS. 2011. Slow protein evolutionary rates are dictated by surface-core association. Proc Natl Acad Sci U S A 108(27):11151-6.

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.

UniProt-Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res 43(Database issue):D204-12.

Ye J, Pavlicek A, Lunney E, Rejto P, Teng C-H. 2010. Statistical method on nonrandom clustering with application to somatic mutations in cancer. BMC Bioinformatics 11:11.

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

Supp. Table S1. The inherited disease associated amino acid substitutions from HGMD shown in Figure 1a for aromatase (*CYP19A1* gene, transcript: NM_031226.2)

Nucleotide Change	Amino Acid Substitution
c.254T>G	M85R
c.380T>G	M127R
c.1094G>A	R365Q
c.1108G>A	V370M
c.1123C>T	R375C
c.1124G>A	R375H
c.1232A>G	N411S
c.1303C>T	R435C
c.1310G>A	C437Y

Nucleotides are indexed in coding sequences, using the A of the ATG translation initiation start site as nucleotide 1.

Supp. Table S2. The SNPs and their associated amino acid substitutions as shown in Figure 1a of aromatase (*CYP19A1* gene, transcript: NM_031226.2)

dbSNP ID	Nucleotide Change	Amino Acid Substitution	ESP 6500 Allele Frequency
rs28757184	c.602C>T	T201M	0.04446
rs700519	c.790C>T	R264C	0.077374

Both SNPs are predicted to be benign by PolyPhen-2. Nucleotides are indexed in coding sequences, using the A of the ATG translation initiation start site as nucleotide 1.

Supp. Table S3. The distribution of amino acid substitutions in Ras GTPase codons reported in COSMIC v67

		Protein Product		
		HRAS (P01112)	KRAS (P01116)	NRAS (P01111)
Codon	12	419 (G>V)	23,742 (G>D)	746 (G>D)
	13	118 (G>R)	4,101 (G>D)	373 (G>D)
	61	367 (Q>R)	367 (Q>H)	1,863 (Q>R)
	All	940	28,444	3,036

Given in parentheses are the most common substitutions for each codon in each protein.

Supp. Table S4. Computational parameters for main Figure 3

Figure	Input Data	Unique amino acid indices	Amino acids	ModBase MPQS	Maximum cluster diameter (Å)	*P*-value	Output Measurement
3a/b	HGMD and ESP	≥ 2	≥ 3	≥ 1.1	15	n/a	Fraction of substitutions clustered
3c	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	5, 10, 15, 20, 25	n/a	Fraction of known cancer genes
3d	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	25	10^{0}, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}	Fraction of known cancer genes
3e	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	5, 15, 25	n/a	PolyPhen-2 scores of clustered mutations
3f	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	5, 15, 25	n/a	Frequency of clustered mutations in COSMIC
4a	COSMIC WGS	≥ 2	≥ 3	≥ 1.1	15	≤ 0.01	Top genes predicted by mutation3D

Each panel 3a-f is designed to test only one parameter at a time (i.e. protein model set, *P*-value, or cluster diameter). Figure 4b combines recommended parameters to predict cancer genes.

Note: Header ‘Unique amino acid indices’ refers to the minimum number of unique amino acid positions needed to be mutated to constitute a cluster. Header ‘Amino acids’ refers to the minimum total number of amino acids (including at the same indices) needed to be mutated to constitute a cluster. Default values for these parameters are used throughout the study (see Supp. Note S3).

Supp. Table S5. Genes whose protein products contain clusters of mutations in at least 4 COSMIC WGS studies

Gene	UniProt	# Studies	CGC?	Mutsig?	Most common		Most Significant Cluster			
					Tissue	Mutations	Tissue	Cluster	PMID	*P*-value
TP53	P04637	101	Y	Y	Central Nervous System(12)	248(51), 273(47), 175(45), 245(31), 179(26)	Lung	R175L(6), M246I(5), R248L(5)	23856246	<6.67E-05
KRAS	P01116	33	Y	Y	Large Intestine(10)	12(30), 13(21), 61(17), 146(13), 117(5)	Pancreas	G12V(80), G12D(74), G12R(24), Q61H(10), G12C(8), Q61R(2), G12S(2), Q61K(2), G13D(2)	25855536	<6.67E-05
PIK3CA	P42336	24	Y	Y	Large Intestine(4)	545(16), 542(15), 546(10), 1047(8), 420(5)	Ovary	E545K(4), E542K(2), Q546K(1), E542V(1), E545V(1), E545A(1)	20826764	<6.67E-05
FBXW7	Q969H0	16	Y	Y	Large Intestine(6)	505(11), 465(10), 479(4), 626(2), 423(2)	Large Intestine	V464M(5), R465H(5)	24599305	<6.67E-05
NRAS	P01111	10	Y	Y	Haematopoietic And Lymphoid Tissue(5)	12(10), 13(8), 61(7), 68(1), 66(1)	Haematopoietic And Lymphoid Tissue	G12S(3), G13D(2), G12V(1), G13R(1), Q61R(1), Q61K(1), G12A(1)	25381062	<6.67E-05
HRAS	P01112	9	Y	Y	Upper Aerodigestive Tract(5)	61(9), 12(8), 13(7), 58(1), 117(1)	Skin	Q61L(4), T58I(2), G12S(2), Q61K(2), G13D(2)	25303977	<6.67E-05
SMAD4	Q13485	9	Y	Y	Large Intestine(5)	361(7), 356(6), 351(4), 386(4), 524(2)	Oesophagus	G386D(3), D351H(1)	23525077	<6.67E-05
CDKN2A	Q8N726	8	Y	Y	Large Intestine(2)	94(5), 102(4), 97(3), 98(2), 107(1)	Oesophagus	H107Q(4), R98L(4), G102V(1), P94L(1)	25839328	<6.67E-05
FRG1B	Q9BZ01	8	N	N	Large Intestine(2)	80(3), 101(3), 50(2), 60(2), 88(2)	Skin	Y46H(4), L50P(4), M11V(2)	24265154	<6.67E-05
GRIK2	Q13002	8	N	N	Skin(3)	830(3), 763(2), 216(1), 213(1), 764(1)	Lung	V526L(2), E524D(2)	22941189	<6.67E-05
NRXN1	Q9ULB1	8	N	N	Large Intestine(3)	1216(2), 1077(2), 1151(1), 856(1), 1167(1)	Large Intestine	A660T(4), C643Y(4)	24211491	<6.67E-05
EGFR	P00533	7	Y	Y	Lung(3)	858(3), 62(2), 289(2), 719(2), 768(1)	Lung	L858R(11), L833V(1), R889G(1)	25189529	<6.67E-05
GRIA2	P42262	7	N	N	Large Intestine(3)	692(1), 752(1), 696(1), 718(1), 335(1)	Oesophagus	T501N(2), G752C(2)	25839328	<6.67E-05
BRAF	P15056	6	Y	Y	Lung(2)	469(5), 600(4), 601(3), 597(2), 466(2)	Urinary Tract	G596R(1), F595L(1), D594G(1), G469A(1)	24121792	<6.67E-05
CDKN2A	P42771	6	Y	Y	Oesophagus(2)	83(3), 114(2), 84(2), 102(2), 21(1)	Skin	P114L(6), E88K(3), H83R(2)	22817889	<6.67E-05
EPHA3	P29320	6	N	N	Large Intestine(3)	48(1), 83(1), 75(1), 94(1), 357(1)	Oesophagus	N79I(2), D75G(2)	23525077	0.00085
GRM8	O00222	6	N	N	Large Intestine(2)	38(1), 212(1), 465(1), 261(1), 53(1)	Liver	G465E(3), D407N(3)	25822088	<6.67E-05
PLXNA4	Q9HCM2	6	N	N	Large Intestine(4)	333(2), 591(1), 459(1), 318(1), 1326(1)	Large Intestine	A437T(4), R459W(2)	25344691	0.00415
DPP10	Q8N608	5	N	N	Large Intestine(2)	152(2), 144(1), 140(1), 603(1), 571(1)	Skin	E220K(2), S152L(2)	21984974	<6.67E-05
EPHB1	P54762	5	N	N	Stomach(1)	846(1), 691(1), 616(1), 190(1), 117(1)	Ovary	T117N(2), A176T(2)	21720365	<6.67E-05
FCAR	P24071	5	N	N	Large Intestine(4)	178(2), 128(2), 155(1), 195(1), 51(1)	Large Intestine	L128P(2), F178L(2)	23856246	<6.67E-05
ITGB8	P26012	5	N	N	Large Intestine(2)	146(1), 339(1), 140(1), 330(1), 333(1)	Upper Aerodigestive Tract	C481Y(2), N474K(2)	25275298	<6.67E-05
KCNMA1	Q12791	5	N	N	Large Intestine(2)	630(1), 604(1), 1096(1), 909(1), 905(1)	Skin	R865C(3), D1096N(3)	22842228	<6.67E-05
LCK	P06239	5	Y	N	Large Intestine(3)	458(2), 484(2), 151(1), 197(1), 291(1)	Haematopoietic And Lymphoid Tissue	A289D(2), A396V(2)	23856246	0.000133
MGAM	O43451	5	N	N	Skin(2)	246(2), 115(1), 123(1), 1045(1), 790(1)	Ovary	K281R(2), V263M(2)	21720365	<6.67E-05
MYH13	Q9UKX3	5	N	N	Lung(1)	590(1), 58(1), 712(1), 108(1), 109(1)	Oesophagus	D58E(2), P79L(2)	25151357	<6.67E-05
NTRK3	Q16288	5	Y	N	Skin(2)	584(1), 583(1), 746(1), 747(1), 382(1)	Stomach	F747V(2), K746T(2)	24816253	<6.67E-05
OR1L8	Q8NGR8	5	N	N	Large Intestine(2)	201(1), 200(1), 127(1), 247(1), 123(1)	Skin	H6Y(2), N5S(2)	25303977	<6.67E-05
PDE10A	Q9Y233	5	N	N	Large Intestine(2)	752(1), 338(1), 719(1), 704(1), 405(1)	Liver	M704I(2), A722V(2)	25822088	<6.67E-05
PIK3R1	P27986	5	Y	Y	Haematopoietic	567(2), 573(1), 464(1), 452(1),	Haematopoietic	L573P(3), N564K(3), K567E(3)	23143597	<6.67E-05

Gene	UniProt	# Studies	CGC?	Mutsig?	Most common		Most Significant Cluster			
					Tissue	Mutations	Tissue	Cluster	PMID	*P*-value
					And Lymphoid Tissue(2)	564(1)	And Lymphoid Tissue			
SYK	P43405	5	Y	N	Large Intestine(2)	33(1), 330(1), 29(1), 52(1), 428(1)	Lung	K104R(1), K105N(1), K105E(1)	22980975	<6.67E-05
ABCB5	Q2M3G0	4	N	N	Soft Tissue(2)	678(1), 581(1), 560(1), 596(1), 680(1)	Pancreas	K560E(3), M525I(3)	25855536	<6.67E-05
ACO1	P21399	4	N	Y	Large Intestine(2)	378(2), 263(2), 448(1), 41(1), 61(1)	Large Intestine	T263I(2), D378N(2)	23856246	<6.67E-05
ADSL	P30566	4	N	N	Large Intestine(4)	300(2), 296(2), 24(1), 77(1), 354(1)	Large Intestine	R300H(6), A291V(2), R296Q(2)	24755471	<6.67E-05
CASP8	Q14790	4	Y	Y	Large Intestine(3)	228(2), 232(2), 233(2), 236(1), 237(1)	Large Intestine	M228I(3), P232H(3)	23856246	<6.67E-05
CHAT	P28329	4	N	N	Large Intestine(1)	217(1), 618(1), 174(1), 669(1), 610(1)	Liver	A217V(2), C669S(2)	25822088	<6.67E-05
CLVS2	Q5SYC1	4	N	N	Large Intestine(1)	135(1), 277(1), 99(1), 109(1), 265(1)	Liver	A109T(2), L265M(2)	25822088	<6.67E-05
CTBP2	P56545	4	N	N	Large Intestine(2)	157(1), 336(1), 82(1), 64(1), 341(1)	Haematopoietic And Lymphoid Tissue	R157W(4), T160K(4)	24970810	0.00017
CTNNA2	P26232	4	N	N	Stomach(2)	753(1), 323(1), 893(1), 756(1), 754(1)	Stomach	D241E(2), R240P(2)	25042771	<6.67E-05
CTNNA3	Q9UI47	4	N	N	Large Intestine(3)	214(1), 842(1), 834(1), 415(1), 158(1)	Large Intestine	R842L(2), R834Q(2)	24211491	<6.67E-05
EPHA4	P54764	4	N	N	Oesophagus(1)	775(1), 604(1), 771(1), 154(1), 773(1)	Lung	D773N(2), D604Y(2)	22980975	<6.67E-05
ERCC2	P18074	4	Y	Y	Urinary Tract(1)	312(1), 44(1), 42(1), 463(1), 181(1)	Bone	K181T(2), R185Q(2)	25186949	<6.67E-05
FN1	P02751	4	N	N	Large Intestine(2)	414(1), 447(1), 1658(1), 2252(1), 1861(1)	Oesophagus	R2266H(2), H2252R(2)	25839328	<6.67E-05
GCK	P35557	4	N	N	Large Intestine(2)	147(1), 21(1), 156(1), 63(1), 304(1)	Skin	F330Y(2), G328R(2)	22842228	<6.67E-05
GNAS	Q5JWF2	4	Y	N	Upper Aerodigestive Tract(1)	844(2), 740(1), 869(1), 961(1), 874(1)	Upper Aerodigestive Tract	R844C(2), N740S(2)	21798893	<6.67E-05
GNAS	P63092	4	Y	N	Upper Aerodigestive Tract(1)	201(2), 318(1), 338(1), 227(1), 226(1)	Peritoneum	R201C(22), R201H(8), R201L(2), Q227H(2)	24944587	<6.67E-05
GRIA4	P48058	4	N	N	Large Intestine(2)	627(1), 634(1), 561(1), 45(1), 342(1)	Skin	G367R(3), G342R(3)	25303977	0.00302
GRIK1	P39086	4	N	N	Large Intestine(2)	797(1), 857(1), 610(1), 218(1), 122(1)	Stomach	V128G(3), I122M(3)	24816253	8.99E-05
GRIK3	Q13003	4	N	N	Large Intestine(1)	450(1), 301(1), 440(1), 447(1), 455(1)	Large Intestine	T455M(4), R450Q(2)	25344691	0.000321
GRM1	Q13255	4	N	N	Large Intestine(2)	309(1), 44(1), 369(1), 43(1), 460(1)	Large Intestine	V309M(2), R275H(2), E311K(2)	25344691	0.000234
GRM3	Q14832	4	N	N	Large Intestine(2)	59(1), 49(1), 46(1), 460(1), 306(1)	Skin	F187L(2), S182L(2), G460E(1)	25303977	0.00302
HK3	P52790	4	N	N	Large Intestine(2)	214(1), 215(1), 772(1), 676(1), 777(1)	Haematopoietic And Lymphoid Tissue	P676S(2), Y673C(2)	23292937	<6.67E-05
HLA-B	P30486	4	N	Y	Upper Aerodigestive Tract(1)	201(1), 202(1), 140(1), 119(1), 107(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-B	Q29836	4	N	Y	Upper Aerodigestive Tract(1)	91(2), 201(1), 202(1), 155(1), 140(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-B	P01889	4	N	Y	Upper Aerodigestive Tract(1)	140(2), 91(2), 155(1), 137(1), 69(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-B	Q31610	4	N	Y	Upper Aerodigestive Tract(1)	91(2), 201(1), 202(1), 155(1), 140(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-B	Q31612	4	N	Y	Upper Aerodigestive Tract(1)	91(2), 155(1), 195(1), 197(1), 176(1)	Bone	Q94K(1), A93T(1), Y91F(1)	25496518	<6.67E-05
HLA-DRB1	P01911	4	N	N	Upper Aerodigestive Tract(1)	135(2), 195(2), 127(2), 133(2), 197(1)	Pancreas	K127Q(5), S133A(2)	23912084	<6.67E-05
HLA-DRB1	Q9GIY3	4	N	N	Upper Aerodigestive	195(2), 11(1), 60(1), 12(1), 59(1)	Pancreas	K127Q(5), S133A(2)	23912084	<6.67E-05

Gene	UniProt	# Studies	CGC?	Mutsig?	Most common		Most Significant Cluster			
					Tissue	Mutations	Tissue	Cluster	PMID	*P*-value
					Tract(1)					
HS3ST4	Q9Y661	4	N	N	Large Intestine(2)	200(1), 313(1), 301(1), 218(1), 411(1)	Lung	T299M(2), K411N(2), T301I(2)	22980975	<6.67E-05
HTR4	Q13639	4	N	N	Large Intestine(2)	302(1), 306(1), 77(1), 183(1), 242(1)	Large Intestine	Y302C(3), G306R(2)	22810696	<6.67E-05
INSRR	P14616	4	N	N	Skin(2)	995(1), 1164(1), 1138(1), 1160(1), 1171(1)	Skin	S1163F(2), T1171P(2)	22842228	<6.67E-05
IPO11	Q9UI26	4	N	N	Large Intestine(2)	797(1), 587(1), 835(1), 337(1), 589(1)	Bone	R835Q(2), R797Q(2)	25186949	<6.67E-05
MCM7	P33993	4	N	N	Large Intestine(3)	415(1), 611(1), 455(1), 445(1), 356(1)	Large Intestine	L356F(2), R611H(2)	22895193	<6.67E-05
MSRB3	Q8IXL7	4	N	N	Large Intestine(2)	184(1), 53(1), 77(1), 63(1), 71(1)	Large Intestine	S161L(2), H77D(2), F71I(2)	22810696	<6.67E-05
NCK2	O43639	4	N	N	Large Intestine(2)	39(1), 205(1), 213(1), 40(1), 273(1)	Large Intestine	D257E(6), I225T(2)	24755471	<6.67E-05
NRXN1	P58400	4	N	N	Skin(2)	181(2), 109(1), 115(1), 132(1), 116(1)	Large Intestine	G225D(4), R132Q(4)	22895193	<6.67E-05
OPRM1	P35372	4	N	N	Large Intestine(2)	347(1), 322(1), 63(1), 46(1), 259(1)	Skin	R369C(3), R347Q(3)	21984974	<6.67E-05
PDE1C	Q14123	4	N	N	Skin(2)	280(1), 468(1), 443(1), 454(1), 176(1)	Lung	R522T(2), K524T(2)	22980975	0.000606
PFKP	Q01813	4	N	N	Large Intestine(2)	467(1), 463(1), 60(1), 177(1), 130(1)	Lung	W463C(2), G467A(2)	22980975	<6.67E-05
PLCB1	Q9NQ66	4	N	N	Skin(2)	720(2), 740(1), 569(1), 768(1), 696(1)	Large Intestine	I767N(3), E302G(2)	24755471	0.00121
POLD1	P28340	4	N	N	Oesophagus(1)	101(1), 339(1), 697(1), 455(1), 306(1)	Lung	I101T(2), V455L(2), G422C(2)	22980975	<6.67E-05
POT1	Q9NUX5	4	Y	N	Haematopoietic And Lymphoid Tissue(2)	105(2), 36(1), 77(1), 116(1), 137(1)	Liver	T105M(2), P116S(2)	25822088	<6.67E-05
PRSS3	P35030	4	N	N	Skin(2)	146(2), 265(2), 148(2), 270(2), 179(2)	Haematopoietic And Lymphoid Tissue	T86I(4), K216Q(4)	24970810	<6.67E-05
PSMB11	A5LHX3	4	N	N	Large Intestine(3)	206(2), 215(1), 65(1), 169(1), 106(1)	Large Intestine	R169C(3), F161V(3)	25344691	<6.67E-05
PTPRD	P23468	4	N	N	Skin(2)	1898(2), 1647(1), 1828(1), 49(1), 1843(1)	Oesophagus	L1377V(4), N1388K(4), A1387T(4)	23525077	0.00662
SEMA3C	Q99985	4	N	N	Large Intestine(3)	558(2), 522(2), 528(2), 526(1), 471(1)	Large Intestine	A522T(4), G558E(2)	24755471	<6.67E-05
SKIV2L	Q15477	4	N	N	Large Intestine(3)	752(1), 751(1), 769(1), 745(1), 754(1)	Large Intestine	L754F(1), I751M(1), L752P(1)	22810696	0.0002
SLC6A2	P23975	4	N	N	Large Intestine(2)	318(1), 140(1), 442(1), 562(1), 148(1)	Large Intestine	A145T(2), A562T(2)	22810696	<6.67E-05
SPAM1	P38567	4	N	N	Large Intestine(2)	207(1), 415(1), 132(1), 416(1), 130(1)	Kidney	T416P(2), F415L(2)	25401301	<6.67E-05
TGFBR2	P37173	4	N	Y	Large Intestine(3)	452(2), 454(2), 446(2), 524(2), 528(2)	Large Intestine	L452P(2), L454P(2)	23856246	<6.67E-05
TMPRSS3	P57727	4	N	N	Large Intestine(1)	400(1), 307(1), 404(1), 417(1), 326(1)	Skin	L307F(3), G417R(2)	25303977	<6.67E-05
TPO	P07202	4	N	N	Large Intestine(1)	153(1), 152(1), 335(1), 613(1), 461(1)	Lung	I613L(2), A640D(2)	22696596	<6.67E-05
TTN	Q8WZ42	4	N	N	Large Intestine(2)	32451(2), 32459(1), 2082(1), 32452(1), 2080(1)	Large Intestine	E32425G(12), K32459E(8), Y32452H(8)	24755471	0.00662

Clustering was performed using default parameters (see Supplementary Note 3), with a *P*-value cutoff of 0.01.

For each gene, reported are: the UniProt protein product, number of WGS studies in which clusters for the gene were identified, whether (Y) or not (N) the gene is in the Cancer Gene Census, whether (Y) or not (N) the gene was identified by MutSig, the primary tissue in which most clusters were found (and the number of studies in which clusters were found in that tissue), the top 5 most commonly mutated amino acid positions (and the number of studies in which mutations were found), and the most significant cluster found. The most significant cluster contains the following associated information: the tissue in which the cluster was found, the mutations in the cluster, the PMID of the publication reporting the mutations, and the *P*-value of the cluster.

Note: a *P*-value of <6.67E-05 indicates that bootstrapping produced no random mutation arrangements as tightly clustered as the observed data.