Full title:
Prominent features of the amino acid mutation landscape in cancer

Short title:

R>H and E>K mutations are prominent features in many cancers

Zachary A. Szpiech[1,*], Nicolas B. Strauli[1,2], Katharine A. White[3], Diego Garrido Ruiz[4], Matthew P. Jacobson[1,4], Diane L. Barber[3], Ryan D. Hernandez[1,5,6,*]

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco.

[2]Biomedical Sciences Graduate Program, University of California, San Francisco.

[3]Department of Cell and Tissue Biology, University of California, San Francisco.

[4]Department of Pharmaceutical Chemistry, University of California, San Francisco.

[5]Quantitative Biosciences Institute, University of California, San Francisco.

[6]Institute for Human Genetics, University of California, San Francisco.

*Corresponding authors

Ryan D Hernandez, PhD
UCSF MC 2530
Byers Hall Room 308
1700 4th Street
San Francisco, CA 94143
Email: ryan.hernandez@ucsf.edu

Zachary A Szpiech, PhD
UCSF MC 2530
Byers Hall Room 308
1700 4th Street
San Francisco, CA 94143
Email: zachary.szpiech@ucsf.edu

ABSTRACT

Cancer can be viewed as a set of different diseases with distinctions based on tissue origin, driver mutations, and genetic signatures. Accordingly, each of these distinctions have been used to classify cancer subtypes and to reveal common features. Here, we present a different analysis of cancer based on amino acid mutation signatures. Non-negative Matrix Factorization and principal component analysis of 29 cancers revealed six amino acid mutation signatures, including four signatures that were dominated by either arginine to histidine (Arg>His) or glutamate to lysine (Glu>Lys) mutations. Sample-level analyses reveal that while some cancers are heterogeneous, others are largely dominated by one type of mutation. Using a non-overlapping set of samples from the COSMIC somatic mutation database, we validate five of six mutation signatures, including signatures with prominent arginine to histidine (Arg>His) or glutamate to lysine (Glu>Lys) mutations. This suggests that our classification of cancers based on amino acid mutation patterns may provide avenues of inquiry pertaining to specific protein mutations that may generate novel insights into cancer biology.

INTRODUCTION

Cancers have been described as open, complex, and adaptive systems [1]. Reflecting this, cancer progression is determined in part by genetic diversification and clonal selection within complex tissue landscapes and with changing tumor properties and microenvironment features [2, 3]. Genetic sequencing of tumor samples has been critical in developing the evolutionary theory of cancer. While cancers traditionally have been—and continue to be—classified by tissue of origin, genetic sequencing has allowed

60 for classification based on driver mutations [4] or nucleotide mutation signatures [5].
61 However, cancer cell adaptation is mediated by changes at the protein level that alter cell
62 biology and enable cancer cell behaviors such as increased proliferation and cell survival.
63 Existing cancer classifications by nucleotide mutation signatures lack a link between the
64 underlying genetic landscape and effects on cancer cell phenotypes. Analysis of cancers
65 by amino acid mutations could provide important connections between cancer evolution
66 and adaptive biological phenotypes as well as provide insight into how specific classes of
67 amino acid mutations may generally alter the function of the proteins in which they are
68 found. There have been some studies to examine amino acid mutations across cancers [6-
69 8], but these have relied on simple mutation counting methods.

70 Here we take a machine-learning approach to analyze amino acid mutations
71 across 29 cancers in order to identify characteristic amino acid mutation signatures. Our
72 analyses reveal that some cancer types have mutation signatures dominated by arginine to
73 histidine (Arg>His) mutations, some have signatures dominated by glutamate to lysine
74 (Glu>Lys), and others have more complex signatures that lack a single dominant amino
75 acid mutation. These signatures were further validated in a non-overlapping set of
76 samples from the COSMIC somatic mutation database. Importantly, this approach
77 identifies not only which amino acid mutations are prevalent among cancers but also
78 which amino acid mutations tend to occur together. For example, cancers with strong
79 Arg>His signatures will also frequently have many Ala>Thr mutations but are unlikely to
80 have many Glu>Lys mutations (despite all of these amino acid transitions resulting from
81 a G>A nucleotide mutation).

82 **RESULTS**

83 **Several cancers are enriched for R>H and E>K amino acid mutations**

84 Multiple studies have interrogated nucleotide mutation biases by analyzing
85 somatic variation across a wide range of cancers [4, 5]. However, in protein coding
86 regions of the genome (i.e. the exome), it is essential to study patterns of amino acid
87 variation to reveal information about potential functional effects at the protein level. We
88 characterized the global properties of amino acid mutations encoded by somatic
89 mutations across a range of cancers by analyzing a tumor-normal paired mutation
90 database [5] consisting of 6,931 samples across 29 cancer types. We applied filtering to
91 remove sequencing artifacts and restricted mutation data to nonsynonymous amino acid
92 mutations (see Materials and Methods, S1 Table and S2 Table for details).

93 Using this amino acid mutation database, we performed an unbiased
94 characterization of mutation signatures across cancer types using Non-negative Matrix
95 Factorization (NMF), which has proven to be a useful tool for pattern discovery in cancer
96 tissue mutation datasets [5] and other biological systems [9]. Applying NMF to the
97 pooled mutation data reveals six mutation signatures at the amino acid level (S1G Fig),
98 including two with strong Arg>His components and two with strong Glu>Lys
99 components (Fig 1A, S1 Fig). Although the cancers are comprised of a mixture of the
100 signatures identified, ten cancers (AML, colorectal, esophageal, low grade glioma,
101 kidney chromophobe, medulloblastoma, pancreatic, prostate, stomach, and uterine) have
102 majority contributions from Arg>His-prominent mutation signatures (R>H and
103 A>T/R>H). We also identify four cancers (bladder, cervix, head and neck, and
104 melanoma) that have majority contributions from Glu>Lys-prominent mutation
105 signatures (E>K and E>K/E>Q). Additionally, there are two complex signatures not

106 dominated by any particular amino acid mutation. Glioblastoma, kidney papillary, liver,
107 and thyroid cancers have majority contribution from the Complex 1 signature, and lung
108 adenocarcinoma, small cell lung, squamous cell lung, and neuroblastoma cancers all have
109 majority contribution from the Complex 2 signature. Finally, seven cancers from a
110 variety of tissues (ALL, breast, CLL, clear cell kidney, B-cell lymphoma, myeloma, and
111 ovarian) have heterogeneous mutation signature contributions.
112
113 **Fig 1. Arg>His and Glu>Lys mutations define mutation signatures of a subset of**
114 **cancers.**
115 (A) Heatmap representation of six-component NMF clustering. Of the six amino acid
116 mutation signatures identified, four have prominent charge-changing mutations: Arg>His
117 (R>H), Glu>Lys (E>K), or Glu>Gln (E>Q). Two complex signatures were also
118 identified. Color scale represents scaled contribution of each signature for a given cancer
119 type. Signature and NMF fit details can be found in S1 Fig. (B) Principal component
120 analysis of nonsynonymous amino acid mutations. PC1 separates cancers with high R>H
121 from cancers with high E>K; PC2 separates cancers with complex signatures. Colors
122 represent the greatest mutation signature contributing to a given cancer. Individual PC
123 loadings can be found in S2 Fig.
124
125 **Visualizing Amino Acid Mutation Properties with Principal Component Analysis**
126 To alternatively visualize the amino acid mutation spectrum, we use principal
127 component analysis to reveal cancers clustering by dominant mutation classes (Fig 1B).
128 We find that PC1 separates Arg>His dominant cancers from Glu>Lys dominant cancers

129 and that PC2 separates cancers with more complex signatures (S2 Fig). This result
130 reinforces our observation that Arg>His and Glu>Lys mutations are characteristic
131 signatures of several cancers.
132
133 **Individual Cancer Samples Recapitulate Amino Acid Mutation Patterns**
134 We also analyze samples individually with NMF and find that Arg>His and
135 Glu>Lys features continue to dominate (Fig 2A and S3 Fig). For many cancer subtypes
136 (melanoma, bladder, uterine, colorectal, low-grade glioma, cervix, neuroblastoma, and
137 the three different lung cancers), individual patients within each
138 cancer exhibit consistent amino acid signatures (Fig 2B). This is true even within
139 clinically diverse cancers such as bladder, uterine, colorectal, and lung cancer, which all
140 have multiple identified driver mutations. This suggests that the amino acid signatures we
141 identified may be independent of underlying driver mutations and may instead be a
142 consequence of common features of the cancer, tumor microenvironment, or selective
143 pressures, all of which may be targeted therapeutically.
144
145 **Fig 2. Amino acid mutation signatures for individual samples.**
146 (A) A heatmap representation of the six-component NMF clustering results for individual
147 cancer samples (only those with >10 total nonsynonymous mutations). Samples with the
148 same maximum signature component were grouped and sorted. Four amino acid
149 mutation signatures identified (R>H, E>K, E>K/E>Q, Complex 2) overlap with
150 signatures in Fig 1A. Color scale represents scaled contribution of each signature for a
151 given sample. Signature and NMF fit details can be found in S3 Fig. (B) Bars show the

152 total fraction of individual samples with a majority of a particular signature within
153 each cancer. Within cancers, a large fraction of individual samples tend to have similar
154 signature components.
155
156 As NMF decomposes a sample into a mixture of characteristic signatures, we can
157 further visualize the normalized mixture coefficients from the individual-level NMF
158 along the three mutation signatures with dominant Arg>His or Glu>Lys components
159 (R>H, E>K, and E>K/E>Q signatures; Fig 3) to determine whether samples tend to be an
160 equal mixture of several signatures or whether they tend to be exclusively composed of a
161 single signature. Indeed, Fig 3 shows a clear separation of samples with a high proportion
162 of Glu>Lys from other signatures.
163
164 **Fig 3. Normalized NMF mixture coefficients for individual samples.**
165 Plot of the normalized mixture coefficients across the three mutation signatures with high
166 R>H or E>K components for every individual sample. Colors represent the greatest
167 contributing mutation signature for each sample based on the full individual-level NMF
168 analysis. Here we see a dramatic separation of samples in the E>K component to the near
169 exclusion of other signatures.
170
171 **Mutation Signature Validation**
172 We validated the NMF signatures with an orthogonal data set (see Materials and
173 Methods) from the COSMIC database [10]. The six mutation signatures identified from
174 COSMIC (S4 Fig) overlap substantially with previously identified signatures (S3 Fig).

We calculated correlation coefficients between all COSMIC Data signatures and each Alexandrov Data signature. When the correlations are very high, this indicates that NMF has identified the same general mutation signature in the two different data sets. Indeed, we found high correlation between the COSMIC signatures and our initially identified signatures for five of the six (Fig 4): R>H, E>K, E>K/E>Q, Complex 2, and Complex A are replicated. The Complex B signature does not replicate as a separate signature, but appears to be largely incorporated into the other complex signatures. Interestingly, a new R>Q/R>W signature is identified as a separate component in the COSMIC data. On inspection of the Alexandrov R>H component we identified, we see that R>Q and R>W are prominent components. Increased sample size in our replicate data set likely enabled NMF to discriminate between these two signatures in COSMIC.

Fig 4. Correlations of COSMIC and Alexandrov mutation signatures. For each COSMIC mutation signature we calculated the correlation with each Alexandrov mutation signature. Five of six Alexandrov mutation signatures replicate in the COSMIC data for k = 6 mutation signatures. Alexandrov Signature R>H is replicated by COSMIC Signature 1, although a subset of mutation types that clustered in the original R>H are identified in the larger COSMIC data set as Signature 2. Alexandrov Signature E>K is replicated by COSMIC Signature 4. Alexandrov Signature E>K, E>Q is replicated by COSMIC Signature 5, although it is also correlated with COSMIC Signature 4 as each of these signatures share mutations. Alexandrov Signature Complex 2 is replicated by COSMIC Signature 6. Alexandrov Signature Complex A is replicated by COSMIC Signature 3. Alexandrov Signature Complex B is not faithfully replicated by any single

198 COSMIC Signature, as the signal appears to be spread amongst COSMIC Signatures 3, 5
199 and 6.
200
201 **DISCUSSION**
202 Proteomic changes can allow cancer cells to adapt to dynamic pressures including
203 changes in matrix composition, oxygen and nutrient availability, intracellular
204 metabolism, as well as increased intracellular pH (pHi), the latter enabling tumorigenic
205 cell behaviors [11-15]. Our analyses reveal that a subset of all possible amino acid
206 mutations dominate the mutation landscape of cancers, with Glu>Lys and Arg>His
207 mutations being the most prominent features of identified mutation signatures.
208 Charge-changing mutations, whether buried or surface-exposed, can alter protein
209 charge, electrostatics, and conformation [16]. Electrostatics of surface residues have been
210 shown to play a key role in protein-protein interactions [17], protein-membrane
211 interactions [18, 19], and kinase substrate recognition [20]. While it is important to note
212 that our analyses are agnostic to the location of the mutation within the proteome and
213 within a protein, the strong bias towards amino acid mutations that alter charge in our
214 identified mutation signatures may suggest an adaptive advantage conferred by these
215 mutations.
216 Glu>Lys mutations swap a negatively charged amino acid for a positively charged
217 amino acid, which may in some cases affect protein function. Indeed, in some cases
218 buried lysine mutations can induce global protein unfolding upon charging that alters
219 mutant protein stability and function [21]. Furthermore, Glu>Lys mutations have been
220 known to affect the function of PIK3CA [22-24].

221 Arg>His mutations swap a positively charged amino acid for a titratable amino
222 acid. Whereas arginine (pKa ~12) should always be protonated, histidine (pKa ~6.5) can
223 titrate within the narrow physiological pH range. Indeed, the pH-sensitive function of
224 many wild-type proteins has been shown to be mediated by titratable histidine residues
225 [25-27]. Moreover, recent work has shown that some Arg>His mutations can confer pH
226 sensitivity to the mutant protein and alter function [28]. We predict that some Arg>His
227 mutations may be adaptive to increased pHi, conferring a gain in pH sensing to the
228 mutant protein.
229 From our analyses, Arg>His mutations define the mutation landscape of a diverse
230 set of cancers across a range of tissues including brain (low-grade glioma), digestive
231 (colorectal), reproductive (uterine), and blood (AML) cancers. Importantly, these cancers
232 do not have overlapping nucleotide mutation signatures [5], which suggests that the
233 amino acid mutation signatures we identified may reflect other aspects of the cancers
234 including distinct physiological pressures, microenvironment features, or functional
235 requirements. Indeed, these results may help inform studies in the emerging field of
236 Molecular Pathologic Epidemiology (MPE) [29, 30], which seeks to integrate knowledge
237 across disciplines to inform personalized approaches to cancer prevention and therapy.
238 Linking amino acid signatures to physiological or pathological features of the cancer
239 could be important for identifying selective pressures that may be driving or sustaining
240 the cancer as well as for limiting disease progression, particularly where targeted
241 approaches fail [31-33].
242
243
244 **MATERIALS AND METHODS**

245
246 **Mutation Dataset Filtering**

247 We validated the dataset [5] by comparing known frequencies of well-studied
248 cancer driver genes with observed frequencies in the dataset. Specifically, BRAF is
249 mutated in 40–50% of melanoma samples, and IDH1 is mutated in 75–85%, low-grade
250 glioma, AML, and glioblastoma samples are mutated 75-85%, 8–12%, and 1–5% of the
251 time, respectively. We used the p53 database (http://p53.fr/index.html) to find expected
252 p53 mutation frequency for various cancers: colorectal, head and neck, pancreatic,
253 stomach, liver, and breast cancer have 43%, 42%, 34%, 32%, 31%, and 22% p53
254 mutation rates, respectively. The observed mutation frequencies were consistently lower
255 than expected for the genes/cancers we assessed, which suggests that the dataset authors
256 [5] were perhaps too stringent in quality control (QC) filtering. Different levels of QC
257 filtering were performed, and we systematically relaxed filters in order to recapitulate the
258 expected mutation frequencies of the selected canonical driver genes. Applying only the
259 'sequencing artifact' QC filter (from [5]) most closely recapitulated expected mutation
260 frequencies for the canonical driver genes, and this filter alone was used for the
261 remainder of the bioinformatics analyses.

262

263 **Mapping somatic SNPs**

264 After filtering we used part of the PolyPhen2 [34] pipeline to map mutations to
265 UCSC Canonical transcripts and restricted to nonsynonymous amino acid changes. The
266 following cancers had reduced sample sizes after filtering and nonsynonymous mutation
267 restriction: AML: one sample eliminated through QC filtering, two samples eliminated
268 because all mutations were synonymous; low grade glioma: one sample eliminated

because after QC filtering all remaining mutations were synonymous; glioblastoma: two samples eliminated because all mutations were synonymous. All Pilocytic Astrocytoma samples were excluded from future analysis due to low total nonsynonymous mutations per sample.

Mutation frequency data sets

For the individual sample data, we represent each sample as a row vector with elements giving the mutation counts observed for each nonsynonymous mutation (e.g. Ala>Cys, Ala>Asp, etc.) and removing all samples with <10 total observed mutations. For the aggregated data set, we sum the mutation counts across all samples of the same cancer type (including samples with <10 mutations), giving one row vector for each cancer type where each element represents the total number of observed nonsynonymous mutations across all samples. For non-negative matrix factorization and principal component analysis, we divide each row by the row sum.

NMF is an unsupervised learning method used to decompose a data matrix into a product of two non-negative matrices representing a set of k signals and mixture coefficients. For example if $\mathbf{X}$ is an $m \times n$ matrix representing the nonsynonymous mutation frequency data, then the NMF of the data is given by

$$X = WH$$

where $\mathbf{W}$ is an $m \times k$ matrix with the k columns representing mutation signatures and $\mathbf{H}$ is a $k \times n$ matrix representing the mixture coefficients that best reconstruct $\mathbf{X}$. Often it is not possible to factor $\mathbf{X}$ exactly, so a typical approach to solving the decomposition will optimize

$$\min_{w,H\geq 0}[D(X,WH)+R(W,H)]$$

where D() is a loss function (often the Frobenius norm or the Kullback-Leibler divergence) and R() is a regularization function. For our NMF analyses, we utilize the R package *NMF* [35] with default choices for D() and R().

Principal Component Analysis (PCA)

PCA is a dimension reducing learning method designed to decompose a data matrix into a set of orthogonal bases defined along the major axes of variation within the data. Here we compute the first two principal components from our mutation frequency matrix **X**. The kth principal component is represented by a vector of loadings, $w_{(k)}$. The first PC is then calculated as

$$w_{(1)} = \arg\max\left\{\frac{w^T X^T X w}{w^T w}\right\}$$

and subsequent PCs are calculated as

$$w_{(k)} = \arg\max\left\{\frac{w^T \widehat{X}_k^T \widehat{X}_k w}{w^T w}\right\}$$

where

$$\widehat{X}_k = X - \sum_{s}^{k-1} X w_{(s)} w_{(s)}^T.$$

We use the R package *prcomp* to perform all PCA analyses.

Validation of NMF Mutation Signatures

In order to validate the mutation signatures that we discovered in our data, we sought an orthogonal data set in which to replicate our analysis. We used the COSMIC

v81 database of somatic mutations [10]. We first filtered all mutations that were not marked as confirmed somatic mutations. Next, as our original data set ("Alexandrov Data") had overlapping samples within the COSMIC database, we excluded all samples that were included in our original analysis. Finally, we excluded samples with fewer than 10 total non-synonymous mutations. This filtering resulted in a final data set of 2,236,176 non-synonymous mutations across 15,868 samples. We named this final data set the "COSMIC Data." We then ran NMF with $k = 6$ signatures on the matrix of individual sample mutation frequencies as described above. Results are shown in S4 Fig. We found that five of the six mutation signatures we originally discovered were replicated in the COSMIC data (Fig 4).

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health grants CA178706 to Diane L. Barber and Ryan D. Hernandez; CA197855 to Diane L. Barber, and HG007644 to Ryan D. Hernandez; and National Institutes of Health F32 grant CA177085 to Katharine A. White.

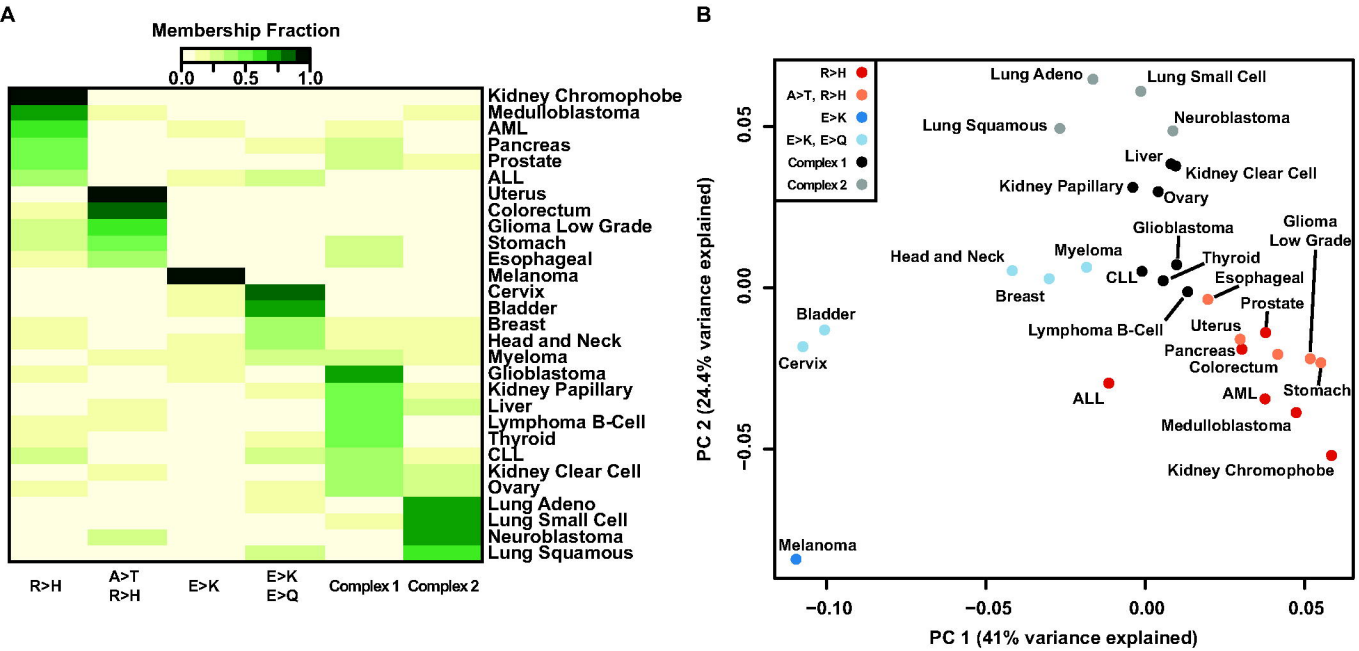
Bibliography

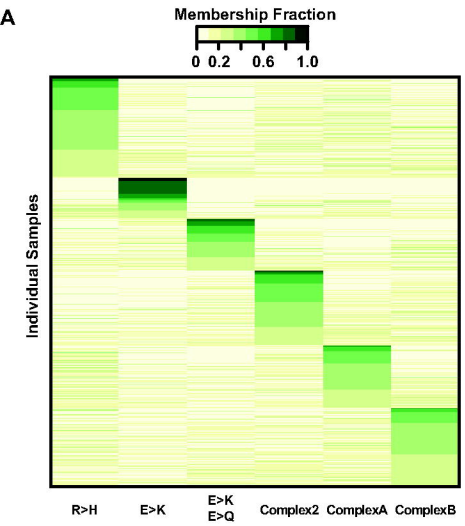
1. Gillies RJ, Gatenby RA. Metabolism and its sequelae in cancer evolution and therapy. Cancer J. 2015;21(2):88-96. doi: 10.1097/PPO.0000000000000102. PubMed PMID: 25815848; PubMed Central PMCID: PMCPMC4446699.
2. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381):306-13. doi: 10.1038/nature10762. PubMed PMID: 22258609; PubMed Central PMCID: PMCPMC3367003.

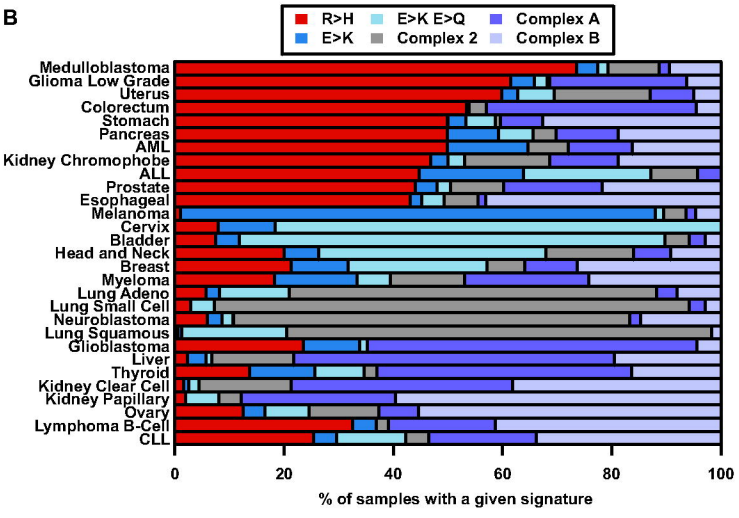
336 3. Nowell PC. The clonal evolution of tumor cell populations. Science.
337 1976;194(4260):23-8. PubMed PMID: 959840.
338 4. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, et al.
339 Signatures of mutation and selection in the cancer genome. Nature. 2010;463(7283):893-
340 8. doi: 10.1038/nature08768. PubMed PMID: 20164919; PubMed Central PMCID:
341 PMCPMC3145113.
342 5. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV,
343 et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-
344 21. doi: 10.1038/nature12477. PubMed PMID: 23945592; PubMed Central PMCID:
345 PMCPMC3776390.
346 6. Tan H, Bao J, Zhou X. Genome-wide mutational spectra analysis reveals
347 significant cancer-specific heterogeneity. Sci Rep. 2015;5:12566. doi:
348 10.1038/srep12566. PubMed PMID: 26212640; PubMed Central PMCID:
349 PMCPMC4515826.
350 7. Anoosha P, Sakthivel R, Michael Gromiha M. Exploring preferred amino acid
351 mutations in cancer genes: Applications to identify potential drug targets. Biochim
352 Biophys Acta. 2016;1862(2):155-65. doi: 10.1016/j.bbadis.2015.11.006. PubMed PMID:
353 26581171.
354 8. Tsuber V, Kadamov Y, Brautigam L, Berglund UW, Helleday T. Mutations in
355 Cancer Cause Gain of Cysteine, Histidine, and Tryptophan at the Expense of a Net Loss
356 of Arginine on the Proteome Level. Biomolecules. 2017;7(3). doi:
357 10.3390/biom7030049. PubMed PMID: 28671612.
358 9. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern
359 discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004;101(12):4164-9.
360 doi: 10.1073/pnas.0308531101. PubMed PMID: 15016911; PubMed Central PMCID:
361 PMCPMC384712.
362 10. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC:
363 somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1):D777-D83.
364 doi: 10.1093/nar/gkw1121. PubMed PMID: 27899578; PubMed Central PMCID:
365 PMCPMC5210583.
366 11. White KA, Grillo-Hill BK, Barber DL. Cancer cell behaviors mediated by
367 dysregulated pH dynamics at a glance. J Cell Sci. 2017;130(4):663-9. doi:
368 10.1242/jcs.195297. PubMed PMID: 28202602; PubMed Central PMCID:
369 PMCPMC5339414.
370 12. Cardone RA, Casavola V, Reshkin SJ. The role of disturbed pH dynamics and the
371 Na+/H+ exchanger in metastasis. Nat Rev Cancer. 2005;5(10):786-95. doi:
372 10.1038/nrc1713. PubMed PMID: 16175178.
373 13. Grillo-Hill BK, Choi C, Jimenez-Vidal M, Barber DL. Increased H(+) efflux is
374 sufficient to induce dysplasia and necessary for viability with oncogene expression. Elife.
375 2015;4. doi: 10.7554/eLife.03270. PubMed PMID: 25793441; PubMed Central PMCID:
376 PMCPMC4392478.
377 14. Parks SK, Chiche J, Pouyssegur J. Disrupting proton dynamics and energy
378 metabolism for cancer therapy. Nat Rev Cancer. 2013;13(9):611-23. doi:
379 10.1038/nrc3579. PubMed PMID: 23969692.

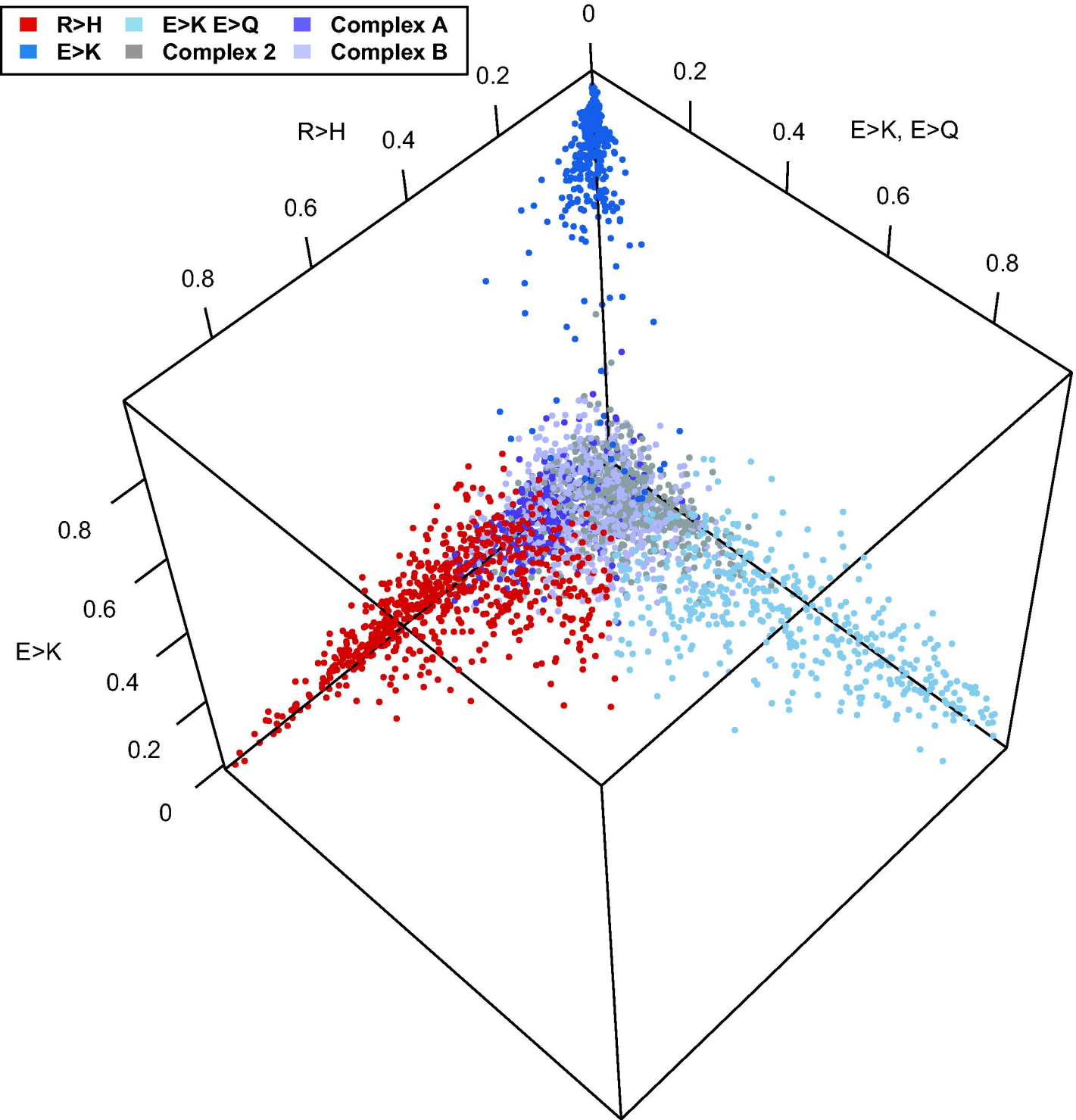
380 15. Webb BA, Chimenti M, Jacobson MP, Barber DL. Dysregulated pH: a perfect
381 storm for cancer progression. Nat Rev Cancer. 2011;11(9):671-7. doi: 10.1038/nrc3110.
382 PubMed PMID: 21833026.
383 16. Zheng Y, Cui Q. Microscopic mechanisms that govern the titration response and
384 pKa values of buried residues in staphylococcal nuclease mutants. Proteins.
385 2017;85(2):268-81. doi: 10.1002/prot.25213. PubMed PMID: 27862310.
386 17. Zhang Z, Witham S, Alexov E. On the role of electrostatics in protein-protein
387 interactions. Phys Biol. 2011;8(3):035001. doi: 10.1088/1478-3975/8/3/035001. PubMed
388 PMID: 21572182; PubMed Central PMCID: PMCPMC3137121.
389 18. Cho W, Stahelin RV. Membrane-protein interactions in cell signaling and
390 membrane trafficking. Annu Rev Biophys Biomol Struct. 2005;34:119-51. doi:
391 10.1146/annurev.biophys.33.110502.133337. PubMed PMID: 15869386.
392 19. Stahelin RV. Lipid binding domains: more than simple lipid effectors. J Lipid
393 Res. 2009;50 Suppl:S299-304. doi: 10.1194/jlr.R800078-JLR200. PubMed PMID:
394 19008549; PubMed Central PMCID: PMCPMC2674730.
395 20. Ubersax JA, Ferrell JE, Jr. Mechanisms of specificity in protein phosphorylation.
396 Nat Rev Mol Cell Biol. 2007;8(7):530-41. doi: 10.1038/nrm2203. PubMed PMID:
397 17585314.
398 21. Chimenti MS, Khangulov VS, Robinson AC, Heroux A, Majumdar A,
399 Schlessman JL, et al. Structural reorganization triggered by charging of Lys residues in
400 the hydrophobic interior of a protein. Structure. 2012;20(6):1071-85. doi:
401 10.1016/j.str.2012.03.023. PubMed PMID: 22632835; PubMed Central PMCID:
402 PMCPMC3373022.
403 22. Carson JD, Van Aller G, Lehr R, Sinnamon RH, Kirkpatrick RB, Auger KR, et al.
404 Effects of oncogenic p110alpha subunit mutations on the lipid kinase activity of
405 phosphoinositide 3-kinase. Biochem J. 2008;409(2):519-24. doi: 10.1042/BJ20070681.
406 PubMed PMID: 17877460.
407 23. Huang CH, Mandelker D, Gabelli SB, Amzel LM. Insights into the oncogenic
408 effects of PIK3CA mutations from the structure of p110alpha/p85alpha. Cell Cycle.
409 2008;7(9):1151-6. doi: 10.4161/cc.7.9.5817. PubMed PMID: 18418043; PubMed Central
410 PMCID: PMCPMC3260475.
411 24. Meyer DS, Koren S, Leroy C, Brinkhaus H, Muller U, Klebba I, et al. Expression
412 of PIK3CA mutant E545K in the mammary gland induces heterogeneous tumors but is
413 less potent than mutant H1047R. Oncogenesis. 2013;2:e74. doi: 10.1038/oncsis.2013.38.
414 PubMed PMID: 24080956; PubMed Central PMCID: PMCPMC3816227.
415 25. Choi CH, Webb BA, Chimenti MS, Jacobson MP, Barber DL. pH sensing by
416 FAK-His58 regulates focal adhesion remodeling. J Cell Biol. 2013;202(6):849-59. doi:
417 10.1083/jcb.201302131. PubMed PMID: 24043700; PubMed Central PMCID:
418 PMCPMC3776353.
419 26. Frantz C, Barreiro G, Dominguez L, Chen X, Eddy R, Condeelis J, et al. Cofilin is
420 a pH sensor for actin free barbed end formation: role of phosphoinositide binding. J Cell
421 Biol. 2008;183(5):865-79. doi: 10.1083/jcb.200804161. PubMed PMID: 19029335;
422 PubMed Central PMCID: PMCPMC2592832.
423 27. Webb BA, White KA, Grillo-Hill BK, Schonichen A, Choi C, Barber DL. A
424 Histidine Cluster in the Cytoplasmic Domain of the Na-H Exchanger NHE1 Confers pH-
425 sensitive Phospholipid Binding and Regulates Transporter Activity. J Biol Chem.

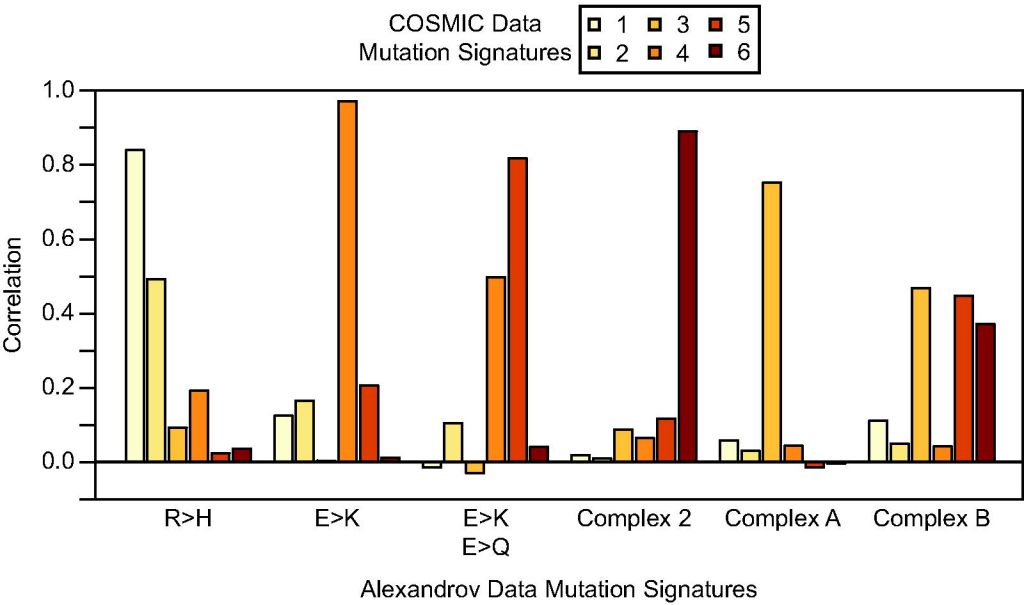
426 2016;291(46):24096-104. doi: 10.1074/jbc.M116.736215. PubMed PMID: 27650500;
427 PubMed Central PMCID: PMCPMC5104935.
428 28. DiGiammarino EL, Lee AS, Cadwell C, Zhang W, Bothner B, Ribeiro RC, et al.
429 A novel mechanism of tumorigenesis involving pH-dependent destabilization of a mutant
430 p53 tetramer. Nat Struct Biol. 2002;9(1):12-6. doi: 10.1038/nsb730. PubMed PMID:
431 11753428.
432 29. Ogino S, Chan AT, Fuchs CS, Giovannucci E. Molecular pathological
433 epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary
434 field. Gut. 2011;60(3):397-411. doi: 10.1136/gut.2010.217182. PubMed PMID:
435 21036793; PubMed Central PMCID: PMCPMC3040598.
436 30. Nishi A, Milner DA, Jr., Giovannucci EL, Nishihara R, Tan AS, Kawachi I, et al.
437 Integration of molecular pathology, epidemiology and social science for global precision
438 medicine. Expert Rev Mol Diagn. 2016;16(1):11-23. doi:
439 10.1586/14737159.2016.1115346. PubMed PMID: 26636627; PubMed Central PMCID:
440 PMCPMC4713314.
441 31. Alfarouk KO, Stock CM, Taylor S, Walsh M, Muddathir AK, Verduzco D, et al.
442 Resistance to cancer chemotherapy: failure in drug response from ADME to P-gp. Cancer
443 Cell Int. 2015;15:71. doi: 10.1186/s12935-015-0221-1. PubMed PMID: 26180516;
444 PubMed Central PMCID: PMCPMC4502609.
445 32. Alfarouk KO, Verduzco D, Rauch C, Muddathir AK, Adil HH, Elhassan GO, et
446 al. Glycolysis, tumor metabolism, cancer growth and dissemination. A new pH-based
447 etiopathogenic perspective and therapeutic approach to an old cancer question.
448 Oncoscience. 2014;1(12):777-802. doi: 10.18632/oncoscience.109. PubMed PMID:
449 25621294; PubMed Central PMCID: PMCPMC4303887.
450 33. Gillies RJ, Verduzco D, Gatenby RA. Evolutionary dynamics of carcinogenesis
451 and why targeted therapy does not work. Nat Rev Cancer. 2012;12(7):487-93. doi:
452 10.1038/nrc3298. PubMed PMID: 22695393; PubMed Central PMCID:
453 PMCPMC4122506.
454 34. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al.
455 A method and server for predicting damaging missense mutations. Nat Methods.
456 2010;7(4):248-9. doi: 10.1038/nmeth0410-248. PubMed PMID: 20354512; PubMed
457 Central PMCID: PMCPMC2855889.
458 35. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization.
459 BMC Bioinformatics. 2010;11:367. doi: 10.1186/1471-2105-11-367. PubMed PMID:
460 20598126; PubMed Central PMCID: PMCPMC2912887.
461


A
Membership Fraction
0.0
0.5
1.0
Kidney Chromophobe
Medulloblastoma
AML
Pancreas
Prostate
ALL
Uterus
Colorectum
Glioma Low Grade
Stomach
Esophageal
Melanoma
Cervix
Bladder
Breast
Head and Neck
Myeloma
Glioblastoma
Kidney Papillary
Liver
Lymphoma B-Cell
Thyroid
CLL
Kidney Clear Cell
Ovary
Lung Adeno
Lung Small Cell
Neuroblastoma
Lung Squamous
R>H
A>T
R>H
E>K
E>K
E>Q
Complex 1
Complex 2
B
R>H
A>T, R>H
E>K
E>K, E>Q
Complex 1
Complex 2
Lung Adeno
Lung Small Cell
Lung Squamous
Neuroblastoma
Liver
Kidney Clear Cell
Kidney Papillary
Ovary
Glioblastoma
Glioma
Low Grade
Head and Neck
Myeloma
Thyroid
CLL
Esophageal
Breast
Prostate
Bladder
Lymphoma B-Cell
Uterus
Pancreas
Cervix
Colorectum
ALL
AML
Stomach
Medulloblastoma
Kidney Chromophobe
Melanoma
0.05
0.00
−0.05
PC 2 (24.4% variance explained)
−0.10
−0.05
0.00
0.05
PC 1 (41% variance explained)


A
Membership Fraction
0 0.2 0.6 1.0
Individual Samples
R>H
E>K
E>K
E>Q
Complex2
ComplexA
ComplexB


B
R>H
E>K
E>K E>Q
Complex 2
Complex A
Complex B
Medulloblastoma
Glioma Low Grade
Uterus
Colorectum
Stomach
Pancreas
AML
Kidney Chromophobe
ALL
Prostate
Esophageal
Melanoma
Cervix
Bladder
Head and Neck
Breast
Myeloma
Lung Adeno
Lung Small Cell
Neuroblastoma
Lung Squamous
Glioblastoma
Liver
Thyroid
Kidney Clear Cell
Kidney Papillary
Ovary
Lymphoma B-Cell
CLL
0
20
40
60
80
100
% of samples with a given signature


R>H
E>K
E>K E>Q
Complex 2
Complex A
Complex B
R>H
E>K, E>Q
E>K
0
0.2
0.4
0.6
0.8


COSMIC Data
Mutation Signatures
1
2
3
4
5
6
1.0
0.8
0.6
0.4
0.2
0.0
Correlation
R>H
E>K
E>K
E>Q
Complex 2
Complex A
Complex B
Alexandrov Data Mutation Signatures