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

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

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

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

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

[5]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 has 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 with either prominent 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. 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 critically important 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 for classification based on driver mutations [4] or nucleotide mutation signatures [5]. However, cancer cell adaptation is mediated by changes at the protein level that alter cell biology and enable cancer cell behaviors such as increased proliferation and cell survival. Existing cancer classifications by nucleotide mutation

signatures lack a link between the underlying genetic landscape and effects on cancer cell phenotypes. Analysis of cancers by amino acid mutations could provide important connections between cancer evolution and adaptive biological phenotypes as well as provide insight into how specific classes of amino acid mutations may generally alter the function of the proteins in which they are found. There have been some studies to examine amino acid mutations across cancers [6, 7], but these have relied on simple mutation counting methods.

Here we take a machine-learning approach to analyze amino acid mutations across 29 cancers in order to identify characteristic amino acid mutation signatures. Our analyses reveal that some cancer types have mutation signatures dominated by arginine to histidine (Arg>His) mutations, some have signatures dominated by glutamate to lysine (Glu>Lys), and others have more complex signatures that lack a single dominant amino acid mutation. Importantly, this approach identifies not only which amino acid mutations are prevalent among cancers but also which amino acid mutations tend to occur together. For example, cancers with strong Arg>His signatures will also frequently have many Ala>Thr mutations but are unlikely to have many Glu>Lys mutations (despite all of these amino acid transitions resulting from a G>A nucleotide mutation).

RESULTS

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

Multiple studies have interrogated nucleotide mutation biases by analyzing somatic variation across a wide range of cancers [4, 5]. However, in protein coding regions of the genome (i.e. the exome), it is essential to study patterns of amino acid variation to reveal information about potential functional effects at the protein level. We

characterized the global properties of amino acid mutations encoded by somatic mutations across a range of cancers by analyzing a tumor-normal paired mutation database [5] consisting of 6,931 samples across 29 cancer types. We applied filtering to remove sequencing artifacts and restricted mutation data to nonsynonymous amino acid mutations (see Methods, Tables S1and S2 for details).

Using this amino acid mutation database, we performed an unbiased characterization of mutation signatures across cancer types using Non-negative Matrix Factorization (NMF), which has proven to be a useful tool for pattern discovery in cancer tissue mutation datasets [5] and other biological systems [8]. Applying NMF to the pooled mutation data reveals six mutation signatures at the amino acid level (Fig S1G), including two with strong Arg>His components and two with strong Glu>Lys components (Fig 1A, Fig S1). Although the cancers are comprised of a mixture of the signatures identified, ten cancers (AML, colorectal, esophageal, low grade glioma, kidney chromophobe, medulloblastoma, pancreatic, prostate, stomach, and uterine) have majority contributions from Arg>His-prominent mutation signatures (R>H and A>T/R>H). We also identify four cancers (bladder, cervix, head and neck, and melanoma) that have majority contributions from Glu>Lys-prominent mutation signatures (E>K and E>K/E>Q). Additionally, there are two complex signatures not dominated by any particular amino acid mutation. Glioblastoma, kidney papillary, liver, and thyroid cancers have majority contribution from the Complex 1 signature, and lung adenocarcinoma, small cell lung, squamous cell lung, and neuroblastoma cancers all have majority contribution from the Complex 2 signature. Finally, seven cancers from a

variety of tissues (ALL, breast, CLL, clear cell kidney, B-cell lymphoma, myeloma, and ovarian) have heterogeneous mutation signature contributions.

Fig 1. Arg>His and Glu>Lys mutations define mutation signatures of a subset of cancers.

(A) Heatmap representation of six-component NMF clustering. Of the six amino acid mutation signatures identified, four have prominent charge-changing mutations: Arg>His (R>H); Ala>Thr (A>T) and R>H; Glu>Lys (E>K); and E>K and Glu>Gln (E>Q). Two complex signatures were also identified. Color scale represents scaled contribution of each signature for a given cancer type. Signature and NMF fit details can be found in Fig S1. (B) Principal component analysis of nonsynonymous amino acid mutations. PC1 separates cancers with high R>H from cancers with high E>K; PC2 separates cancers with complex signatures. Individual PC loadings can be found in Fig S2.

Visualizing Amino Acid Mutation Properties with Principal Component Analysis

To alternatively visualize the amino acid mutation spectrum, we use principal component analysis to reveal cancers clustering by dominant mutation classes (Fig 1B). We find that PC1 separates Arg>His dominant cancers from Glu>Lys dominant cancers and that PC2 separates cancers with more complex signatures (Fig S2). This result reinforces our observation that Arg>His and Glu>Lys mutations are characteristic signatures of several cancers.

Individual Cancer Samples Recapitulate Amino Acid Mutation Patterns

We also analyze samples individually with NMF and find that Arg>His and Glu>Lys features continue to dominate (Figs 2A and S3A). For many cancer subtypes (melanoma, bladder, uterine, colorectal, low-grade glioma, cervix, neuroblastoma, and the three different lung cancers), individual patients within each cancer exhibit consistent amino acid signatures (Fig 2B). This is true even within clinically diverse cancers such as bladder, uterine, colorectal, and lung cancer, which all have multiple identified driver mutations. This suggests that the amino acid signatures we identified may be independent of underlying driver mutation and may instead be reporting on common features of the cancer, tumor microenvironment, or selective pressures, which may be targeted therapeutically.

Fig 2. Amino acid mutation signatures for individual samples.
(A) A heatmap representation of the six-component NMF clustering results for individual cancer samples (only those with >10 total nonsynonymous mutations). Samples with the same maximum signature component were grouped and sorted. Four amino acid mutation signatures identified (R>H, E>K, E>K/E>Q, Complex 2) overlap with signatures in Fig 1A. Color scale represents scaled contribution of each signature for a given sample. Signature and NMF fit details can be found in Fig S3. (B) Bars show the total fraction of individual samples with a majority of a particular signature within each cancer. Within cancers, a large fraction of individual samples tend to have similar signature components.

As NMF decomposes a sample into a mixture of characteristic signatures, we can further visualize the normalized mixture coefficients from the individual-level NMF along the three mutation signatures with dominant Arg>His or Glu>Lys components (R>H, E>K, and E>K/E>Q signatures; Fig 3) to determine whether samples tend to be an equal mixture of several signatures or whether they tend to be exclusively composed of a single signature. Indeed, Fig 3 shows a clear separation of samples with a high proportion of Glu>Lys from other signatures.

Fig 3. Normalized NMF mixture coefficients for individual samples.
Plot of the normalized mixture coefficients across the three mutation signatures with high R>H or E>K components for every individual sample. Colors represent the major component for each sample based on the full individual-level NMF analysis. Here we see a dramatic separation of samples in the E>K component to the near exclusion of other signatures.

DISCUSSION

Our analyses reveal that a subset of all possible amino acid mutations dominate the mutation landscape of cancers, with Glu>Lys and Arg>His mutations being the most prominent features of identified mutation signatures. Proteomic changes can allow cancer cells to adapt to dynamic pressures including changes in matrix composition, oxygen and nutrient availability, intracellular metabolism, as well as increased intracellular pH (pHi), the latter enabling tumorigenic cell behaviors [9-13]. The strong bias towards amino acid mutations that alter charge in our identified mutation signatures may suggest an adaptive

advantage to one or more of these pressures. Glu>Lys mutations swap a negatively charged amino acid for a positively charged amino acid, which may in some cases effect protein function. Furthermore, whereas arginine (pKa ~12) should always be protonated, histidine (pKa ~6.5) can titrate within the narrow physiological pH range. Indeed, the pH-sensitive function of many wild-type proteins has been shown to be mediated by titratable histidine residues [14-16]. Therefore, while both of these mutation classes could substantially affect protein function, we predict that some Arg>His mutations may be adaptive to increased pHi, conferring a gain in pH sensing to the mutant protein.

Given the potential adaptive gain in pH sensing that Arg>His mutations could confer, it is interesting to note that Arg>His mutations define the mutation landscape of a diverse set of cancers across a range of tissues including brain (low-grade glioma), digestive (colorectal), reproductive (uterine), and blood (AML) cancers. These cancers do not have overlapping nucleotide mutation signatures [5], which suggests that the amino acid mutation signatures we identified may reflect other aspects of the cancers including distinct physiological pressures, microenvironment features, or functional requirements that could be important for limiting disease progression, particularly where targeted approaches fail [17-19].

MATERIALS AND METHODS

Mutation Dataset Filtering

We validated the dataset [5] by comparing known frequencies of well-studied cancer driver genes with observed frequencies in the dataset. Specifically, BRAF is mutated in 40–50% of melanoma samples, and IDH1 is mutated in 75–85%, low-grade glioma, AML, and glioblastoma samples are mutated 75-85%, 8–12%, and 1–5% of the

time, respectively. We used the p53 database (http://p53.fr/index.html) to find expected p53 mutation frequency for various cancers: colorectal, head and neck, pancreatic, stomach, liver, and breast cancer have 43%, 42%, 34%, 32%, 31%, and 22% p53 mutation rates, respectively. The observed mutation frequencies were consistently lower than expected for the genes/cancers we assessed, which suggests that the dataset authors [5] were perhaps too stringent in quality control (QC) filtering. Different levels of QC filtering were performed, and we systematically relaxed filters in order to recapitulate the expected mutation frequencies of the selected canonical driver genes. Applying only the 'sequencing artifact' QC filter (from [5]) most closely recapitulated expected mutation frequencies for the canonical driver genes, and this filter alone was used for the remainder of the bioinformatics analyses.

Mapping somatic SNPs

After filtering we used part of the PolyPhen2 [20] pipeline to map mutations to UCSC Canonical transcripts and restricted to nonsynonymous amino acid changes. The following cancers had reduced sample sizes after filtering and nonsynonymous mutation restriction: AML: one sample eliminated through QC filtering, two samples eliminated 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 **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 **W** is an $m \times k$ matrix with the k columns representing mutation signatures and **H** is a $k \times n$ matrix representing the mixture coefficients that best reconstruct **X**. Often it is not possible to factor **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* [21] 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 k[th] 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

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

where

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

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

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.
3. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8. PubMed PMID: 959840.
4. Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, et al. Signatures of mutation and selection in the cancer genome. Nature. 2010;463(7283):893-8. doi: 10.1038/nature08768. PubMed PMID: 20164919; PubMed Central PMCID: PMCPMC3145113.
5. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-21. doi: 10.1038/nature12477. PubMed PMID: 23945592; PubMed Central PMCID: PMCPMC3776390.
6. Tan H, Bao J, Zhou X. Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. Sci Rep. 2015;5:12566. doi: 10.1038/srep12566. PubMed PMID: 26212640; PubMed Central PMCID: PMCPMC4515826.
7. Anoosha P, Sakthivel R, Michael Gromiha M. Exploring preferred amino acid mutations in cancer genes: Applications to identify potential drug targets. Biochim Biophys Acta. 2016;1862(2):155-65. doi: 10.1016/j.bbadis.2015.11.006. PubMed PMID: 26581171.
8. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004;101(12):4164-9. doi: 10.1073/pnas.0308531101. PubMed PMID: 15016911; PubMed Central PMCID: PMCPMC384712.
9. White KA, Grillo-Hill BK, Barber DL. Cancer cell behaviors mediated by dysregulated pH dynamics at a glance. J Cell Sci. 2017;130(4):663-9. doi: 10.1242/jcs.195297. PubMed PMID: 28202602; PubMed Central PMCID: PMCPMC5339414.
10. Cardone RA, Casavola V, Reshkin SJ. The role of disturbed pH dynamics and the Na+/H+ exchanger in metastasis. Nat Rev Cancer. 2005;5(10):786-95. doi: 10.1038/nrc1713. PubMed PMID: 16175178.
11. Grillo-Hill BK, Choi C, Jimenez-Vidal M, Barber DL. Increased H(+) efflux is sufficient to induce dysplasia and necessary for viability with oncogene expression. Elife. 2015;4. doi: 10.7554/eLife.03270. PubMed PMID: 25793441; PubMed Central PMCID: PMCPMC4392478.
12. Parks SK, Chiche J, Pouyssegur J. Disrupting proton dynamics and energy metabolism for cancer therapy. Nat Rev Cancer. 2013;13(9):611-23. doi: 10.1038/nrc3579. PubMed PMID: 23969692.

13. Webb BA, Chimenti M, Jacobson MP, Barber DL. Dysregulated pH: a perfect storm for cancer progression. Nat Rev Cancer. 2011;11(9):671-7. doi: 10.1038/nrc3110. PubMed PMID: 21833026.

14. Choi CH, Webb BA, Chimenti MS, Jacobson MP, Barber DL. pH sensing by FAK-His58 regulates focal adhesion remodeling. J Cell Biol. 2013;202(6):849-59. doi: 10.1083/jcb.201302131. PubMed PMID: 24043700; PubMed Central PMCID: PMCPMC3776353.

15. Frantz C, Barreiro G, Dominguez L, Chen X, Eddy R, Condeelis J, et al. Cofilin is a pH sensor for actin free barbed end formation: role of phosphoinositide binding. J Cell Biol. 2008;183(5):865-79. doi: 10.1083/jcb.200804161. PubMed PMID: 19029335; PubMed Central PMCID: PMCPMC2592832.

16. Webb BA, White KA, Grillo-Hill BK, Schonichen A, Choi C, Barber DL. A Histidine Cluster in the Cytoplasmic Domain of the Na-H Exchanger NHE1 Confers pH-sensitive Phospholipid Binding and Regulates Transporter Activity. J Biol Chem. 2016;291(46):24096-104. doi: 10.1074/jbc.M116.736215. PubMed PMID: 27650500; PubMed Central PMCID: PMCPMC5104935.

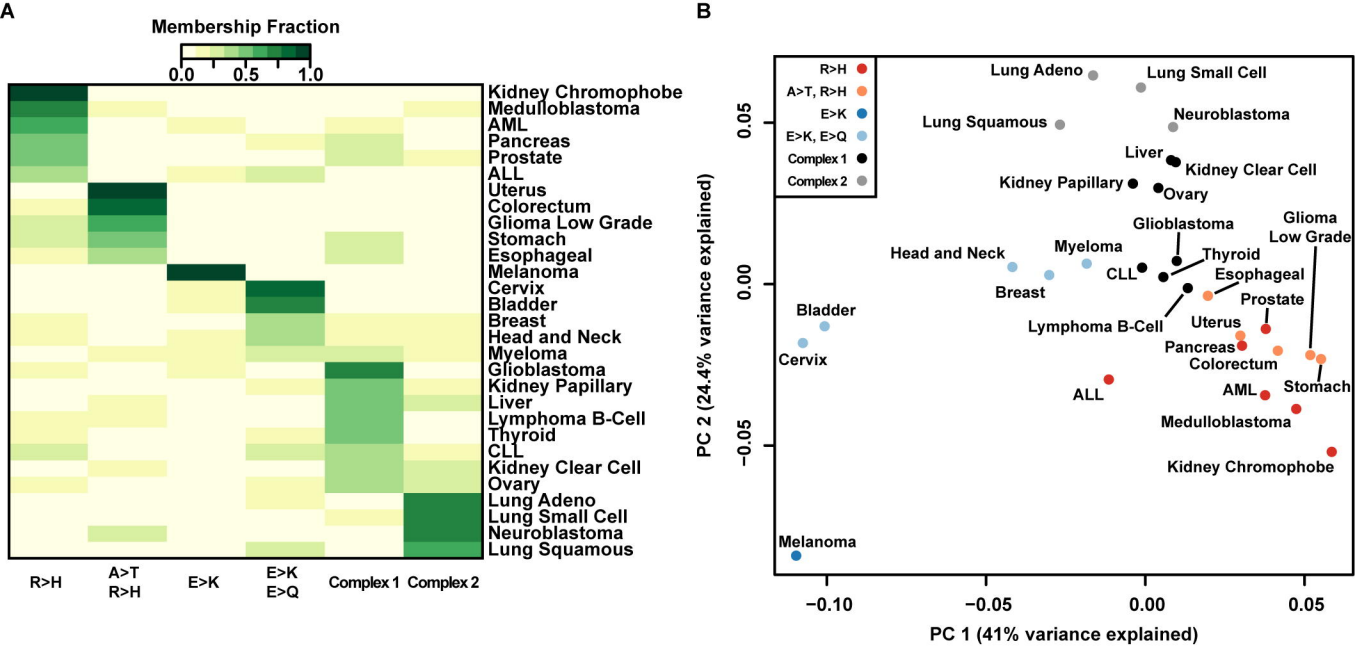
17. Alfarouk KO, Stock CM, Taylor S, Walsh M, Muddathir AK, Verduzco D, et al. Resistance to cancer chemotherapy: failure in drug response from ADME to P-gp. Cancer Cell Int. 2015;15:71. doi: 10.1186/s12935-015-0221-1. PubMed PMID: 26180516; PubMed Central PMCID: PMCPMC4502609.

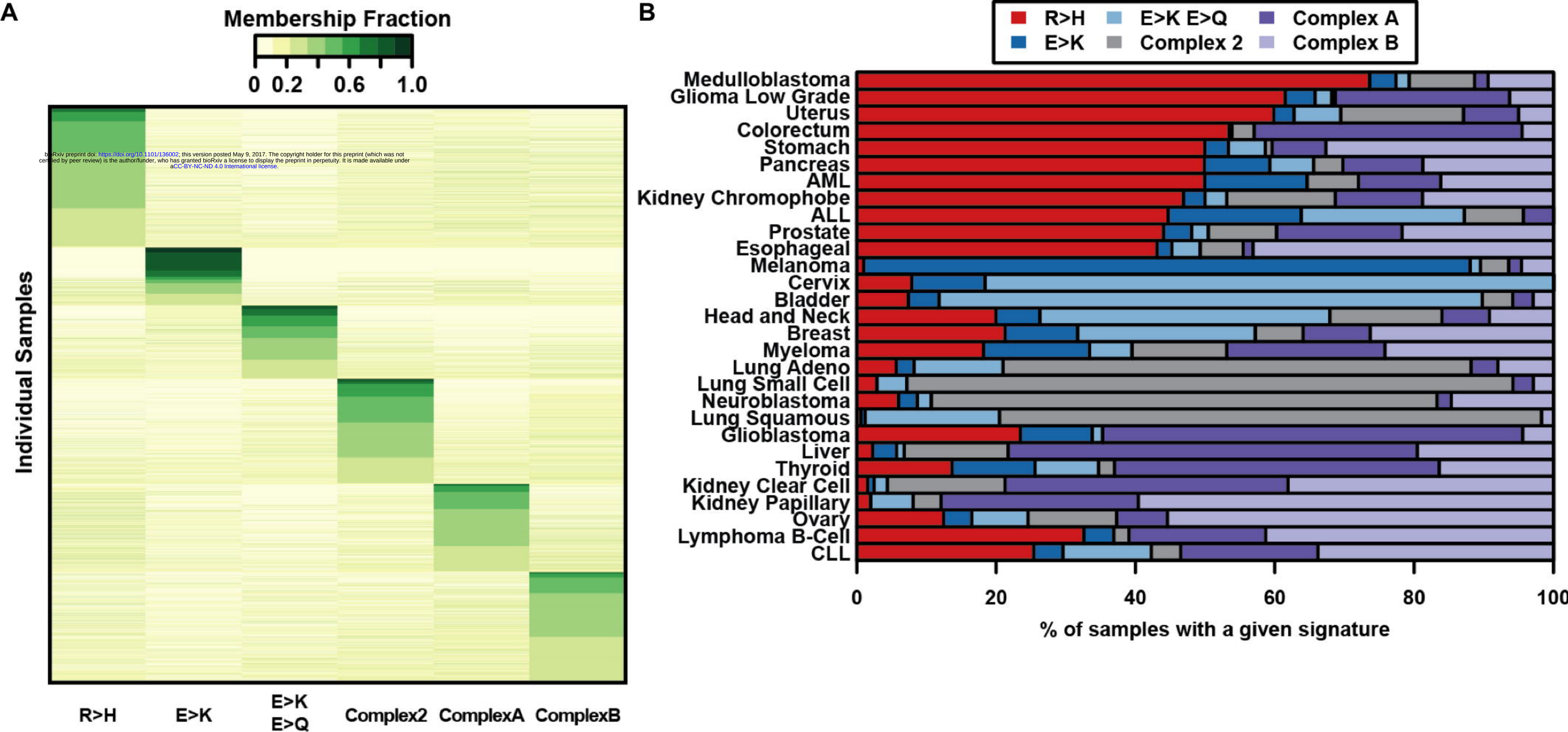
18. Alfarouk KO, Verduzco D, Rauch C, Muddathir AK, Adil HH, Elhassan GO, et al. Glycolysis, tumor metabolism, cancer growth and dissemination. A new pH-based etiopathogenic perspective and therapeutic approach to an old cancer question. Oncoscience. 2014;1(12):777-802. doi: 10.18632/oncoscience.109. PubMed PMID: 25621294; PubMed Central PMCID: PMCPMC4303887.

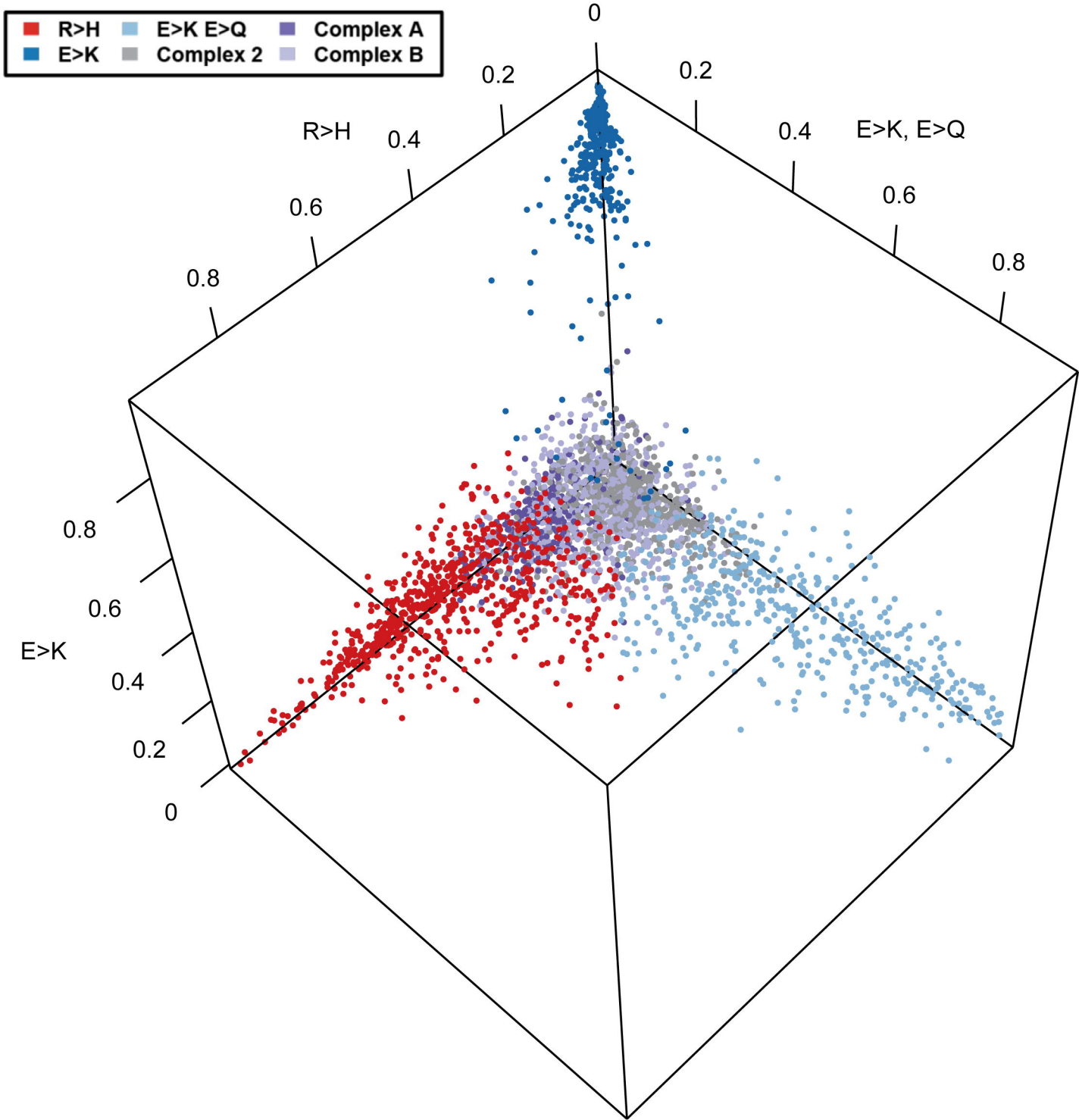
19. Gillies RJ, Verduzco D, Gatenby RA. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. Nat Rev Cancer. 2012;12(7):487-93. doi: 10.1038/nrc3298. PubMed PMID: 22695393; PubMed Central PMCID: PMCPMC4122506.

20. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9. doi: 10.1038/nmeth0410-248. PubMed PMID: 20354512; PubMed Central PMCID: PMCPMC2855889.

21. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010;11:367. doi: 10.1186/1471-2105-11-367. PubMed PMID: 20598126; PubMed Central PMCID: PMCPMC2912887.


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
Cervix
Pancreas
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)




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