RESEARCH ARTICLE

Prominent features of the amino acid mutation landscape in cancer

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, United States of America, **2** Biomedical Sciences Graduate Program, University of California, San Francisco, United States of America, **3** Department of Cell and Tissue Biology, University of California, San Francisco, United States of America, **4** Department of Pharmaceutical Chemistry, University of California, San Francisco, United States of America, **5** Quantitative Biosciences Institute, University of California, San Francisco, United States of America, **6** Institute for Human Genetics, University of California, San Francisco, United States of America

* ryan.hernandez@ucsf.edu (RDH); zachary.szpiech@ucsf.edu (ZAS)

OPEN ACCESS

Citation: Szpiech ZA, Strauli NB, White KA, Ruiz DG, Jacobson MP, Barber DL, et al. (2017) Prominent features of the amino acid mutation landscape in cancer. PLoS ONE 12(8): e0183273. https://doi.org/10.1371/journal.pone.0183273

Editor: Alvaro Galli, CNR, ITALY

Received: May 10, 2017

Accepted: August 1, 2017

Published: August 24, 2017

Data Availability Statement: Data are available at https://figshare.com/articles/mutation_data_zip/5276908.

Funding: This work was supported by National Institutes of Health (https://www.nih.gov/) grant CA178706 to DLB and RDH; National Institutes of Health grant CA197855 to DLB; National Institutes of Health grant HG007644 to RDH; and National Institutes of Health F32 grant CA177085 to KAW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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 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

Competing interests: The authors have declared that no competing interests exist.

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 [6–8], 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. These signatures were further validated in a non-overlapping set of samples from the COSMIC somatic mutation database. 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 non-synonymous amino acid mutations (see Materials and Methods, S1 Table and S2 Table 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 [9]. Applying NMF to the pooled mutation data reveals six mutation signatures at the amino acid level (S1G Fig), including two with strong Arg>His components and two with strong Glu>Lys components (Fig 1A, S1 Fig). 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.

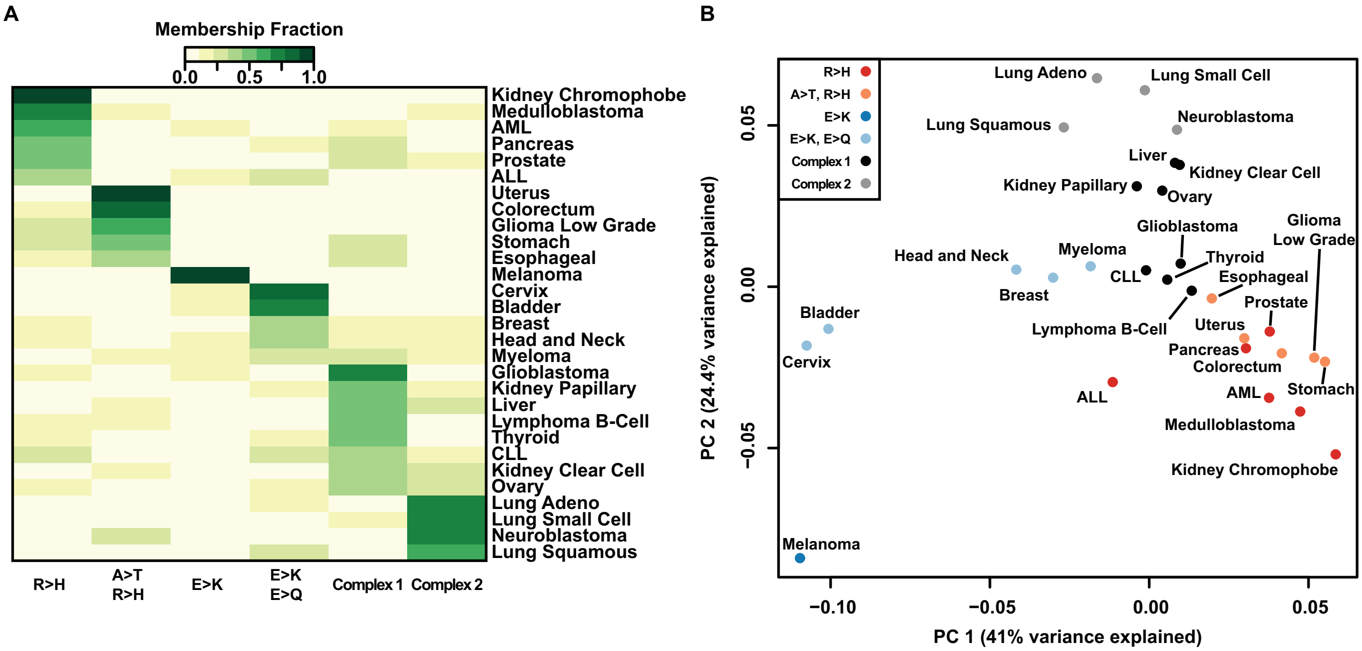

Fig 1. Arg>His and Glu>Lys mutations define mutation signatures of a subset of cancers. (A) Heatmap representation of six-component NMF clustering. Of the six amino acid mutation signatures identified, four have prominent charge-changing mutations: Arg>His (R>H), Glu>Lys (E>K), or Glu>Gln (E>Q). Two 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 S1 Fig. (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. Colors represent the greatest mutation signature contributing to a given cancer. Individual PC loadings can be found in S2 Fig.

https://doi.org/10.1371/journal.pone.0183273.g001

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 (S2 Fig). 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 (Fig 2A and S3 Fig). 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 mutations and may instead be a consequence of common features of the cancer, tumor microenvironment, or selective pressures, all of which may be targeted therapeutically.

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

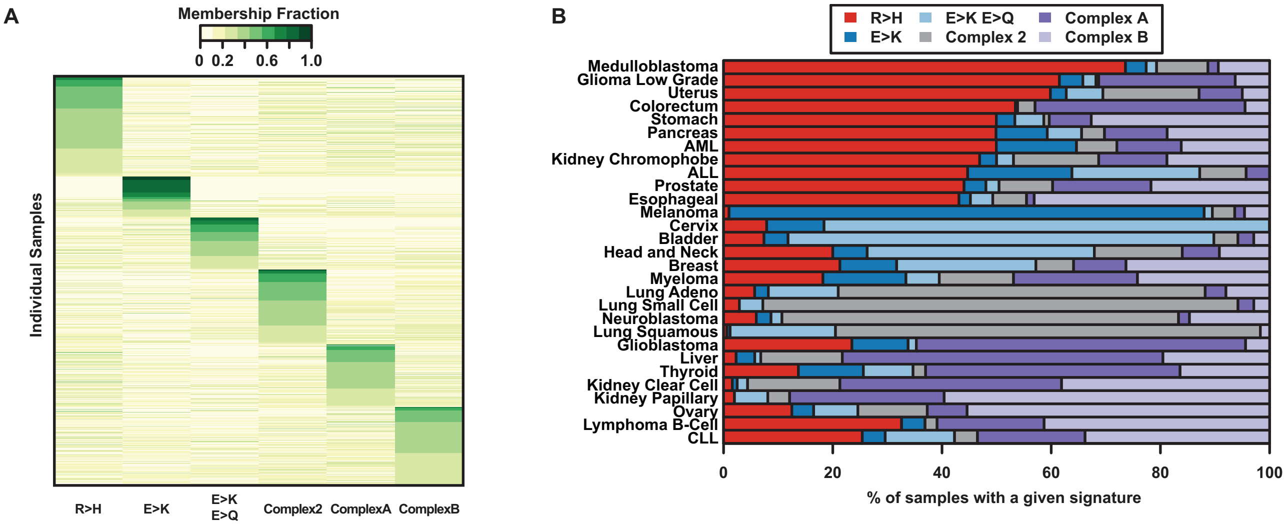

Fig 2. Amino acid mutation signatures for individual samples. (A) A heatmap representation of the six-component NMF clustering results for individual cancer samples (only those with >10 total nonsynonymous mutations). Samples with the same maximum signature component were grouped and sorted. Four amino 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 S3 Fig. (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.

https://doi.org/10.1371/journal.pone.0183273.g002

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.

Mutation signature validation

We validated the NMF signatures with an orthogonal data set (see Materials and Methods) from the COSMIC database [10]. The six mutation signatures identified from 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.

Discussion

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 [11–15]. 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.

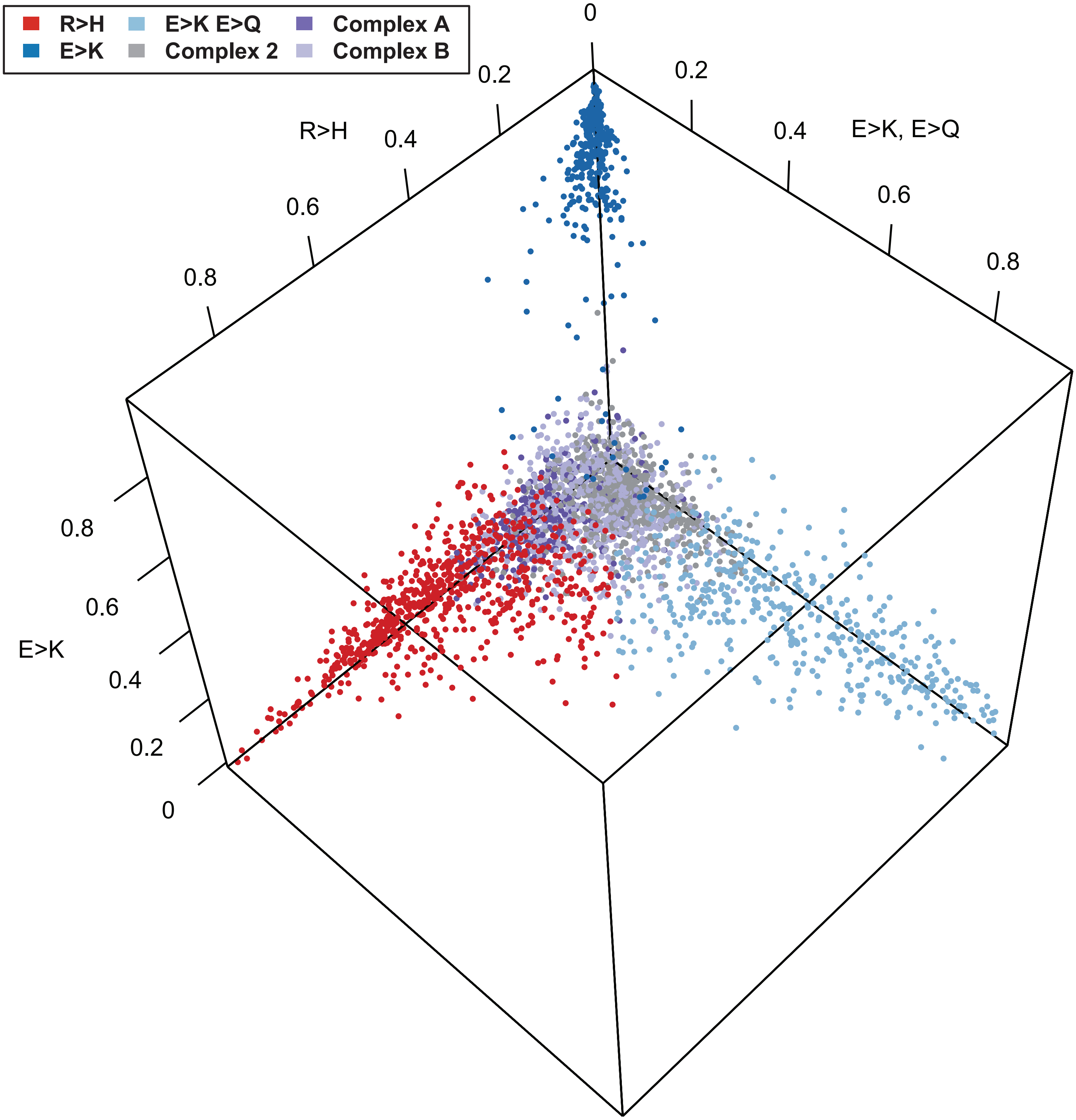

Fig 3. Normalized NMF mixture coefficients for individual samples. Plot of the normalized mixture coefficients across the three mutation signatures with high R>H or E>K components for every individual sample. Colors represent the greatest contributing mutation signature for each sample based on the full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

https://doi.org/10.1371/journal.pone.0183273.g003

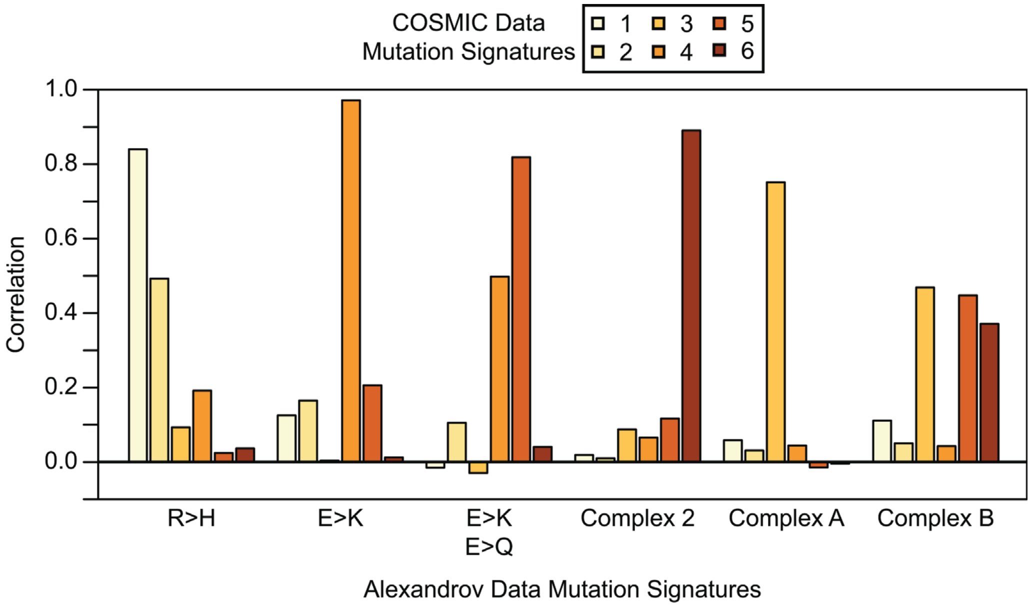

Fig 4. Correlations of COSMIC and Alexandrov mutation signatures. For each COSMIC mutation signature we calculated the correlation with each Alexandrov mutation signature. Five of six Alexandrov mutation signatures replicate in the COSMIC data for k = 6 mutation signatures. Alexandrov Signature R>H is 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 COSMIC Signature, as the signal appears to be spread amongst COSMIC Signatures 3, 5 and 6.

https://doi.org/10.1371/journal.pone.0183273.g004

Charge-changing mutations, whether buried or surface-exposed, can alter protein charge, electrostatics, and conformation [16]. Electrostatics of surface residues have been shown to play a key role in protein-protein interactions [17], protein-membrane interactions [18, 19], and kinase substrate recognition [20]. While it is important to note that our analyses are agnostic to the location of the mutation within the proteome and within a protein, the strong bias towards amino acid mutations that alter charge in our identified mutation signatures may suggest an adaptive advantage conferred by these mutations.

Glu>Lys mutations swap a negatively charged amino acid for a positively charged amino acid, which may in some cases affect protein function. Indeed, in some cases buried lysine mutations can induce global protein unfolding upon charging that alters mutant protein stability and function [21]. Furthermore, Glu>Lys mutations have been known to affect the function of PIK3CA [22–24].

Arg>His mutations swap a positively charged amino acid for a titratable amino acid. 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 [25–27]. Moreover, recent work has shown that some Arg>His mutations can confer pH sensitivity to the mutant protein and alter function [28]. We predict that some Arg>His mutations may be adaptive to increased pHi, conferring a gain in pH sensing to the mutant protein.

From our analyses, 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. Importantly, 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. Indeed, these results may help inform studies in the emerging field of Molecular Pathologic Epidemiology (MPE) [29, 30], which seeks to integrate knowledge across disciplines to inform personalized approaches to cancer prevention and therapy. Linking amino acid signatures to physiological or pathological features of the cancer could be important for identifying selective pressures that may be driving or sustaining the cancer as well as for limiting disease progression, particularly where targeted approaches fail [31–33].

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 [34] 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

$$\min_{w,H \geq 0} [D(X, WH) + R(W, H)] \quad k = 6$$

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 k$^{\text{th}}$ 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

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

where

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

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).

Supporting information

S1 Fig. Six nonsynonymous mutation signatures identified by NMF on aggregated data. NMF identifies six mutation signatures from data aggregated by cancer subtype including signatures with (A) high R>H weights, (B) high A>T and R>H weights, **(**C) high E>K weights, (D) high E>K and E>Q weights, and complex signatures (E) and (F). Amino acid mutations to or from X represent mutations to or from stop codons. (G) The cophenetic distance and change in residual sum of squares for different choices for the number (*k*) of mutation signatures for the NMF analyses of data aggregated by cancer subtype. Solid lines represent NMF on the original data and dotted lines represent NMF on this data after randomization. Six to seven components offer a consistently good fit.
(PDF)

S2 Fig. Principal components analysis loadings on aggregated nonsynonymous mutations for the first two principal components. (A) The first principal component separates cancers with many R>H mutations from those with many E>K mutations, and (B) the second principal component separates these cancers from those with more complex signatures. Amino acid mutations to or from X represent mutations to or from stop codons.
(PDF)

S3 Fig. Six nonsynonymous mutation signatures identified by NMF on individual samples. NMF identifies six mutation signatures from individual samples (only those with >10 total nonsynonymous mutations) including four signatures (A), (B), (C), (D) identified in the aggregated analysis and two other complex signatures, (E) and (F). Amino acid mutations to or from X represent mutations to or from stop codons. (G) The cophenetic distance and change in residual sum of squares for different choices for the number (*k*) of mutation signatures for the NMF analyses of data aggregated by individual samples. Solid lines represent NMF on the original data and dotted lines represent NMF on this data after randomization. Six to seven components offer a consistently good fit.
(PDF)

S4 Fig. Six nonsynonymous mutation signatures identified by NMF on individual samples from the COSMIC database. Six mutation signatures identified by NMF from individual samples in the COSMIC database (only those with >10 total nonsynonymous mutations) used for validation of NMF results from Alexandrov data. Amino acid mutations to or from X represent mutations to or from stop codons. (A), (C), (E), and (F) match with previously found signatures from the Alexandrov data.
(PDF)

S1 Table. Expected and observed mutation frequency for cancer driver genes. All Filters: entire filtering schema (Alexandrov *et al*., 2013) applied; No Filtering: none of the filters applied; Seq. Artifact Filter: only mutations defined as sequencing artifacts (Alexandrov *et al*., 2013) filtered. The expected range of mutation counts for the driver genes in cancers are based on reported mutation frequencies in the cancer genetics field and reported p53 mutation rates (see Methods).
(PDF)

S2 Table. Total nonsynonymous mutation counts by cancer type. Data was filtered using sequencing artifacts filter and any samples containing only synonymous mutations were eliminated from analysis (see Methods). Calculated is the mean number of nonsynonymous mutations per sample. Pilocytic astrocytoma samples were eliminated from analyses due to low

frequency of nonsynonymous mutations.
(PDF)

Acknowledgments

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.

Author Contributions

Conceptualization: Zachary A. Szpiech, Nicolas B. Strauli, Katharine A. White, Diego Garrido Ruiz, Matthew P. Jacobson, Diane L. Barber, Ryan D. Hernandez.

Data curation: Zachary A. Szpiech, Nicolas B. Strauli.

Formal analysis: Zachary A. Szpiech, Nicolas B. Strauli.

Funding acquisition: Katharine A. White, Matthew P. Jacobson, Diane L. Barber, Ryan D. Hernandez.

Investigation: Zachary A. Szpiech, Nicolas B. Strauli.

Methodology: Zachary A. Szpiech, Nicolas B. Strauli, Katharine A. White, Diego Garrido Ruiz, Matthew P. Jacobson, Diane L. Barber, Ryan D. Hernandez.

Project administration: Zachary A. Szpiech, Nicolas B. Strauli, Katharine A. White, Diego Garrido Ruiz, Matthew P. Jacobson, Diane L. Barber, Ryan D. Hernandez.

Resources: Zachary A. Szpiech, Nicolas B. Strauli.

Supervision: Matthew P. Jacobson, Diane L. Barber, Ryan D. Hernandez.

Validation: Zachary A. Szpiech, Nicolas B. Strauli.

Visualization: Zachary A. Szpiech, Nicolas B. Strauli.

Writing – original draft: Zachary A. Szpiech, Nicolas B. Strauli, Katharine A. White, Diego Garrido Ruiz, Matthew P. Jacobson, Diane L. Barber, Ryan D. Hernandez.

Writing – review & editing: Zachary A. Szpiech, Nicolas B. Strauli, Katharine A. White, Diego Garrido Ruiz, Matthew P. Jacobson, Diane L. Barber, Ryan D. Hernandez.

References

1. Gillies RJ, Gatenby RA. Metabolism and its sequelae in cancer evolution and therapy. Cancer J. 2015; 21(2):88–96. https://doi.org/10.1097/PPO.0000000000000102 PMID: 25815848; PubMed Central PMCID: PMCPMC4446699.
2. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012; 481(7381):306–13. https://doi.org/10.1038/nature10762 PMID: 22258609; PubMed Central PMCID: PMCPMC3367003.
3. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976; 194(4260):23–8. 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. https://doi.org/10.1038/nature08768 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. https://doi.org/10.1038/nature12477 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. https://doi.org/10.1038/srep12566 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. https://doi.org/10.1016/j.bbadis.2015.11.006 PMID: 26581171.

8. Tsuber V, Kadamov Y, Brautigam L, Berglund UW, Helleday T. Mutations in Cancer Cause Gain of Cysteine, Histidine, and Tryptophan at the Expense of a Net Loss of Arginine on the Proteome Level. Biomolecules. 2017; 7(3). https://doi.org/10.3390/biom7030049 PMID: 28671612.

9. 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. https://doi.org/10.1073/pnas.0308531101 PMID: 15016911; PubMed Central PMCID: PMCPMC384712.

10. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017; 45(D1):D777–D83. https://doi.org/10.1093/nar/gkw1121 PMID: 27899578; PubMed Central PMCID: PMCPMC5210583.

11. 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. https://doi.org/10.1242/jcs.195297 PMID: 28202602; PubMed Central PMCID: PMCPMC5339414.

12. 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. https://doi.org/10.1038/nrc1713 PMID: 16175178.

13. 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. https://doi.org/10.7554/eLife.03270 PMID: 25793441; PubMed Central PMCID: PMCPMC4392478.

14. Parks SK, Chiche J, Pouyssegur J. Disrupting proton dynamics and energy metabolism for cancer therapy. Nat Rev Cancer. 2013; 13(9):611–23. https://doi.org/10.1038/nrc3579 PMID: 23969692.

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

16. Zheng Y, Cui Q. Microscopic mechanisms that govern the titration response and pKa values of buried residues in staphylococcal nuclease mutants. Proteins. 2017; 85(2):268–81. https://doi.org/10.1002/prot.25213 PMID: 27862310.

17. Zhang Z, Witham S, Alexov E. On the role of electrostatics in protein-protein interactions. Phys Biol. 2011; 8(3):035001. https://doi.org/10.1088/1478-3975/8/3/035001 PMID: 21572182; PubMed Central PMCID: PMCPMC3137121.

18. Cho W, Stahelin RV. Membrane-protein interactions in cell signaling and membrane trafficking. Annu Rev Biophys Biomol Struct. 2005; 34:119–51. https://doi.org/10.1146/annurev.biophys.33.110502.133337 PMID: 15869386.

19. Stahelin RV. Lipid binding domains: more than simple lipid effectors. J Lipid Res. 2009; 50 Suppl:S299–304. https://doi.org/10.1194/jlr.R800078-JLR200 PMID: 19008549; PubMed Central PMCID: PMCPMC2674730.

20. Ubersax JA, Ferrell JE Jr. Mechanisms of specificity in protein phosphorylation. Nat Rev Mol Cell Biol. 2007; 8(7):530–41. https://doi.org/10.1038/nrm2203 PMID: 17585314.

21. Chimenti MS, Khangulov VS, Robinson AC, Heroux A, Majumdar A, Schlessman JL, et al. Structural reorganization triggered by charging of Lys residues in the hydrophobic interior of a protein. Structure. 2012; 20(6):1071–85. https://doi.org/10.1016/j.str.2012.03.023 PMID: 22632835; PubMed Central PMCID: PMCPMC3373022.

22. Carson JD, Van Aller G, Lehr R, Sinnamon RH, Kirkpatrick RB, Auger KR, et al. Effects of oncogenic p110alpha subunit mutations on the lipid kinase activity of phosphoinositide 3-kinase. Biochem J. 2008; 409(2):519–24. https://doi.org/10.1042/BJ20070681 PMID: 17877460.

23. Huang CH, Mandelker D, Gabelli SB, Amzel LM. Insights into the oncogenic effects of PIK3CA mutations from the structure of p110alpha/p85alpha. Cell Cycle. 2008; 7(9):1151–6. https://doi.org/10.4161/cc.7.9.5817 PMID: 18418043; PubMed Central PMCID: PMCPMC3260475.

24. Meyer DS, Koren S, Leroy C, Brinkhaus H, Muller U, Klebba I, et al. Expression of PIK3CA mutant E545K in the mammary gland induces heterogeneous tumors but is less potent than mutant H1047R. Oncogenesis. 2013; 2:e74. https://doi.org/10.1038/oncsis.2013.38 PMID: 24080956; PubMed Central PMCID: PMCPMC3816227.

25. 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. https://doi.org/10.1083/jcb.201302131 PMID: 24043700; PubMed Central PMCID: PMCPMC3776353.

26. 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. https://doi.org/10.1083/jcb.200804161 PMID: 19029335; PubMed Central PMCID: PMCPMC2592832.

27. 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. https://doi.org/10.1074/jbc.M116.736215 PMID: 27650500; PubMed Central PMCID: PMCPMC5104935.

28. DiGiammarino EL, Lee AS, Cadwell C, Zhang W, Bothner B, Ribeiro RC, et al. A novel mechanism of tumorigenesis involving pH-dependent destabilization of a mutant p53 tetramer. Nat Struct Biol. 2002; 9 (1):12–6. https://doi.org/10.1038/nsb730 PMID: 11753428.

29. Ogino S, Chan AT, Fuchs CS, Giovannucci E. Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field. Gut. 2011; 60(3):397–411. https://doi.org/10.1136/gut.2010.217182 PMID: 21036793; PubMed Central PMCID: PMCPMC3040598.

30. Nishi A, Milner DA Jr., Giovannucci EL, Nishihara R, Tan AS, Kawachi I, et al. Integration of molecular pathology, epidemiology and social science for global precision medicine. Expert Rev Mol Diagn. 2016; 16(1):11–23. https://doi.org/10.1586/14737159.2016.1115346 PMID: 26636627; PubMed Central PMCID: PMCPMC4713314.

31. 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. https://doi.org/10.1186/s12935-015-0221-1 PMID: 26180516; PubMed Central PMCID: PMCPMC4502609.

32. 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. https://doi.org/10.18632/oncoscience.109 PMID: 25621294; PubMed Central PMCID: PMCPMC4303887.

33. 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. https://doi.org/10.1038/nrc3298 PMID: 22695393; PubMed Central PMCID: PMCPMC4122506.

34. 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. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512; PubMed Central PMCID: PMCPMC2855889.

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