Dynamic clustering of genomics cohorts beyond race, ethnicity—and ancestry

Hussein Mohsen[1,2,*], Kim Blenman[3,4], Prashant S. Emani[5], Quaid Morris[1,6], Jian Carrot-Zhang[2], Lajos Pusztai[3]

[1] Computational and Systems Biology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[2] Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[3] Breast Medical Oncology, School of Medicine, Yale University, New Haven, CT 06511, USA

[4] Computer Science, Yale University, New Haven, CT 06511, USA

[5] Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA

[6] Computational Biology and Medicine, Weill-Cornell Medical College, New York, NY 10065, USA

[*] Corresponding author

Abstract

Background. Recent decades have witnessed a steady decrease in the use of race categories in genomic studies. While studies that still include race categories vary in goal and type, these categories already build on a history during which racial color lines have been enforced and adjusted in the service of social and political systems of power and disenfranchisement. For early modern biological classification systems, data collection was also considerably arbitrary and

limited. Fixed, discrete classifications have limited the study of human biodiversity and disrupted widely spread genetic and phenotypic continuums across geographic scales. Relatedly, the use of broad and predefined classification schemes—e.g. continent-based—across traits can risk missing important trait-specific genomic signals.

Results. To address these issues, we introduce a dynamic approach to clustering human genomics cohorts on a trait-specific level and without using a set of predefined categories. We tested the approach on whole-exome sequencing datasets in ten cancer types and partitioned them based on germline variants in cancer-relevant genes that could confer cancer type-specific disease predisposition. Results demonstrate clustering patterns that transcend discrete continent-based categories across cancer types. Functional analysis based on cancer type-specific clusterings also captures the fundamental biological processes underlying cancer, differentiates between dynamic clusters on a functional level, and identifies novel potential drivers overlooked by a continent-based clustering model.

Conclusions. Through a trait-based lens, the dynamic clustering approach reveals genomic patterns that transcend predefined classification categories. We propose that coupled with diverse data collection, new clustering approaches have the potential to draw a more complete portrait of genomic variation and to address, in parallel, technical and social aspects of studying human biodiversity.

Keywords: genetic variation, cancer genomics, classification, ancestry, ethnicity, race

Introduction

1.1. Historical perspective

The earliest scientific attempt to use race as a category to classify human beings dates back to the seventeenth century. In a 1684 essay titled, "A New Division of the Earth, According to the Different Species or Races of Men Who Inhabit it," French physician Francois Bernier categorized human beings into five types, the last of which, the Sámi people, he described as "little stunted creatures" and "wretched animals" [1]. Swedish botanist Carl Linnaeus, dubbed as the founder of modern taxonomy, published decades later (1735) the first edition of *Systema Naturae*, in which he created a system with four categories instead, and based his categories on geography and skin color. In the tenth edition (1758), he expanded the system and confounded physical with personality and social traits based on his interpretation of the humoral theory that links geography and climate to skin color and good and bad character [2]. In this work, Linnaeus loaded his classifications with prejudice and crafted a hierarchy placing *Homo sapiens europaeus*, a category he color-coded as white, on top, while using descriptions such as "harsh face," "careless," "stubborn," "lazy," "greedy," and "ruled by caprice" to describe *Homo sapiens americanus*, *afer* and *asiaticus*, color-coded respectively as red, black, and yellow [1, 3, 4]. Linnaeus also added a separate category he called *Homo sapiens monstrosus*, in which he mostly included humans with various birth defects and mythical "humans" such as giants from Patagonia [2-5].

Bernier and Linnaeus suggested different sets of categories and imbued their human classification systems with imagined hierarchical value judgement. Their imposition of a few, fixed, distinct and

discrete categories reduced the complexity of human biodiversity and shaped subsequent classification systems. In early systems of classification, descriptions often depended on the philosophical (and political) choices of the classifiers, technological limitations, and economic factors including trade routes [2, 6]. For instance, the routes involving Sweden and the Netherlands during Linnaeus' time shaped his considerably arbitrary choice to describe peoples of specific geographies but not others. Further, Linnaeus relied on anecdotal and written accounts of his students and of missionaries, mercantilists, travelers, and slave traders, and did not travel himself outside of western Europe [2, 4].

Early biological classification systems were also venues for the use of emerging modern science to demarcate human difference in the service of power during times of colonial expansions and the Atlantic slave trade. Institutions and individuals exerted intentional efforts to create racial classification systems in modern science, which opened the door for racialized hypothesis generation. In a stark yet far from lone example, the Bordeaux Royal Academy of Science announced an essay contest in 1739 to study "the degeneration of Black hair and Black skin." The announcement was made a year after the regional assembly of Bordeaux endorsed the existence of enslaved Black people on French soil, and as recently described in "Who's Black and Why: A Hidden Chapter from the Eighteenth-Century Invention of Race," before members of the Bordeaux Academy decided to invest Academy prize money in the company that ran the French slave trade in the African continent, Compagnie Perpetuelle des Indes [7]. Another example of a (re)defined color line to benefit the interests of chattel slavery before its abolition in the United States is the introduction of "one drop laws," according to which a person was categorized Black if they had a known "trace of Black blood" in their ancestry. As a result, in the words of anthropologist Nina

Jablonski, skin color was “no longer the necessary and sufficient criterion for race classification” [2] (for more examples, see [8, 9] and references listed in Box 1-1 of [10]).

Changing meanings of race categories continued to reflect and drive political and social transformations. Since the beginning of the first U.S. census in 1790, for example, racial groupings in the census have changed more than twenty times [1]. Notably, race categories and their meanings also vary across national borders [1, 10]. Biology and medicine are susceptible to societal and cultural influences, and scientists are engaged in a bi-directional process of being influenced by social and cultural concepts that co-shape interpretations of nature, and scientific interpretations that in turn influence social order [11, 12]. Scientific attempts to formulate biological classification systems of race continued in the nineteenth and twentieth centuries and were muddled, again, with confusion. In the words of anthropologist Fay-Cooper Cole during the opening of the “Conference on Racial Differences,” which was held in 1928 at the National Academy in Washington, D.C. and attended by opponents and proponents of eugenics, the term “race” was “frequently used in three or four ways in the same article,” and there existed “a great deal of confusion in the use of the word.” Eugenics, which propagated race science for more than six decades, declined during the 1930s and 1940s after it faced strong scrutiny and criticism within scientific circles and in response to Nazi eugenical horrors. Yet, the use of racial classification systems to study human biodiversity continued [13]. The understanding of human biodiversity has progressed since then, however, and further highlighted their unreliability.

1.2. Further limitations

The reduction of human biodiversity into a small set of skin color-coded race categories implies the existence of distinct skin color lines. However, skin color is a continuum influenced by climate, genetics, and the intensity and seasonality of ultraviolet radiation [2]. Similar skin colors can also result from convergent adaptation in response to similar selective pressures, rather than from genetic relatedness [2, 14], and further analysis and data collection demonstrated the prevalence of continuous rather than discrete skin color distributions (e.g. [15, 16]). Further, recent comprehensive efforts centered on diverse genomics data have demonstrated (i) the complexity and prevalence of continuums on the genome level (e.g. [17]), and (ii) the sharp limitations of broadly predefined classification—be it at the level of socio-political categories such as race and ethnicity, or broad geographic ones such as continental ancestry (e.g. [18]).

While genetic ancestry is a concept that describes a partial relationship of a person with their genealogical history, it can still be subject to significant limitations as a classification criterion. First, there is no single criterion to define ancestry, and categories can take geographic (e.g. South Asian or Central American), geopolitical (e.g. Zambian or Italian), or cultural (e.g. Brahmin or Lemba) meanings [14]. Second, geographic ancestry categories might be muddled with imprecise conflation with race categories, and their descriptors can be distortive of time and space (e.g. references to Asian ancestry might exclude the entirety of or wide regions within West, Central, East or South Asia, and nationality-based categories might refer to ancestors during the period that preceded the very formation of respective countries). Third, continent-based categories reduce the high levels of genomic complexity within each continent and might inadvertently imply purity

141 within continents when communicating results. Fourth, separate categories impose discreteness on
142 continuums between continents [14, 17, 19-21] that have long been connected by land (e.g. Asia
143 and Europe, or today's Asia and Africa through the Sinai Peninsula), in technology, or both.
144
145 Further, single category assignments to individuals ignore the multitudes of personal belonging.
146 While many individuals choose to affiliate with multiple groups for personal or cultural reasons
147 [14], it is also highly common for individuals to have a "mixed genetic ancestry" with respect to
148 sets of predefined categories (e.g. 97.3% of individuals are associated with a median of four
149 ancestry categories in [18], in consistence with [22]). Importantly, this "mixed ancestry" can be
150 observed on the individual level and does not have to reflect population stratification. Further,
151 significant amounts of genetic ancestry labeled as "Western Asian," for example, is present in
152 samples with origins ranging from present-day Morocco to Mongolia, and from England to
153 Ethiopia, that is, in Asia, Europe and Africa [18].
154
155 *1.3. Dynamic, trait-specific germline clustering*
156
157 Given the limitations of predefined classification systems, and recognizing the wide range of
158 phenotypes and the complexity of genomic variation, we propose a dynamic approach that
159 generates trait-specific clusterings of genomics cohorts. The approach builds on an earlier idea
160 from an exchange between biological anthropologist Frank B. Livingstone and evolutionary
161 biologist Theodosius Dobzhansky on generating clusterings on the gene(s) level (see [23] and
162 Chapter 9 of [13]), and expands it in light of the wide advances in genomic data collection,
163 measurement of genomic variation, and interpretation of the genomic basis of complex traits. The

approach is also motivated by multiple factors. First, the genomic basis of different traits is encoded in different loci of the human genome, and the loci relevant for a single trait, ranging from one to many in number, cover only a small portion of the whole genome. Second, biological and physiological processes are shared among individuals within and across any number of groups—however ancestrally defined. Third, especially when common germline polymorphisms are involved in part or in full in trait predisposition, the genomic variants are significantly shared across continental regions. Relatedly, a single nucleotide polymorphism (SNP) can be concurrently classified "rare" in multiple regions, and the distribution of classifications depends on available data [24].

Fourth, evolutionary forces might be acting on trait-specific genomic regions in parallel in distant geographies. Fifth, as predefined ancestral labels concurrently bear geopolitical, historical, and social meanings, their assignment to categories used to study genomic biodiversity, and particularly predisposition to disease, opens the door for prolonging a history of stigmatizing entire communities [25].

Finally, the disruption of observed continental clines overlooks inter-continental patterns related to a trait of interest. This raises a core question on the goal of clustering cohorts: if two individuals in distant geographies have similar genomic markers and phenotypic expression corresponding to a trait, e.g. both are right-handed, should they be in the same cluster when studying the genomic basis of handedness, or separate ones? Should "populations" be determined based on a gene(s) (or trait) of interest, or the whole genome? What are the limitations of clustering based on the whole genome, in a fragmented and data-scarce continental setting, when a trait is affected by only a

small subset of genomic regions? Further, it is also quantitatively well-established that selected features or clustering criteria strongly affect resulting "populations" (i.e. clusters or clines), and consequently reshape the starting point from which to discover—or miss—patterns and generate hypotheses [26, 27].

The dynamic clustering approach takes a different angle to classifying biodiversity by grouping individuals based on predisposition to a trait under study—herein a cancer type. An individual's membership to a cluster depends on their genomic sequence at a specific set of regions known to be associated with the trait. Number of clusters, which are *de facto* neutrally labelled, is determined according to the dataset and specific trait under study.

In cancer, germline (inherited) predisposition is mediated by deleterious mutations in several dozen high penetrance cancer-relevant genes and probably a combination of individually low penetrance variants. Different genes are associated with different degrees of risk, and with variable cancer-specificity [28]. Further, germline alterations require additional acquired (somatic) mutations for malignant transformation [29-31]. We hypothesize that clustering cancers based on their germline variants in cancer type-specific loci transcends predefined continental categories. We expect clusterings to vary across cancers—in terms of number of clusters and sample-cluster membership—to reflect human biodiversity and heterogeneity across complex traits. We also note that this dynamic (i.e. trait-specific) approach moves beyond the notion of local ancestry at a single locus as it can simultaneously consider any set of coding or non-coding regions associated with a trait, and it can scale to accommodate newly acquired knowledge on the genomic basis of the trait under study.

210 **Results**

211

212 *2.1. Overview*

213

214 To study predisposition to ten of the most common cancer types [32], we utilized germline data
215 from the Cancer Genome Atlas (TCGA) [28, 30] and ancestral category values from the The
216 Cancer Genetic Ancestry Atlas (TCGAA) [33]—which are based on comparisons with the 1000
217 Genomes [34], HGDP [35] and HapMap [36] projects. In studied cohorts, TCGAA uses categories
218 that refer to continental and sub-continental regions. For clarity, we hereafter refer to them
219 collectively as continent-based.

220

221 We first generated the trait-specific clusterings of each cohort, using the TCGA terminology for
222 cancer types, BRCA, COAD, KIRC, LIHC, LUAD, LUSC, OV, PAAD, PRAD, and READ, using
223 all nonsynonymous SNPs within different sets of COSMIC genes known to have germline
224 association with each cancer type (Methods, Supplementary Table 3). Next, we focused on the
225 SNP subset predicted to have high functional impact within each cancer type's samples as a basis
226 for dynamically generating clusters. We then assessed the performance of three algorithmic
227 clustering approaches (K-means, DBSCAN, and HClust) to identify generated clusters in
228 multidimensional scaling (MDS) plots. Finally, we identified potentially overlooked somatic
229 driver genes in each TCGA cohort based on dynamic clustering in comparison with drivers
230 identified using the continent-based model, and performed a functional genomic analysis to assess
231 their biological and clinical importance in the context of cancer.

232

233 *2.2. Beyond continent-based categories*

234

235 Upon dynamically clustering based on trait-specific regions—herein cancer type-specific germline
236 COSMIC genes, a number of visual patterns emerge. First, the number of clusters varies per cancer
237 type (1-8 clusters), strongly reflecting known genomic heterogeneity in cancer [28, 30, 37] (Figure
238 1). Second, and despite the relative lack of diversity in TCGA datasets, clusters transcend
239 continent-based categories in all cancer types to include samples with "African," "East Asian,"
240 "European," and "Other" ancestral labels within clusters. Third, this pattern is also observed in
241 colon and rectum cancers, known to be associated with high disparities in incidence and outcome
242 [38, 39], with one and two clusters (Figure 1b and 1j), respectively.

243

244 *2.3. High functional impact compact clusters*

245

246 Next, we selected subsets of SNP variants with high functional impact on protein function (HFI)
247 within the COSMIC genes corresponding to each TCGA cohort as a basis for clustering (see
248 Methods). This selection led to SNP subsets with n = 1 (PAAD) to 269 (BRCA). HFI-based
249 dynamic clusters transcend continent-based boundaries and exhibit two notable patterns. First, the
250 number of HFI-based clusters is higher, on average, than clusters based on all nonsynonymous
251 SNPs of COSMIC genes (e.g. PAAD with two COSMIC-based vs four HFI-based clusters in
252 Figure 1h and Figure 2a, respectively; HFI-based clusters for other cancer types in Supplementary
253 Figure 1). Second, subsections within select cancer types tend to be compact and often include
254 samples with highly similar or identical HFI variant patterns due to the smaller number of loci
255 used for clustering in these cohorts. Dots corresponding to distinct samples overlay each other,

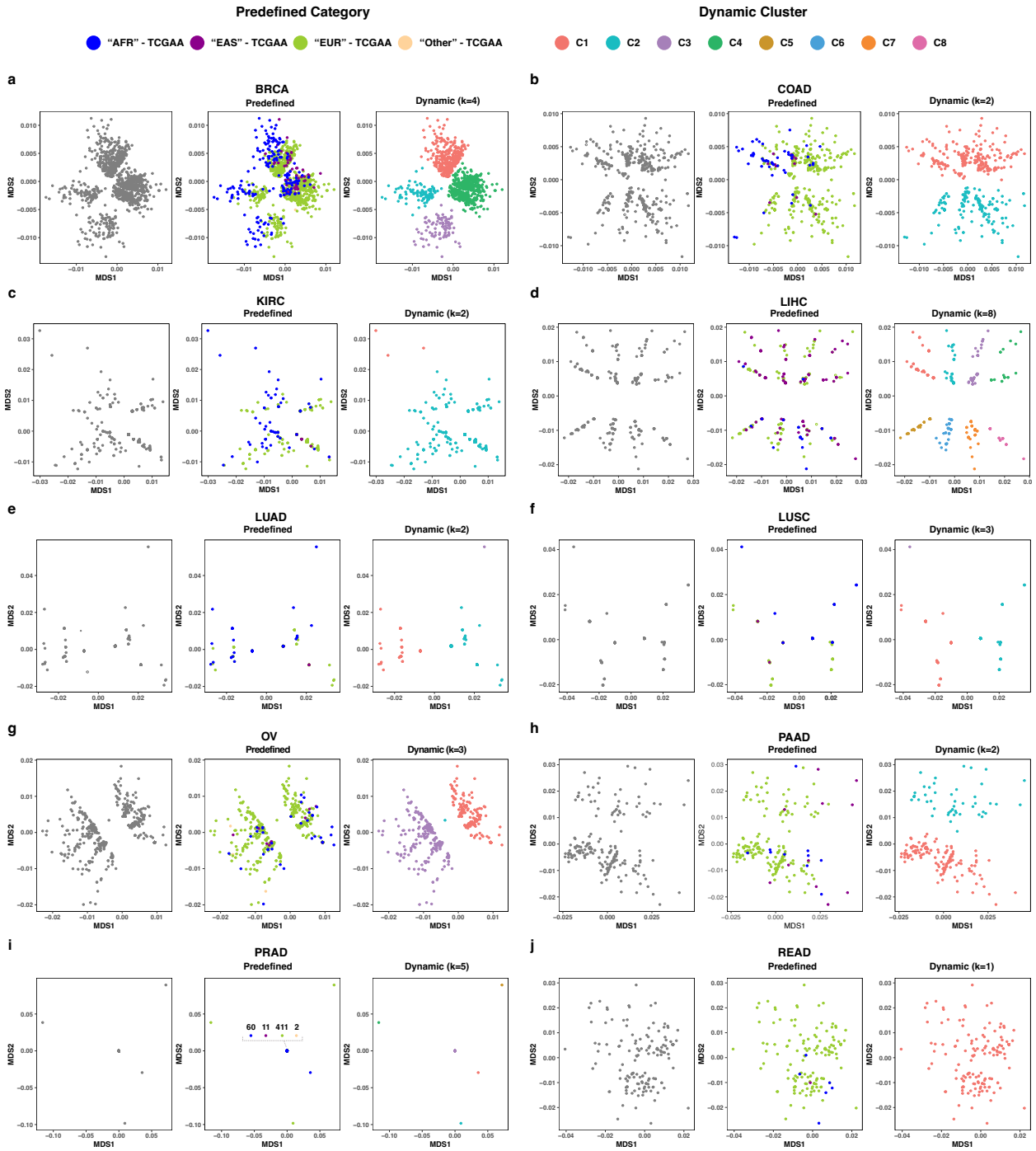

Figure 1. Multidimensional scaling (MDS) plots of dynamically generated clusters for ten TCGA cancer cohorts. Cancer type-specific dynamic clusters transcend predefined continent-based categories. Dynamic cluster numbers (i.e. C1, C2, … C8) correspond to disjoint sample subsets within each cancer cohort.

resulting in single dots each representing a subcluster (e.g. LUSC-HFI in Figure 2a and KIRC-HFI in Supplementary Figure 1).

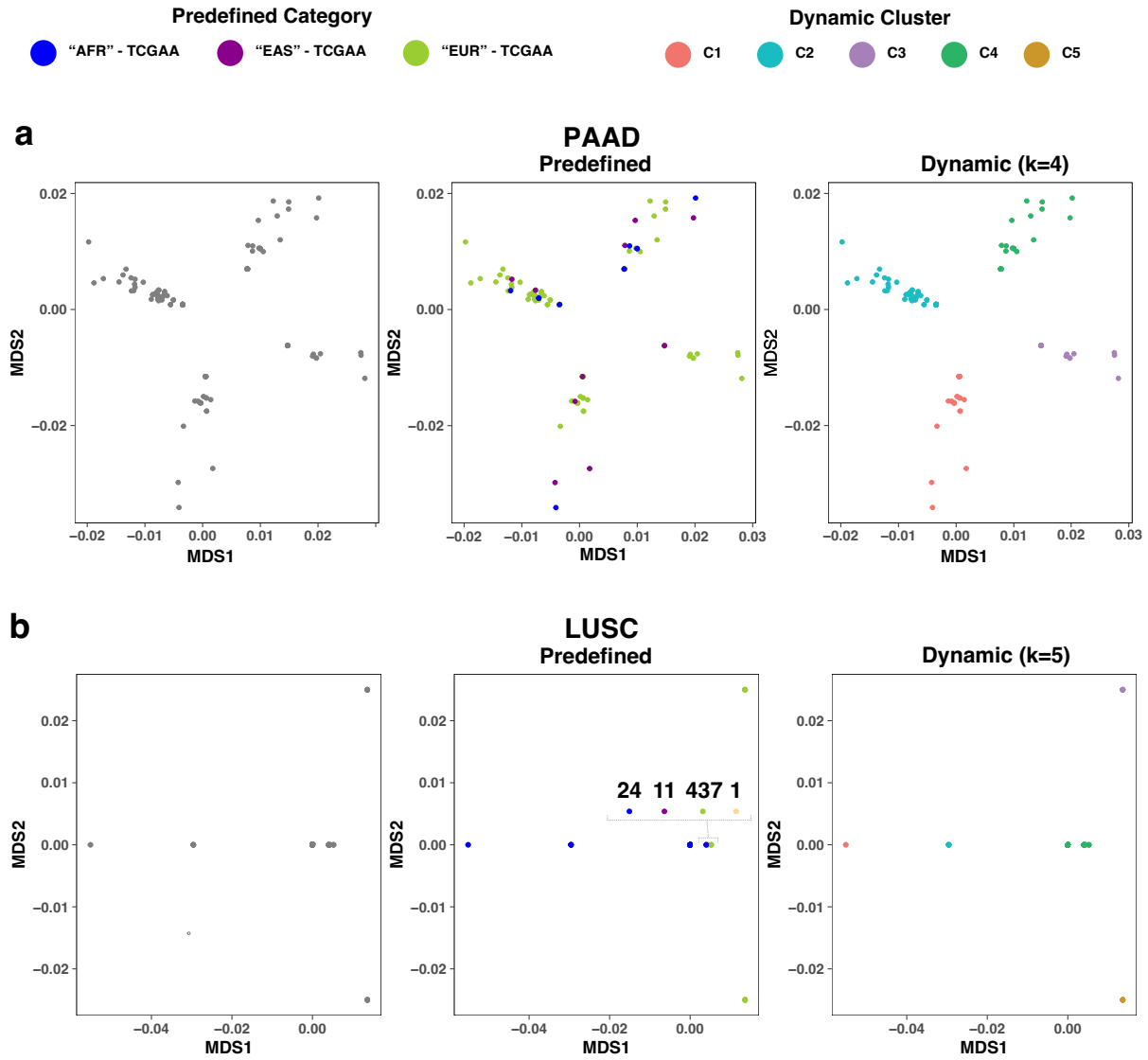

Figure 2. MDS plots of dynamically generated clusters based on high functional impact germline variant subsets. (a) PAAD results demonstrate a higher number of clusters in HFI-based results compared to ones based on all nonsynonymous variants in the COSMIC-based setting in Figure 1h. (b) LUSC results demonstrate compact clusters with a high number of samples demonstrating similar or identical variation patterns in HFI subsets.

270 *2.4. Human aid improves algorithmic clustering*

271

272 Algorithmic approaches can generate different clusterings of the same dataset, and their
273 performance primarily depends on data distribution and cluster definition [27, 40]. To
274 algorithmically identify clusters, we ran the K-means algorithm with a predefined k = 4 (number
275 of selected TCGAA continent-based categories) and an "optimal" k chosen based on visual
276 inspection of cancer type-specific plots. Given the varying number of clusters across cancer types,
277 predefined-k clustering performed poorly across multiple cancer types (Figure 3a-d). Similarly,
278 dynamic-k K-means faced challenges in accurately identifying clusters across cohorts (e.g. LIHC-
279 COSMIC, Figure 3e).

280

281 Similarly, clustering results from two other algorithms, DBSCAN (Density-based Spatial
282 Clustering of Applications with Noise) and HClust, performed poorly in multiple cohorts. HClust,
283 which stands for agglomerative hierarchical clustering—herein used with complete-linkage, is a
284 bottom-up approach that starts with individual points as separate clusters and iteratively merges
285 most similar clusters until a predefined number of clusters is met (e.g. k = 8 clusters for LIHC-
286 COSMIC in Figure 3f). Like K-means, HClust correctly identified only a subset (i.e. two) of the
287 eight clusters in LIHC-COSMIC. DBSCAN, which is known to identify dense clusters and outliers
288 in low-density regions, partitioned LIHC-COSMIC results into three clusters (Figure 3g), a
289 number decided algorithmically based on input parameters (Methods): two large clusters, each
290 roughly with four of the dynamic clusters, and a third cluster that includes distant samples the
291 algorithm considered outliers (in red).

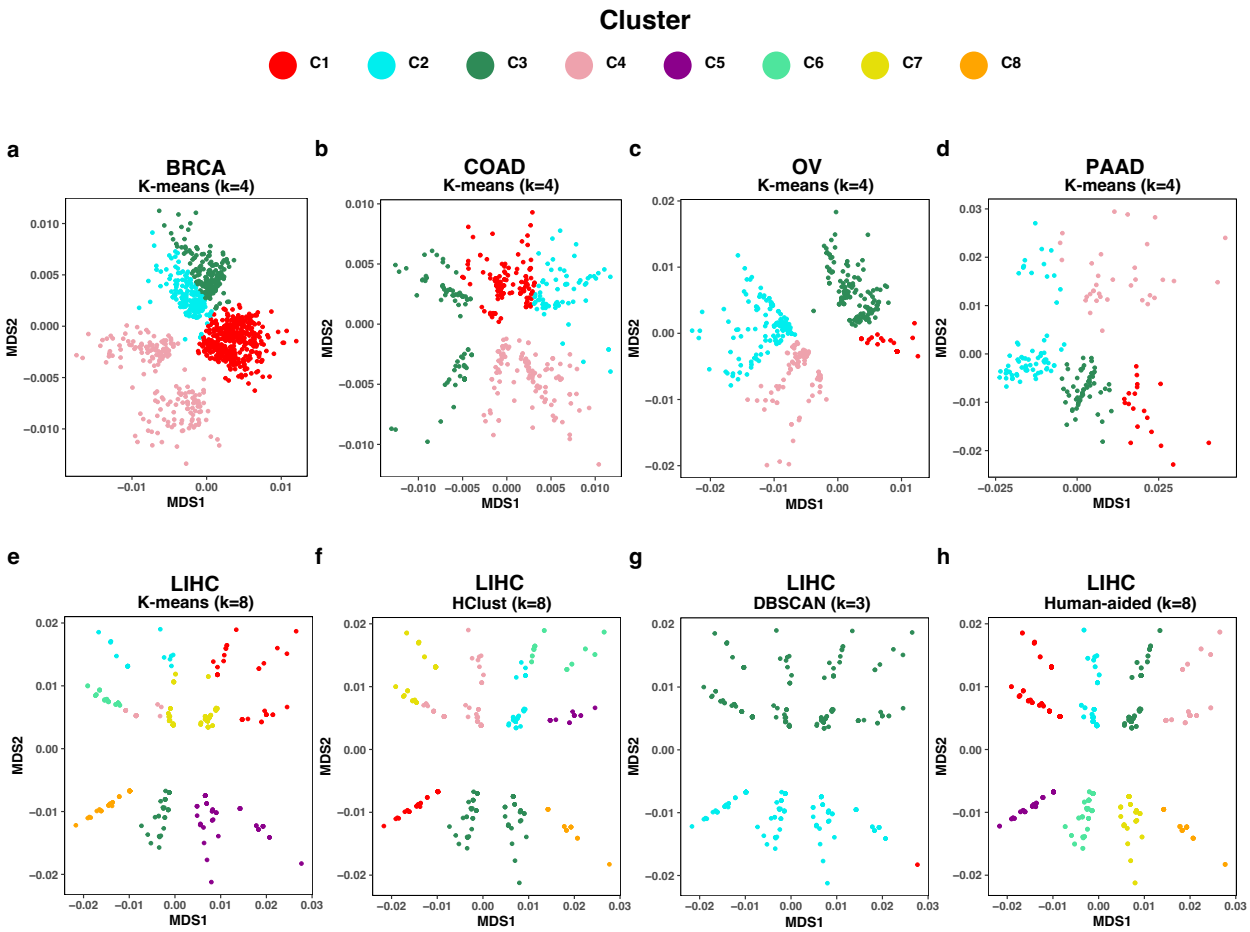

Figure 3. Algorithmic and human-aided identification of dynamic clusters. K-means results with predefined-k = 4 fails to identify COSMIC-based clusters in (a) BRCA, (b) COAD, (c) OV, and (d) PAAD among other cancer types. Dynamic-k results also demonstrate the failure of (e) K-means, (f) HClust, and (g) DBSCAN to identify COSMIC-based clusters in LIHC, highlighting the need for (h) human-aid in cluster identification.

In sum, and while specific algorithms perform better than others at identifying clusters, algorithmic clustering does not seem to suffice for both COSMIC- and HFI-based settings. As a result, human intervention to attain more precise results remains central (e.g. Figure 3h). We also note that in certain instances, multiple “optimal” numbers of clusters in the same plot can exist, and the choice

303 remains centered on experimental goals and the problem under study (e.g. the four dynamic
304 clusters of BRCA-COSMIC in Figure 1a being alternatively considered two larger diagonally-
305 separated clusters in the same plot).
306
307 *2.5. Dynamically-identified cancer drivers*
308
309 We next explored the biological significance of clusters identified by the dynamic clustering
310 approach. Particularly, we focused on identifying potential cancer type-specific somatic driver
311 genes. We used MutSigCV to identify potential somatic drivers based on whole exome sequence
312 data corresponding to each dynamic cluster generated on germline variant sets (COSMIC genes
313 and HFI). COSMIC-based clusters yielded 98 potential drivers across cancer types, and HFI-based
314 clusters 109 drivers. Both lists included tens of known drivers from a more comprehensive list by
315 Bailey *et al.* which relied on multiple computational and experimental tools [41]. These include
316 *KRAS*, *TP53*, *PIK3CA*, *BRCA1*, *PTEN*, *CDH1*, *RB1*, *PTEN*, *FOXA1*, *SPOP*, and *VHL* (for full
317 lists, see Supplementary Tables 1 and 2). COSMIC- and HFI-based lists also include 23 and 36
318 cancer type-specific novel potential drivers, respectively, which are overlooked if analysis is
319 performed on continent-based clusters (Figure 4a). Among these genes are known drivers listed in
320 [41], including *APC*, *CBFB*, *B2M*, *CDKN2A*, and *RPL5*.
321
322 We then investigated the functional importance of COSMIC- and HFI-based driver gene lists.
323 Given their significant coverage of known drivers, enrichment analysis of both full lists point to
324 known oncogenic pathways, including the majority of signaling pathways listed in Sanchez-Vega

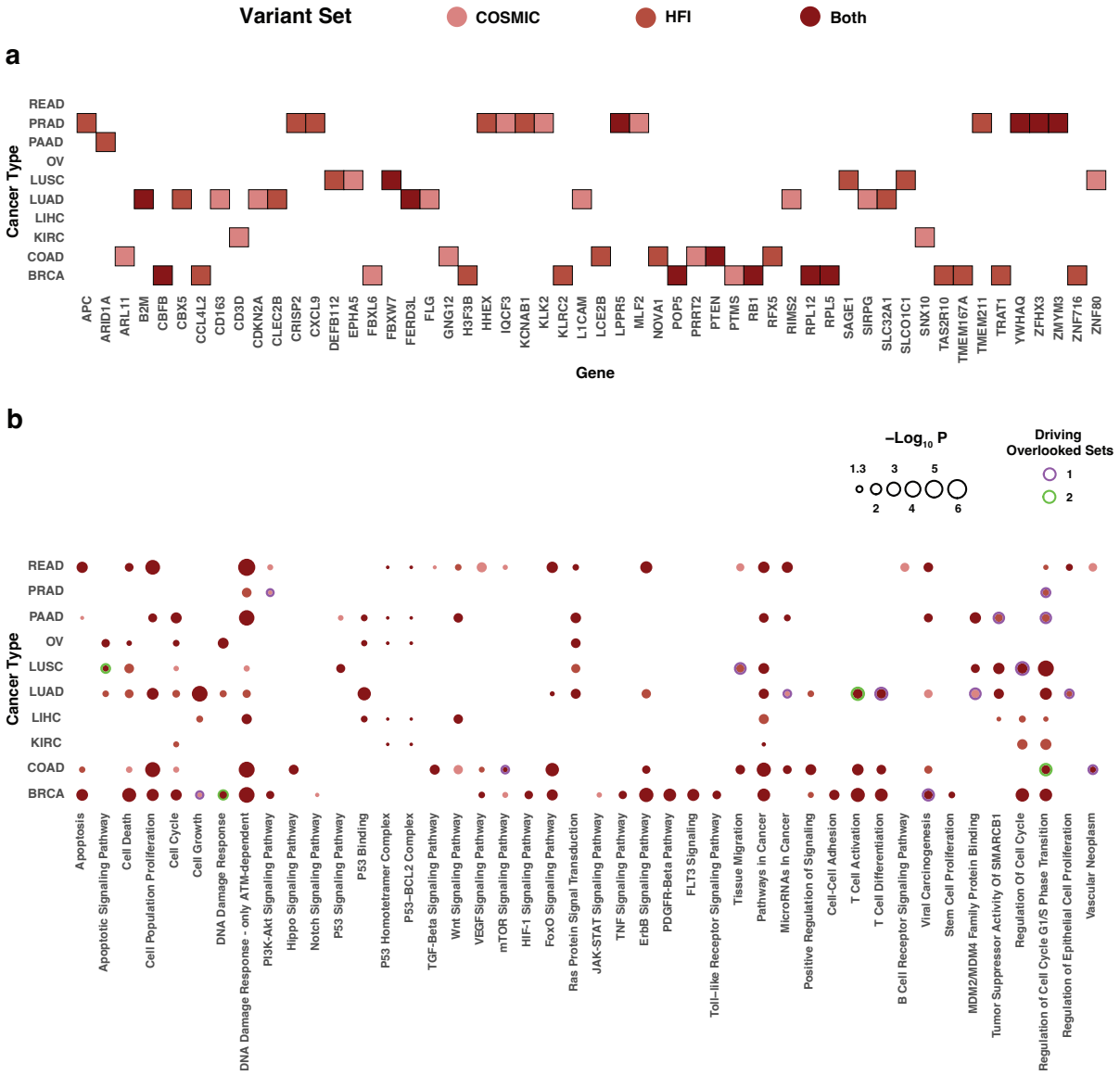

Figure 4. Known and potential driver genes identified based on dynamic clustering. (a) Dynamic cluster-based genes overlooked by the continent-based model. Each of the listed genes was identified in at least one COSMIC- or HFI-based dynamic cluster and none of the clusters based on predefined continent-based categories. (b) Dynamic cluster-based drivers associate widely with known cancer pathways. Genes overlooked by the continent-based model drive a subset of these associations in one (purple border) or both (green) settings centered on the COSMIC- and HFI-based variant sets.

332 *et al.* [42]. These pathways include cell cycle alongside Hippo, Notch, PI3K/Akt, TP53, TGFB
333 and WNT signaling. Among other important pathways and processes are apoptosis, HIF-1, mTOR,
334 TNF, JAK-STAT, VEGF, and FoxO signaling (Figure 4b), and pathways named after several
335 cancer types and other diseases and infections (e.g. Epstein-Barr and Kaposi sarcoma-associated
336 herpesvirus virus infections). We compared enrichment results in the resulting dynamically-
337 generated lists with and without novel drivers overlooked by the continent-based model. The
338 inclusion of novel genes allows dynamic clusters to refer to biological processes, pathways, and
339 entities with strong effect on cancer etiology (purple and green borders) such as apoptotic signaling
340 in LUSC, DNA damage response in BRCA, mTOR signaling in COAD, and multiple terms
341 pertaining to cell adhesion, tissue migration, cell cycle, and cell proliferation across multiple
342 cancer types.
343
344 *2.5. Dynamic clusters are distinguishable by clinical and functional cancer signifiers*
345
346 Dynamic clusters vary by the composition of their underlying germline variants. To investigate
347 the biomedical significance of the resulting clusterings, we tested for the associations between
348 each of the clusters, compared to its all its counterparts within a cancer cohort, with clinical
349 variables and gene expression patterns in TCGA. In the LUAD-COSMIC setting, the age at the
350 first pathologic diagnosis is lower in cluster 1 (C1) than 2 (mean = 64.4 and 66.6 years,
351 respectively; Wilcoxon rank-sum test, $p_{adj} < 0.05$; Figure 5a). In LIHC, C1 of the COSMIC setting
352 shows a significant enrichment for high tumor grade samples, and C1 of the HFI setting for late
353 tumor stage ones (Figure 5b and 5c, respectively; Wilcoxon rank-sum test, $p_{adj} < 0.05$), with similar
354 results that vary among clusters in LIHC and LUSC cohorts (Supplementary Figure 2).

355 We then shifted attention to analyzing differential gene expression patterns across cancer types
356 and settings. At the individual gene level, opposite expression levels are detected among clusters
357 of the same cohort (Figure 5d). Notably, such patterns correspond to genes with reported
358 associations with tumorigenesis, patient survival, metastasis, and cell proliferation. These include
359 *CSN2* [43], *NROB1* [44], and *NR1H4* [45] in both COSMIC and HFI settings in BRCA, and *MT1B*
360 [46], *MUC1* [47], and *KLK11* [48] in LIHC-COSMIC. At the gene program level, we used
361 Moonlight [49] to identify programs that are collectively differentially expressed in each cluster
362 compared to matched "normal" samples. Within different cohorts, the magnitude and direction of
363 expression mark significant differences among clusters. These include, among others (Figure 5e;
364 Supplementary Figure 3), a negative expression (Z-score < 0) of cell death in only one out of four
365 clusters in BRCA-HFI (C1), of migration of cells in only two out of eight in LIHC-COSMIC (C4
366 and C7), and of cell survival and migration of cells in one of the two clusters in COAD-HFI (C1);
367 a significantly higher expression of proliferation of cells in one cluster in each of COAD-COSMIC
368 and COAD-HFI (C2 in each); and a significantly lower expression of cell survival in LIHC-
369 COSMIC (C4), quantity of leukocytes in LUSC-COSMIC (C2), and necrosis and fatty acid
370 metabolism in LIHC-HFI (C1 and C2, respectively).
371
372 *2.6. Dynamic clusters are distinguishable by non-cancer signifiers*
373
374 In addition to cancer-focused gene programs, we investigated the biological significance of specific genes
375 significantly expressed in only one dynamic cluster within each setting ($|Log_2 FC| > 2$). Biological
376 enrichment analysis of these gene sets revealed essential and non-cancer-focused biological processes that
377 distinguish different clusters. These include, among others (Figure 6, Supplementary Figure 4), multiple
378 processes related to neuronal response to stimulus in cluster 2 (C2) of COAD-COSMIC, non-cancerous

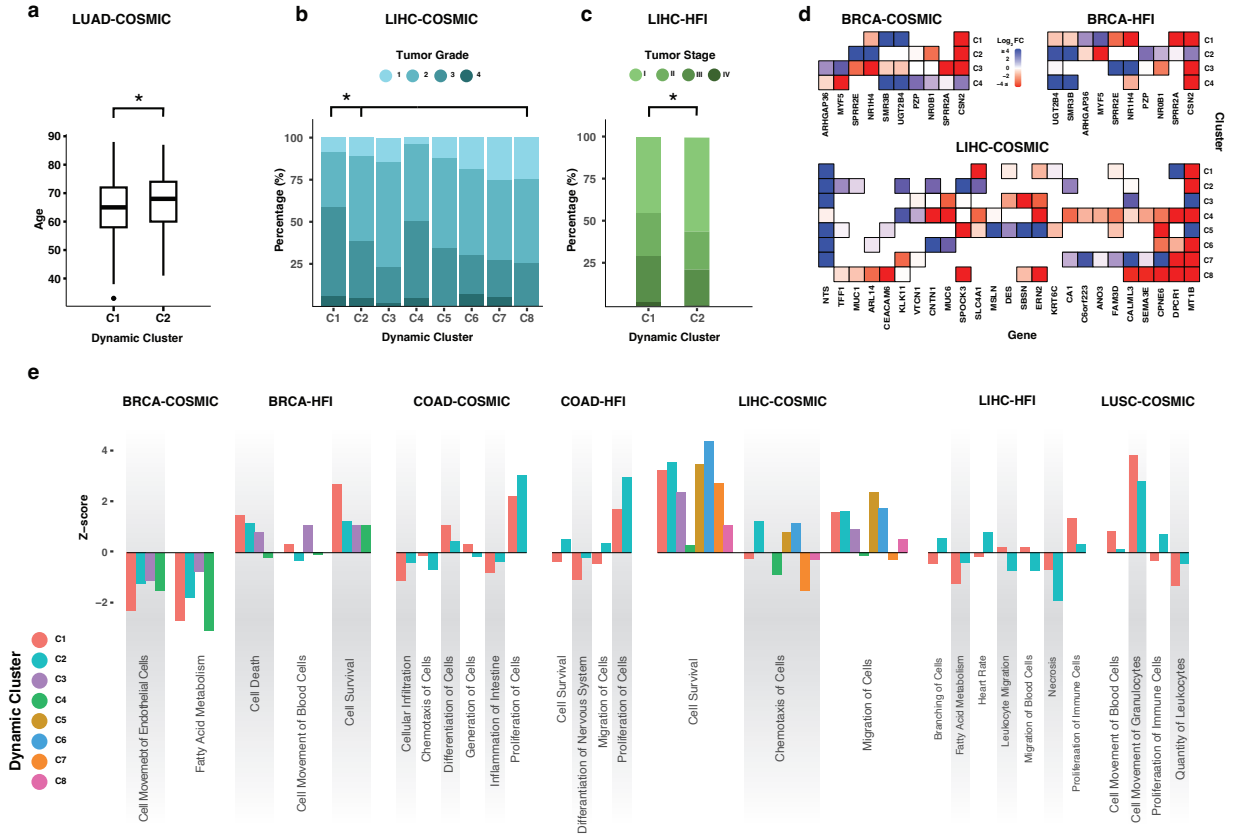

Figure 5. Dynamic clusters across cancer types highlight clinical and functional associations. a) Dynamic cluster 1 (C1) based on the COSMIC subset in LUAD (LUAD-COSMIC) shows statistically significant lower age cancer onset than that of the second cluster (C2; $p_{adj} < 0.05$) b) C1 in LIHC-COSMIC includes samples with higher tumor grade compared to other clusters combined ($p_{adj} < 0.05$). Clusters C5 and C8 include no Grade 4 samples. c) C1 shows more advanced tumor stage samples in LIHC-HFI compared to C2 ($p_{adj} < 0.05$) d) Genes significantly expressed (FDR < 0.1) in opposite directions among clusters of BRCA-COSMIC, BRCA-HFI and LIHC-COSMIC highlight potential functional roles associated with different clusters e) Gene programs with known association to cancer are collectively expressed in different magnitudes and directions between dynamic clusters across cancer types and settings.

diseases and infections in C1 of LUSC-COSMIC, and neuron development in C1 of LIHC-HFI. Other processes that closely pertain to cancer from an essential point of view include ones revolving around

immune response in C2 of READ-HFI, and cell differentiation, which recurrently emerged from individual clusters within multiple cancer types and settings (i.e. label “Across Clusters” in Figure 6).

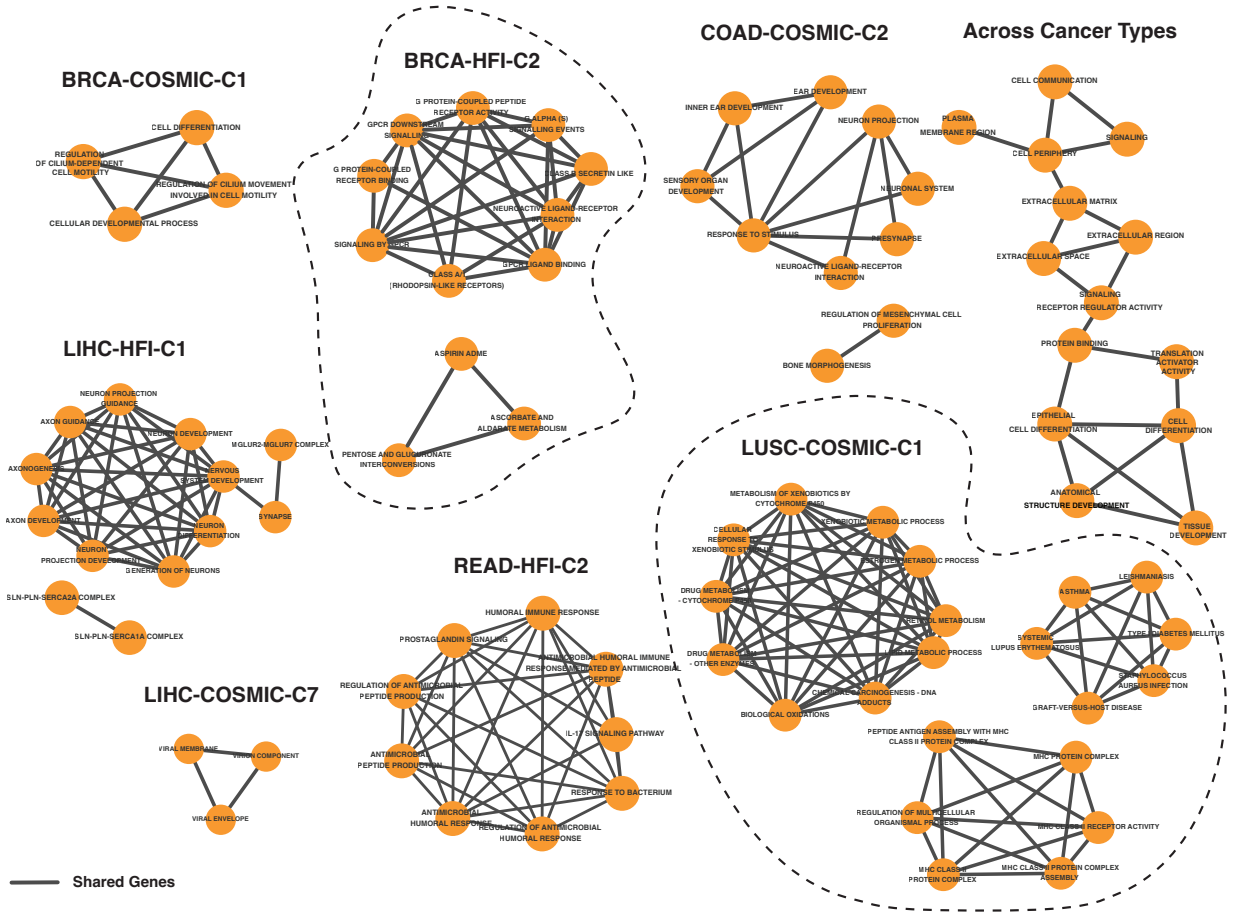

Figure 6. Genes expressed in single clusters within each cancer type and setting highlight related biological and clinical associations beyond cancer (e.g. asthma and lupus in LUSC-COSMIC-C1 and neural development in LIHC-HFI-C1). Resulting associations share a subset of their underlying genes (i.e. edges), and a number of biological processes recurrently emerges across cancer types and settings (i.e. “Across Cancer Types,” top-right).

403 ***Discussion***

404

405 Coupled with recent wide genomic data availability, existing, complex, and ever-changing human
406 biodiversity engenders the need for multiple perspectives to studying genomic traits. We introduce
407 a dynamic clustering approach that centers, in parallel, technical and social aspects of studying
408 human biodiversity by utilizing a lens that focuses on trait-based genetic similarity. This approach
409 does not lock samples within geography-based ancestry categories or average genomic patterns
410 across the whole genome regardless of the trait under study. It also recognizes continuums and
411 clusters when they exist at the trait-specific level (e.g. eight clusters for COSMIC-based LIHC vs
412 one continuous cluster for READ in Figure 1). When we apply the dynamic approach to germline
413 data of the TCGA cancer cohorts, emerging clusters transcend predefined continent-based
414 categories. When we examine the somatic mutational patterns in the resulting dynamic germline-
415 based clusterings, we identify tens of known and potential cancer driver genes, many of which
416 were overlooked by the continent-based model. Results based on trait-specific clusters also capture
417 the fundamental biology associated with the hallmarks of cancer and associate clusters with
418 clinical and biological signifiers [50].

419

420 The dynamic approach has broader implications for how biodiversity is observed and classified.
421 The use of racial classification systems in science has long been contested given their history that
422 is rife with confusion, technological limitations, and enforcement in service of colonialism. Race,
423 itself, is an idea rather than a discovery; an idea invented to impose systems of control and
424 discrimination that continue to shape today's social realities [51]. Enforced color lines disrupted
425 continuous clines of variation and often shifted to serve political goals rather than to describe

patterns of biodiversity. Further, race categories are usually collected to comply with civil rights reporting guidelines or for social and administrative purposes, but the racial categorization systems were not designed for genetic studies [2, 14]. Similarly, ethnicity categories can be centered on culture, social norms, religious beliefs, or language rather than genetic ancestry, and their use in genetic studies can lead to inaccurately reported results. Ethnicity categories are malleable concepts that can change in different times or circumstances irrespective of hereditary lines [14, 18].

A sharp decreasing trend in using race categories in genomic studies has been reported [52]. While genetic ancestry categories resemble, in their biological aspect, a direct reflection of a partial inheritance of genetic material across generational lines, they can also suffer from limitations in semantics, in space, and in time. Ancestry categories can be based on geographic, geopolitical, social, or cultural elements, and the process of imposing clear lines among predefined populations is rife with social and technical limitations. These limitations are heightened when categories are quite broad (e.g. continent-based), when individuals identify with more than one category, and when the labels might further open the door for enforced stigmatic associations of disease on entire communities [25]. As a result, it is generally advisable to cluster genomics cohorts only when justified by the research question rather than by default [10, 19], and to place social implications of the research at the heart of the design process rather than as an afterthought [53].

Given the compound nature of human biodiversity (i.e. continuums and clusters), a broader sampling of human genetic diversity—with the careful selection of categories during the data collection phase, if and as needed or obligated—remains highly central to more clearly understand

449 the patterns of genetic variation (see Chapter 5 of [10] and [54]). In fact, it is through this type of
450 diverse data collection efforts that continuous patterns of variation have been elucidated on a wider
451 scale [17, 20]. Relatedly, alleles that increase the susceptibility to a disease can be present across
452 multiple geographies. The notion of dynamic clustering can be carried over to genome-wide
453 analyses as well. For example, in the case of methods that study one variant at a time—such as
454 QTL identification and GWAS—it is possible to limit the SNP-based clustering to some
455 neighborhood of each variant to consider how the local genetic relatedness of individuals can be
456 accounted for in enrichment analysis.
457
458 While certain genomic patterns might be identified in a given dataset based on a given predefined
459 model, this type of models is not necessarily the only route towards this type of identification.
460 Equally importantly, as we demonstrate, other patterns can be missed when relying on broad
461 categories or when considering complex traits with loci distributed across the genome. Broad
462 stratification, whether driven by discriminatory legacies embedded in genetic practice or
463 normalized experimental design and data collection, might obfuscate genomic patterns that
464 transcend predefined categorical boundaries. Diverse datasets and new and existing quantitative
465 approaches that can transcend predefined categories—by utilizing or being able to incorporate
466 trait-specific regions, clines, estimated relatedness matrices (e.g. [55, 56]), principal components
467 (e.g. [57-60]), or other chosen means—are hence crucial to approach different types of genomic
468 studies (e.g. ones for gene discovery or other types of genomic studies described in [10]), to detect
469 various patterns of genomic variation, to address theoretical and applied challenges (e.g. gene-
470 environment interactions, pleiotropy, false positive control), and to leverage data generated using
471 different technologies (e.g. deep sequencing and GWAS). These efforts have the potential to draw

472 a more complete portrait of the genomic bases of traits all the while navigating the entangled
473 relationships between science and society [11, 61-63].
474
475 **Methods**
476
477 ***Genomic Datasets***
478
479 We used TCGA germline data from the BRCA (n=1072 samples after data processing), COAD
480 (445), KIRC (514), LIHC (360), LUAD(513), LUSC(503), OV (556), PAAD (182), PRAD (488),
481 and READ (164) studies. We used the MC3 somatic dataset [64] for functional enrichment analysis
482 and the ancestral labels from the TCGAA Project (http://fcgportal.org/TCGAA/) [33] filtered
483 according to the recommendations in [65].
484
485 ***SNP Selection***
486
487 For COSMIC-based SNP sets, we selected autosomal SNPs annotated as nonsynonymous, stop-
488 gain, and stop-loss ClinVar database annotations [66]. For HFI subsets, the functional impact of
489 missense germline variants within a cancer type's exome samples was determined using MetaSVM
490 [67] and annotations by ClinVar, when available. We considered a missense variant high-
491 functional impact if classified as Deleterious by MetaSVM or listed as Pathogenic/Likely-
492 Pathogenic in ClinVar. We used MetaSVM scores from the dbNSFP database which includes pre-
493 calculated function impact scores for 75,931,005 human non-synonymous single-nucleotide
494 variants [68]. Loss-of-function (LoF) variants, including frameshift indels, stop gain, and stop loss

495 variants, were also considered high-functional impact as well as variants annotated as high-
496 confidence LoF in gnomAD [69] or Pathogenic/Likely-Pathogenic in ClinVar. We only included
497 autosomal variants that met quality control measures described in the source paper by Huang *et al.*
498 [30]. The number of selected SNPs in each COSMIC and HFI variant set is available in
499 Supplementary Table 4.
500
501 ***Driver Gene Identification***
502
503 Potential driver gene identification was performed on the MutSigCV [70] v1.3.4 server available
504 at https://www.genepattern.org/modules/docs/MutSigCV. Genes with FDR < 0.1 were deemed
505 statistically significant potential drivers.
506
507 ***Clinical Variable Analysis***
508
509 Clinical TCGA data was obtained from [71]. Continuous and ordinal variable comparisons (i.e.
510 for age, tumor grade, and tumor stage) were performed using the Wilcoxon rank-sum test in a one-
511 vs-all configuration on clusters with > 5% of samples within a cohort and COSMIC or HFI setting,
512 with Bonferroni correction and $p_{adj} < 0.05$ significance level.
513
514 ***Gene Expression Analysis***
515
516 Gene expression data corrected for batch effect and study-specific bias were downloaded from
517 RNAseqDB [72] at https://github.com/mskcc/RNAseqDB. Differential expression analysis at the

518 gene program level Moonlight v1.20 [49], and at the gene level was performed using edgeR v3.36
519 [73] ($q < 0.05$, $|\text{Log}_2 \text{FC}| > 2$).
520
521 ***Enrichment Analysis***
522
523 Enrichment analysis to identify pathways and biological processes was performed on g:Profiler
524 available at https://biit.cs.ut.ee/gprofiler/, with entities having g:SCS threshold < 0.05 considered
525 significant [74]. Visualization was done using ggplot2 [75], except for Figure 6 generated using
526 EnrichmentMap v.3.5.0 plugin [76] in Cytoscape v 3.10.3 [77].
527
528 ***Data Availability***
529
530 The results published here are in whole or part based upon data generated by the TCGA Research
531 Network: https://www.cancer.gov/tcga. Controlled-access germline variants of TCGA cohorts
532 were downloaded from the Genomic Data Commons (GDC, https://gdc.cancer.gov/about-
533 data/publications/PanCanAtlas-Germline-AWG) of the National Cancer Institute (NCI) per Huang
534 *et al.* [30]. TCGA variant and meta-data are available through the GDC portal at
535 https://portal.gdc.cancer.gov/. Ancestry categories were obtained from TCGAA [33] available at
536 http://fcgportal.org/TCGAA/. We chose to not analyze samples labeled "Native American [NA]"
537 out of respect for Indigenous sovereignty.
538

Genes with germline associations at the tissue-specific level were downloaded from the COSMIC v90 [78] census list's 'Germline' column. Full gene lists (n = 2 to 11) are available in Supplementary Table 3.

Code Availability

We used PLINK v1.90 [79] available at https://zzz.bwh.harvard.edu/plink/ to calculate IBS matrices used to generate MDS plots. DBSCAN clusters were generated using the "dbscan" package in R (https://cran.r-project.org/web/packages/dbscan/index.html) [80], and HClust and K-Means using "stats" (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html).

References

1. Roberts, D.E., *Fatal invention: How Science, Politics, and Big Business Re-create Race in the Twenty-first Century*. 2011, New York: New Press.
2. Jablonski, N.G., *Skin color and race.* Am J Phys Anthropol, 2021. **175**(2): p. 437-447.
3. Marks, J., *Long shadow of Linnaeus's human taxonomy.* Nature, 2007. **447**(7140): p. 28-28.
4. Anemone, R.L., *Race and Human Diversity: A Biocultural Approach*. 2019, Oxford and New York: Routledge.
5. Sax, B., *When Adam and Eve Were Monkeys: Anthropomorphism, zoomorphism, and other ways of looking at animals*, in *The Routledge companion to animal-human history*, H. Kean and P. Howell, Editors. 2018, Routledge/Taylor & Francis Group,: London ; New York.
6. HoSang, D.M., *On Racial Speculation and Racial Science: A Response to Shiao et al.* Sociological Theory, 2014. **32**(3).
7. Gates, H.L. and A.S. Curran, *Who's Black and Why?: A Hidden Chapter from the Eighteenth-century Invention of Race*. 2022, Cambridge, Massachusetts: The Belknap Press of Harvard University Press.
8. Curran, A.S., *The Anatomy of Blackness: Science & Slavery in an Age of Enlightenment*. 2011, Baltimore: Johns Hopkins University Press.
9. Hogarth, R.A., *Medicalizing Blackness: Making Racial Difference in the Atlantic World, 1780-1840*. 2017, Chapel Hill: The University of North Carolina Press.

10. National Academies of Sciences, Engineering, Medicine, *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. 2023, Washington, DC: The National Academies Press.
11. Reardon, J., *The Human Genome Diversity Project: A case study in Coproduction.* Social Studies of Science, 2001. **31**(3): p. 357-388.
12. Jasanoff, S., *The Idiom of Co-Production*, in *States of Knowledge: The Co-Production of Science and the Social Order*, S. Jasanoff, Editor. 2004, Routledge. p. 1-12.
13. Yudell, M., *Race Unmasked: Biology and Race in the Twentieth Century*. 2014, New York: Columbia University Press.
14. Race, Ethnicity, and Genetics Working Group, *The use of racial, ethnic, and ancestral categories in human genetics research.* Am J Hum Genet, 2005. **77**(4): p. 519-32.
15. Crawford, N.G., et al., *Loci associated with skin pigmentation identified in African populations.* Science, 2017. **358**(6365): p. eaan8433.
16. Jacobs, L.C., et al., *Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans.* Human Genetics, 2013. **132**(2): p. 147-158.
17. Wojcik, G.L., et al., *Genetic analyses of diverse populations improves discovery for complex traits.* Nature, 2019. **570**(7762): p. 514-518.
18. Baker, J.L., C.N. Rotimi, and D. Shriner, *Human ancestry correlates with language and reveals that race is not an objective genomic classifier.* Scientific Reports, 2017. **7**(1).
19. Sohail, M., A. Izarraras-Gomez, and D. Ortega-Del Vecchyo, *Populations, Traits, and Their Spatial Structure in Humans.* Genome Biology and Evolution, 2021. **13**(12).
20. Lewis, A.C.F., et al., *Getting genetic ancestry right for science and society.* Science, 2022. **376**(6590): p. 250-252.
21. Serre, D. and S. Pääbo, *Evidence for gradients of human genetic diversity within and among continents.* Genome Res, 2004. **14**(9): p. 1679-85.
22. Shriner, D., et al., *Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry.* Scientific Reports, 2014. **4**(1).
23. Livingstone, F.B. and T. Dobzhansky, *On the Non-Existence of Human Races.* Current Anthropology, 1962. **3**(3).
24. Cotter, D.J., et al., *A rarefaction approach for measuring population differences in rare and common variation.* Genetics, 2023. **224**(2).
25. Kader, F.Đ., Lan N.; Lee, Matthew; Chin, Matthew K.; Kwon, Simona C.; Yi, Stella S., *Disaggregating Race/Ethnicity Data Categories: Criticisms, Dangers, And Opposing Viewpoints.* Health Affairs Forefront, 2022.
26. Alelyani, S.T., Jiliang; Liu, Huan, *Feature Selection for Clustering: A Review*, in *Data Clustering: Algorithms and Applications*, C.C.R. Aggarwal, Chandan K., Editor. 2014, Chapman and Hall/CRC: New York, NY.
27. Ultsch, A. and J. Lötsch, *The Fundamental Clustering and Projection Suite (FCPS): A Dataset Collection to Test the Performance of Clustering and Data Projection Algorithms.* Data, 2020. **5**(1).
28. ICGC-TCCA Pan-Cancer Analysis of Whole Genomes Consortium, *Pan-cancer analysis of whole genomes.* Nature, 2020. **578**(7793): p. 82-93.
29. Qing, T., et al., *Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden.* Nature Communications, 2020. **11**(1).
30. Huang, K.L., et al., *Pathogenic Germline Variants in 10,389 Adult Cancers.* Cell, 2018. **173**(2): p. 355-370 e14.
31. Carter, H., et al., *Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer.* Cancer Discov, 2017. **7**(4): p. 410-423.
32. NCI SEER. *Cancer Stat Facts: Cancer of Any Site*. 2022; Available from: https://seer.cancer.gov/statfacts/html/all.html.

33. Yuan, J., et al., *Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers.* Cancer Cell, 2018. **34**(4): p. 549-560 e9.
34. The 1000 Genomes Project Consortium, *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
35. Cavalli-Sforza, L.L., et al., *Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project.* Genomics, 1991. **11**(2).
36. International HapMap Consortium, *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
37. Pon, J.R. and M.A. Marra, *Driver and passenger mutations in cancer.* Annu Rev Pathol, 2015. **10**: p. 25-50.
38. Bailey, C.E., et al., *Increasing disparities in the age-related incidences of colon and rectal cancers in the United States, 1975-2010.* JAMA Surg, 2015. **150**(1): p. 17-22.
39. Myer, P.A., et al., *The Genomics of Colorectal Cancer in Populations with African and European Ancestry.* Cancer Discov, 2022. **12**(5): p. 1282-1293.
40. Xu, D. and Y. Tian, *A Comprehensive Survey of Clustering Algorithms.* Annals of Data Science, 2015. **2**(2): p. 165-193.
41. Bailey, M.H., et al., *Comprehensive Characterization of Cancer Driver Genes and Mutations.* Cell, 2018. **173**(2): p. 371-385.e18.
42. Sanchez-Vega, F., et al., *Oncogenic Signaling Pathways in The Cancer Genome Atlas.* Cell, 2018. **173**(2): p. 321-337.e10.
43. Bach, K., et al., *Time-resolved single-cell analysis of Brca1 associated mammary tumourigenesis reveals aberrant differentiation of luminal progenitors.* Nat Commun, 2021. **12**(1): p. 1502.
44. Zhang, H., et al., *The prognostic value of the orphan nuclear receptor DAX-1 (NROB1) in node-negative breast cancer.* Anticancer Res, 2011. **31**(2): p. 443-9.
45. Journe, F., et al., *Association between farnesoid X receptor expression and cell proliferation in estrogen receptor-positive luminal-like breast cancer from postmenopausal patients.* Breast Cancer Research and Treatment, 2009. **115**(3): p. 523-535.
46. Xu, P., et al., *Copy number variation of metallothionein 1 (MT1) associates with MT1X isoform expression and the overall survival of hepatocellular carcinoma patients in Guangxi.* Gene Reports, 2024. **34**: p. 101889.
47. Yuan, S.F., et al., *Expression of MUC1 and its significance in hepatocellular and cholangiocarcinoma tissue.* World J Gastroenterol, 2005. **11**(30): p. 4661-6.
48. Ke, H., et al., *Serum protein biomarkers for HCC risk prediction in HIV/HBV co-infected people: a clinical proteomic study using mass spectrometry.* Front Immunol, 2023. **14**: p. 1282469.
49. Colaprico, A., et al., *Interpreting pathways to discover cancer driver genes with Moonlight.* Nature Communications, 2020. **11**(1): p. 69.
50. Hanahan, D. and Robert A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.
51. McLean, S.-A., *Isolation by Distance and the Problem of the Twenty-First Century.* Human Biology, 2019. **91**(2): p. 81-94.
52. Byeon, Y.J.J., et al., *Evolving use of ancestry, ethnicity, and race in genetics research—A survey spanning seven decades.* The American Journal of Human Genetics, 2021. **108**(12): p. 2215-2223.
53. Reardon, J., *Human Population Genomics and the Dilemma of Difference*, in *Reframing Rights: Bioconstitutionalism in the Genetic Age*, S. Jasanoff, Editor. 2011, MIT Press: Cambridge, Massachusetts; London, Englad. p. 217-238.
54. Oni-Orisan, A., et al., *Embracing Genetic Diversity to Improve Black Health.* New England Journal of Medicine, 2021. **384**(12): p. 1163-1167.

55. Kang, H.M., et al., *Variance component model to account for sample structure in genome-wide association studies.* Nat Genet, 2010. **42**(4): p. 348-54.
56. Zhou, X. and M. Stephens, *Genome-wide efficient mixed-model analysis for association studies.* Nat Genet, 2012. **44**(7): p. 821-4.
57. Jiang, L., et al., *A resource-efficient tool for mixed model association analysis of large-scale data.* Nature Genetics, 2019. **51**(12): p. 1749-1755.
58. Mbatchou, J., et al., *Computationally efficient whole-genome regression for quantitative and binary traits.* Nat Genet, 2021. **53**(7): p. 1097-1103.
59. Zhao, H., et al., *A practical approach to adjusting for population stratification in genome-wide association studies: principal components and propensity scores (PCAPS).* Statistical Applications in Genetics and Molecular Biology, 2018. **17**(6).
60. Zhou, W., et al., *Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies.* Nature Genetics, 2018. **50**(9): p. 1335-1341.
61. Yudell, M., et al., *Taking race out of human genetics.* Science, 2016. **351**(6273): p. 564-565.
62. Thorp, H.H., *Time to look in the mirror*, in *Science*. 2020.
63. Nobles, M.W., Chad; Wonkam, Ambroise; Wathuti, Elizabeth, *Science must overcome its racist legacy: Nature's guest editors speak*, in *Nature*. 2022.
64. Ellrott, K., et al., *Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines.* Cell Syst, 2018. **6**(3): p. 271-281 e7.
65. Bailey, M.H., et al., *Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples.* Nature Communications, 2020. **11**(1): p. 4748.
66. Landrum, M.J., et al., *ClinVar: public archive of interpretations of clinically relevant variants.* Nucleic Acids Res, 2016. **44**(D1): p. D862-8.
67. Kim, S., et al., *Meta-analytic support vector machine for integrating multiple omics data.* BioData Min, 2017. **10**: p. 2.
68. Liu, X., et al., *dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.* Human Mutation, 2016. **37**(3): p. 235-241.
69. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.
70. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-218.
71. Liu, J., et al., *An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.* Cell, 2018. **173**(2): p. 400-416.e11.
72. Wang, Q., et al., *Unifying cancer and normal RNA sequencing data from different sources.* Scientific Data, 2018. **5**(1): p. 180061.
73. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
74. Raudvere, U., et al., *g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update).* Nucleic Acids Research, 2019. **47**(W1): p. W191-W198.
75. Wickman, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016, New York, NY: Springer-Verlag.
76. Reimand, J., et al., *Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap.* Nature Protocols, 2019. **14**(2): p. 482-517.
77. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

78. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer.* Nucleic Acids Res, 2019. **47**(D1): p. D941-D947.
79. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
80. Hahsler, M.P., Matthew; Arya, Sunil; Mount, David, *dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms.* 2022.

Acknowledgements

No acknowledgements to include in this section.

Author Contributions

Conceptual basis: HM. Methods and experimental design: HM, KB and LP. Experiment execution and data analysis: HM. Discussions: HM, KB, PE, QM, JCZ, and LP. First manuscript draft: HM and LP. Final manuscript: HM and LP with input from KB, PE, QM, and JCZ.

Competing Interests

LP has received consulting fees and honoraria for advisory board participation from Pfizer, Astra Zeneca, Merck, Novartis, Bristol-Myers Squibb, GlaxoSmithKline, Genentech, Personalis, Daiichi, Natera, Exact Sciences and institutional research funding from Seagen, GlaxoSmithKline, AstraZeneca, Merck, Pfizer and Bristol Myers Squibb.

Corresponding Author

Hussein Mohsen (MohsenH@mskcc.org)