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 pre-defined 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 pre-defined categories. We tested the approach on whole-exome sequencing datasets in eight 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 analyses based on cancer type-specific clusterings were also able to capture the fundamental biology underlying cancer and to identify 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 pre-defined 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.

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 Lapps of Norway, 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 pre-defined 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 pre-defined 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 pre-defined 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 pre-defined 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

187 small subset of genomic regions? Further, it is also quantitatively well-established that selected
188 features or clustering criteria strongly affect resulting "populations" (i.e. clusters or clines), and
189 consequently reshape the starting point from which to discover—or miss—patterns and generate
190 hypotheses [26, 27].
191
192 The dynamic clustering approach takes a different angle to classifying biodiversity by grouping
193 individuals based on predisposition to a trait under study—herein a cancer type. An individual's
194 membership to a cluster depends on their genomic sequence at a specific set of regions known to
195 be associated with the trait. Number of clusters, which are *de facto* neutrally labelled, is determined
196 according to the dataset and specific trait under study.
197
198 In cancer, germline (inherited) predisposition is mediated by deleterious mutations in several
199 dozen high penetrance cancer-relevant genes and probably a combination of individually low
200 penetrance variants. Different genes are associated with different degrees of risk, and with variable
201 cancer-specificity [28]. Further, germline alterations require additional acquired (somatic)
202 mutations for malignant transformation [29-31]. We hypothesize that clustering cancers based on
203 their germline variants in cancer type-specific loci transcends pre-defined continental categories.
204 We expect clusterings to vary across cancers—in terms of number of clusters and sample-cluster
205 membership—to reflect human biodiversity and heterogeneity across complex traits. We also note
206 that this dynamic (i.e. trait-specific) approach moves beyond the notion of local ancestry at a single
207 locus as it can simultaneously consider any set of coding or non-coding regions associated with a
208 trait, and it can scale to accommodate newly acquired knowledge on the genomic basis of the trait
209 under study.

Results

2.1. Overview

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

We first generated the trait-specific clusterings of each cohort, using the TCGA terminology for cancer types, BRCA, COAD, KIRC, LIHC, OV, PAAD, PRAD, and READ, using SNPs within subsets of COSMIC genes known to have germline association with each cancer type. Next, we focused on the SNP subset predicted to have high functional impact within each cancer type's samples as a basis for dynamically generating clusters. We then assessed the performance of three algorithmic clustering approaches (K-means, DBSCAN, and Hclust) to identify generated clusters in multidimensional scaling (MDS) plots. Finally, we identified potentially overlooked somatic driver genes in each TCGA cohort based on dynamic clustering in comparison with drivers identified using the continent-based model, and performed a functional genomic analysis to assess their biological importance in the context of cancer.

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 rectum and colon cancers, known to be associated with high disparities in incidence and outcome
242 [38-40], with one and two clusters (Supplementary Figures 7 and 2), 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 each TCGA cohort as bases for clustering (see Methods). This selection led to SNP subsets
248 with n= 36 (PAAD) to 138 (BRCA). HFI-based dynamic clusters transcend continent-based
249 boundaries and exhibit two notable patterns. First, subsections within select resulting clusters
250 (subclusters) tend to be highly compact and often include samples with highly similar or identical
251 HFI variant patterns. Dots corresponding to distinct samples overlay each other, resulting in single
252 dots each representing a subcluster (e.g. Figure 2). Second, the number of HFI-based clusters is
253 higher, on average, than clusters based on SNPs of COSMIC genes (e.g. PAAD with one
254 COSMIC-based vs five HFI-based clusters). This is likely due to the increased effect of single

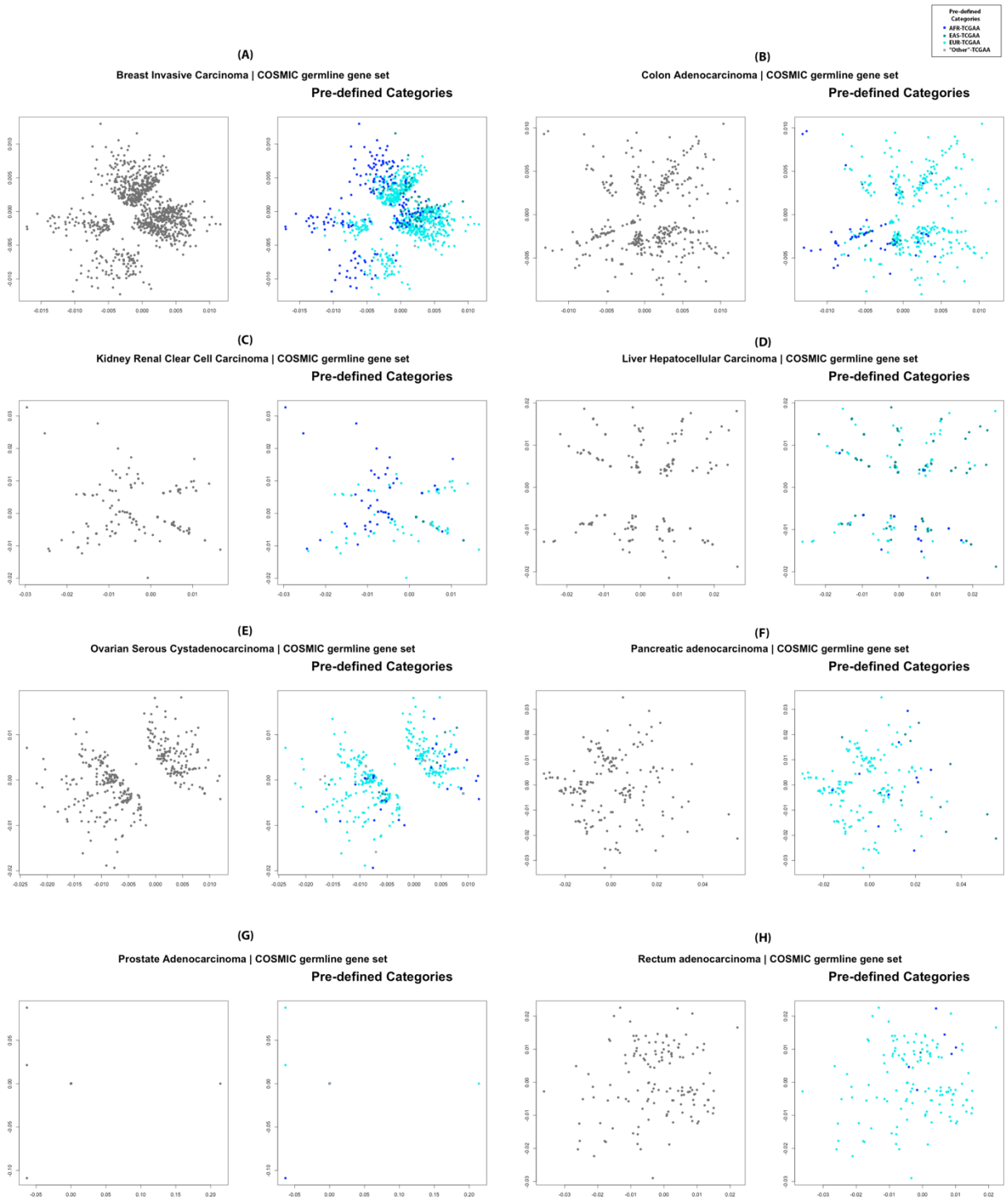


Figure 1. Multidimensional scaling (MDS) plots of dynamically generated clusters for eight TCGA cancer cohorts. Cancer type-specific dynamic clusters transcend pre-defined continent-based categories.

259

260 SNPs on sample-cluster membership when the number of loci used for clustering is small, resulting
261 in higher stratification within the same cohort. HFI-based results for LIHC are shown in Figure 2,
262 and similar patterns are observed for other cancer types (Supplementary Figures 8 to 15).

263

264 *2.4. Human aid improves algorithmic clustering*

265

266 Algorithmic approaches can generate different clusterings of the same dataset, and their
267 performance primarily depends on data distribution and cluster definition [27, 41]. To
268 algorithmically identify clusters, we ran the K-means algorithm with a pre-defined k = 5 (number
269 of ancestral categories in TCGAA) and an "optimal" k chosen based on visual inspection of cancer
270 type-specific plots. Given the varying number of clusters across cancer types, pre-defined-k
271 clustering performed poorly (e.g. Figure 3A). Dynamic-k clustering identified a considerable
272 subset, yet not all, of the dynamic clusters. Namely, it was able to identify clusters in six out of the
273 eight COSMIC-based cases (Supplementary Figures 1, 2 and 4-7), and it faced limitations in the
274 remaining two (e.g. Figure 3B) and in seven out of eight HFI-based clusters (Supplementary
275 Figures 9-15).

276

277 Clustering results from two other algorithms, DBSCAN (Density-based Spatial Clustering of
278 Applications with Noise) and Hclust, faced similar limitations and expectedly varied in
279 performance. Hclust, which stands for agglomerative hierarchical clustering—herein used with
280 complete-linkage, is a bottom-up approach that starts with individual points as separate clusters
281 and iteratively merges most similar clusters until a pre-defined number of clusters is met (e.g. k =

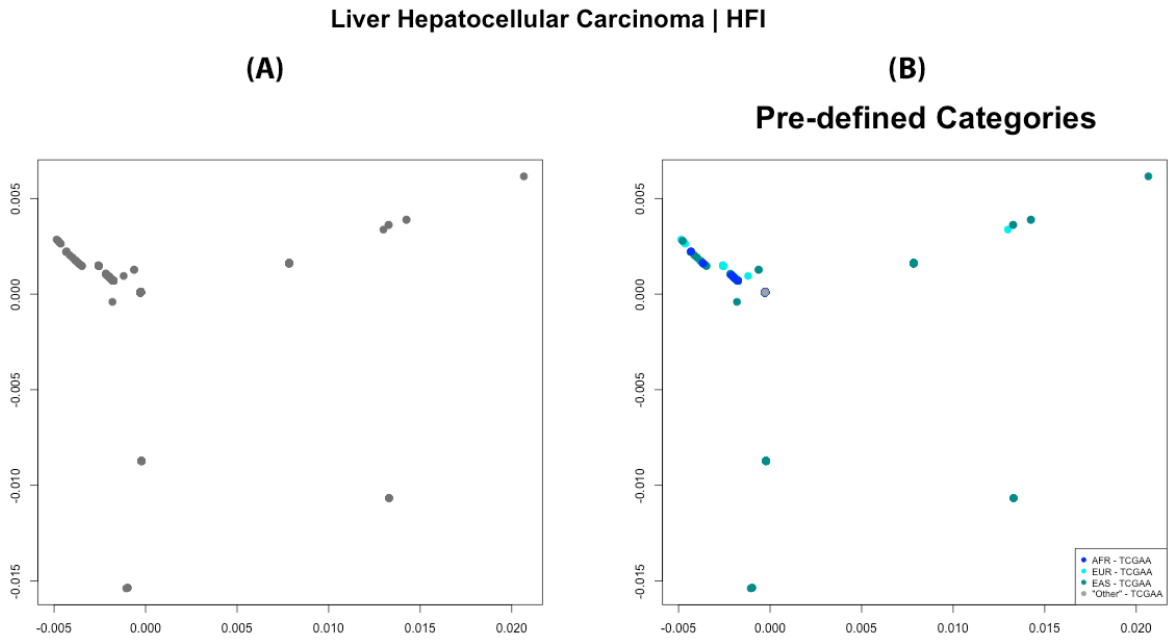


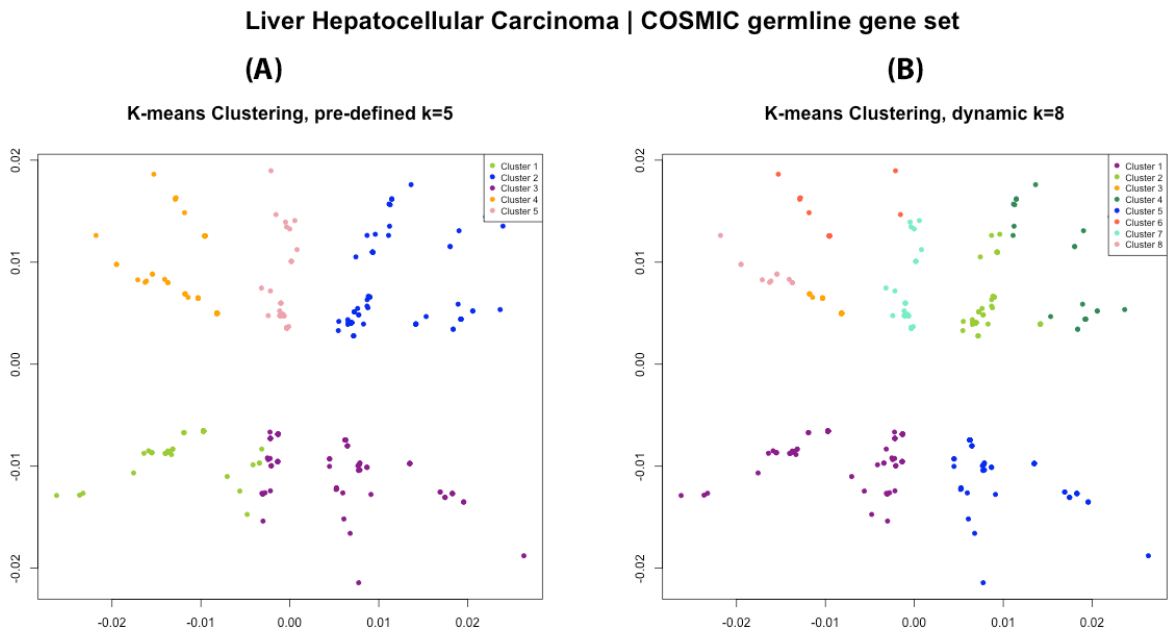
Figure 2. MDS plots of dynamically generated clusters for LIHC based on high functional impact germline variant subsets. (B) is labeled using continent-based categories provided in TCGAA, and single dots in both plots can correspond to multiple overlaid samples. Cancer type-specific stratified clusters transcend pre-defined continent-based categories and are more compact than COSMIC-based ones.

8 clusters for COSMIC-based LIHC, Figure 4). Like K-means, Hclust correctly identified only a subset (i.e. three) of the eight clusters in COSMIC-based LIHC results. DBSCAN is known to identify dense clusters and outliers in low-density regions. The algorithm faced limitations when dynamic clusters in TCGA cohorts had a relatively low density, and it identified three clusters in Figure 4B: two large clusters, each roughly with four subclusters, and a third cluster that includes distant samples the algorithm considered outliers (in green). DBSCAN faced more challenges with

295 identifying dynamic clusters than its two counterparts across other cancer types (Supplementary

296 Figures 15-21).

297



298

299 **Figure 3.** K-means results with (A) pre-defined k = 5 (number of ancestral categories in TCGAA)

300 and (B) k = 8 (number of clusters that emerged with COSMIC-based dynamic clustering).

301

302 While specific algorithms perform better than others at identifying clusters, algorithmic clustering

303 does not seem to suffice. As a result, this highlights a remaining necessity for human intervention

304 to attain more precise results. We highlight the importance of this intervention for COSMIC- and

305 HFI-based clusters for LIHC, respectively with eight and seven clusters, in Figure 5 (for other

306 cancer types, see Supplementary Figures 3 and 9-14). We also note that in certain instances,

307 multiple "optimal" numbers of clusters in the same plot can exist, and the choice remains centered

308 on experimental goals and the problem under study (e.g. the selected four clusters in the BRCA

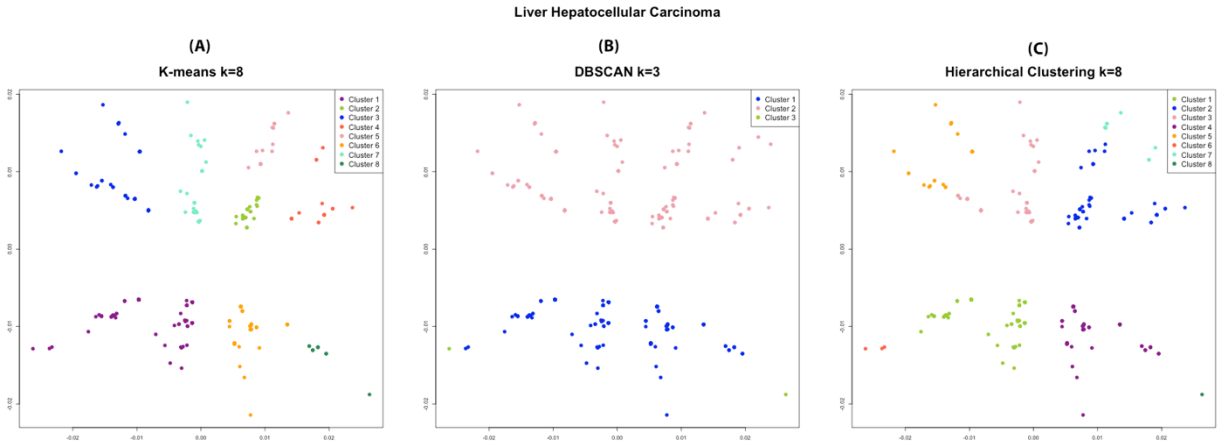


Figure 4. Classification of LIHC samples using (A) K-means, (B) DBSCAN, and (C) Hierarchical Clustering overlaid on COSMIC-based dynamic clusters.

COSMIC-based plots of Supplementary Figure 1, or two larger diagonally-separated clusters in the same plot).

2.5. Dynamically-identified cancer drivers

We next explored the biological significance of clusters identified by the dynamic clustering approach. Particularly, we focused on identifying potential cancer type-specific somatic driver genes and their functional importance.

We used MutSigCV to identify potential somatic drivers based on whole exome sequence data corresponding to each dynamic cluster generated on germline variant sets (COSMIC genes and HFI). COSMIC-based clusters yielded 71 potential drivers across cancer types, and HFI-based clusters 114 drivers. Both lists included tens of known drivers from a more comprehensive list by

Bailey *et al.* which relied on multiple computational and experimental tools [42]. These include *KRAS*, *TP53*, *PIK3CA*, *BRCA1*, *PTEN*, *CDH1*, *RB1*, *PTEN*, *CASP8*, and *VHL* (for full lists, see Supplementary Tables 1 and 2). COSMIC- and HFI-based lists also include 15 and 44 cancer type-specific novel potential drivers, respectively, which are overlooked if analysis is performed on

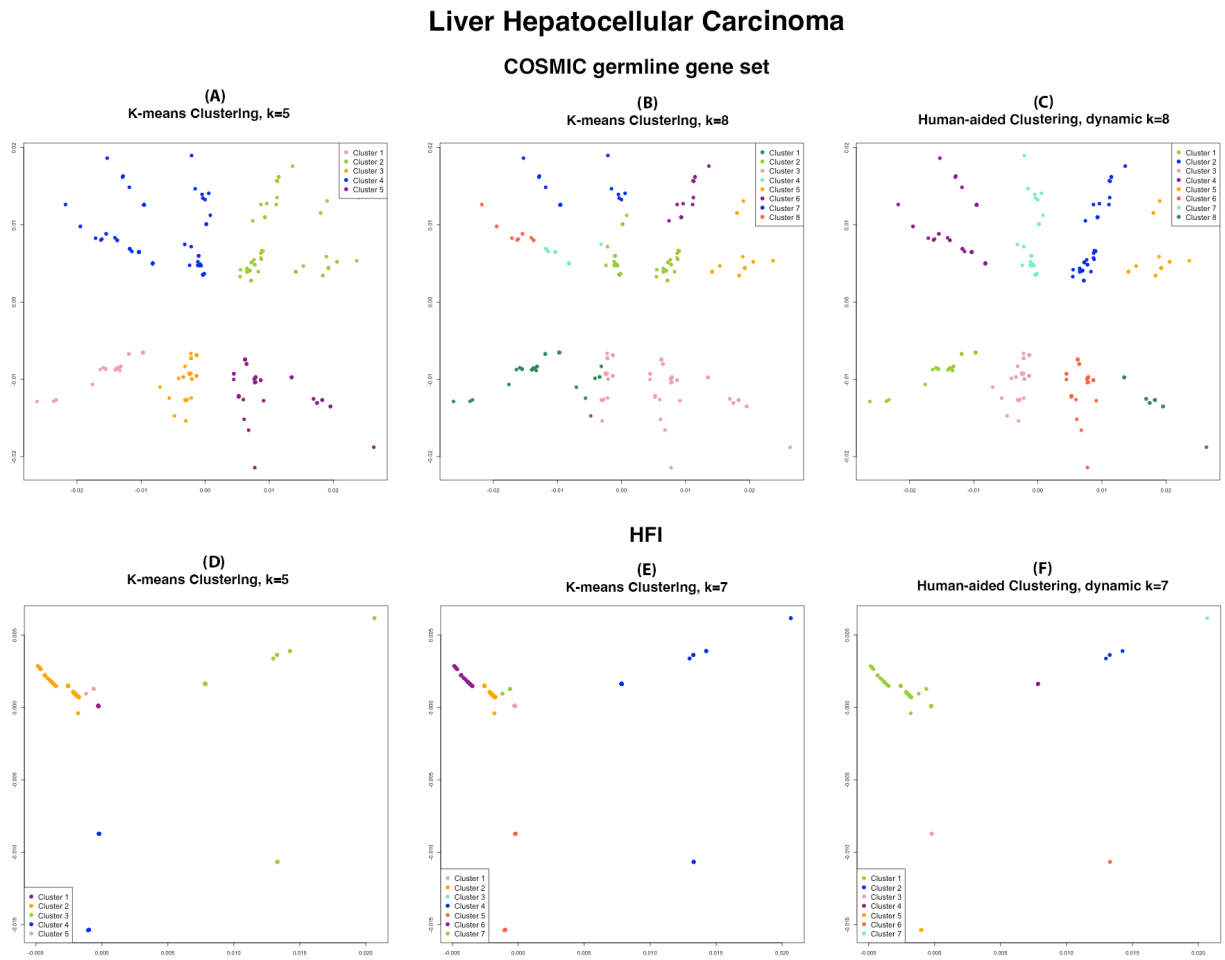


Figure 5. Classification of LIHC samples in COSMIC- (A-C, pre-defined k = 5, dynamic/ "optimal" k = 8) and HFI-based (D-F, pre-defined k = 5, dynamic/"optimal" k = 7) clusterings using K-means. Results highlight the necessity of visual inspection to identify clusters (C and F).

continent-based clusters (Figure 6). Among these genes are known drivers listed in [42], including *APC*, *CBFB*, and *KDM6A*. Notably, HFI-based clusters generated more potential drivers than continent-based clusters across cancer types, while COSMIC-based ones rank third in terms of number of suggested novel driver genes.

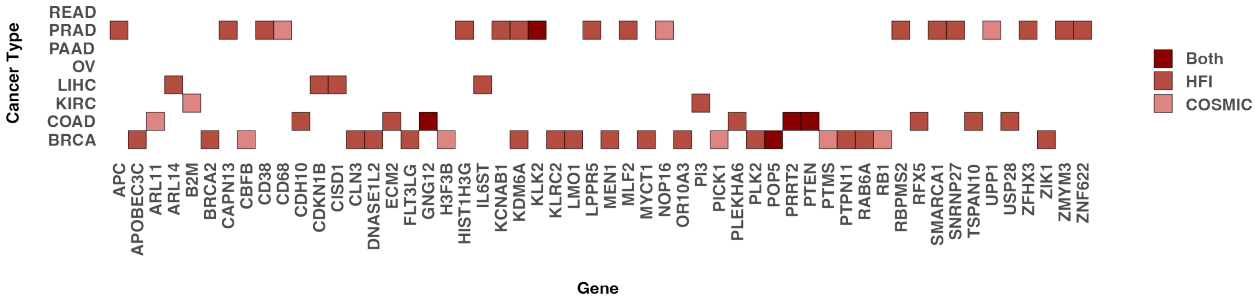


Figure 6. Known and potential driver genes identified based on dynamic clustering and overlooked by the continent-based model. Each of the listed genes was identified in at least one of the cancer type-specific dynamic clusters.

We then investigated the functional importance of COSMIC- and HFI-based driver gene lists. Given their significant coverage of known drivers, enrichment analysis of both full lists point to known oncogenic pathways, including the majority of signaling pathways listed in Sanchez-Vega *et al.* [43]. These pathways include cell cycle alongside Hippo, Notch, PI3K/Akt, RTK-RAS, TP53, TGFB and WNT signaling. Among other important pathways and processes are apoptosis, FoxO signaling, EGFR tyrosine kinase inhibitor resistance, MAPK cascade, and pathways named after several cancer types and other diseases and infections (e.g. Epstein-Barr and Kaposi sarcoma-associated herpesvirus virus infections). Second, we compared enrichment results in the resulting dynamically-generated lists with and without novel drivers overlooked by the continent-based

356 model. The inclusion of novel genes allows dynamic clusters to refer to biological processes,
357 pathways, and entities with strong effect on cancer etiology such as HIF, and mTOR signaling
358 pathways, the SWI/SNF complex, and multiple terms referring to essential and cancer-associated
359 terms pertaining to tissue migration, cell cycle, and cell growth. Several important oncogenic
360 signaling pathways listed in [43] are also pulled from data annotation sources into enrichment
361 results thanks to novel genes, further indicating their potential importance. These pathways include
362 TP53, PI3K/Akt, and WNT signaling.
363
364 ***Discussion***
365
366 Coupled with recent wide genomic data availability, existing, complex, and ever-changing human
367 biodiversity engenders the need for multiple perspectives to studying genomic traits. We introduce
368 a dynamic clustering approach that centers, in parallel, technical and social aspects of studying
369 human biodiversity by utilizing a lens that focuses on trait-based genetic similarity. This approach
370 does not lock samples within geography-based ancestry categories or average genomic patterns
371 across the whole genome regardless of the trait under study. It also recognizes continuums and
372 clusters when they exist at the trait-specific level (e.g. eight clusters for COSMIC-based LIHC in
373 Figure 5 vs one continuous cluster for PAAD in Supplementary Figure 5). When we apply the
374 dynamic approach to germline data of the TCGA cancer cohorts, emerging clusters transcend pre-
375 defined continent-based categories reinforcing that particular SNP patterns co-occur across
376 populations, and that multiple distinct germline SNP patterns exist within each cancer type. When
377 we examined the somatic mutational patterns in the resulting dynamic germline-based clusterings,
378 we identified tens of known and potential cancer driver genes, many of which were overlooked by

379 the continent-based model. Results based on trait-specific clusters also captured the fundamental
380 biology associated with the hallmarks of cancer [44].
381
382 The dynamic approach has broader implications for how biodiversity is observed and classified.
383 The use of racial classification systems in science has long been contested given their history that
384 is rife with confusion, technological limitations, and enforcement in service of colonialism. Race,
385 itself, is an idea rather than a discovery; an idea invented to impose systems of control and
386 discrimination that continue to shape today's social realities [45]. Enforced color lines disrupted
387 continuous clines of variation and often shifted to serve political goals rather than to describe
388 patterns of biodiversity. Further, race categories are usually collected to comply with civil rights
389 reporting guidelines or for social and administrative purposes, but the racial categorization systems
390 were not designed for genetic studies [2, 14]. Similarly, ethnicity categories can be centered on
391 culture, social norms, religious beliefs, or language rather than genetic ancestry, and their use in
392 genetic studies can lead to inaccurately reported results. Ethnicity categories are malleable
393 concepts that can change in different times or circumstances irrespective of hereditary lines [14,
394 18].
395
396 A sharp decreasing trend in using race categories in genomic studies has been reported [46]. While
397 genetic ancestry categories resemble, in their biological aspect, a direct reflection of a partial
398 inheritance of genetic material across generational lines, they can also suffer from limitations in
399 semantics, in space, and in time. Ancestry categories can be based on geographic, geopolitical,
400 social, or cultural elements, and the process of imposing clear lines among pre-defined populations
401 is rife with social and technical limitations. These limitations are heightened when categories are

402 quite broad (e.g. continent-based), when individuals identify with more than one category, and
403 when the labels might further open the door for enforced stigmatic associations of disease on entire
404 communities [25]. As a result, it is generally advisable to cluster genomics cohorts only when
405 justified by the research question rather than by default [10, 19], and to place social implications
406 of the research at the heart of the design process rather than as an afterthought [47].
407
408 Given the compound nature of human biodiversity (i.e. continuums and clusters), a broader
409 sampling of human genetic diversity—with the careful selection of categories during the data
410 collection phase, if and as needed or obligated—remains highly central to more clearly understand
411 the patterns of genetic variation (see Chapter 5 of [10] and [48]). In fact, it is through this type of
412 diverse data collection efforts that continuous patterns of variation have been elucidated on a wider
413 scale [17, 20]. Relatedly, alleles that increase the susceptibility to a disease can be present across
414 multiple geographies. The notion of dynamic clustering can be carried over to genome-wide
415 analyses as well. For example, in the case of methods that study one variant at a time—such as
416 QTL identification and GWAS—it is possible to limit the SNP-based clustering to some
417 neighborhood of each variant to consider how the local genetic relatedness of individuals can be
418 accounted for in enrichment analysis.
419
420 While certain genomic patterns can be identified based on ancestry-based models, these models
421 might not necessarily be the only route towards this type of identification. Equally importantly,
422 certain patterns can also be missed when relying on broad or discrete ancestral categories or when
423 considering complex traits with loci distributed across the genome. Broad stratification, whether
424 driven by discriminatory legacies embedded in genetic practice or normalized experimental design

425 and data collection, might obfuscate genomic patterns that transcend pre-defined categorical
426 boundaries. Diverse datasets and new and existing quantitative approaches that can transcend pre-
427 defined categories—by utilizing or being able to incorporate trait-specific regions, clines,
428 estimated relatedness matrices (e.g. [49, 50]), principal components (e.g. [51-54]), or other chosen
429 means—are hence crucial to approach different types of genomic studies (e.g. ones for gene
430 discovery or other types of genomic studies described in [10]), to detect various patterns of
431 genomic variation, to address theoretical and applied challenges (e.g. gene-environment
432 interactions, pleiotropy, false positive control), and to leverage data generated using different
433 technologies (e.g. deep sequencing and GWAS). These efforts have the potential to draw a more
434 complete portrait of the genomic bases of traits all the while navigating the entangled relationships
435 between science and society [11, 55-57].
436
437 **Methods**
438
439 ***Genomic Datasets***
440
441 We used TCGA germline data from the BRCA (n=1072 samples after data processing), COAD
442 (445), KIRC (514), LIHC (360), OV (554), PAAD (182), PRAD (487), and READ (164) studies.
443 We used the MC3 somatic dataset [58] for functional enrichment analysis and the ancestral labels
444 from the TCGAA Project (http://fcgportal.org/TCGAA/) [33].
445
446
447

448 ***SNP Selection***

449

450 For COSMIC-based SNP sets, we selected autosomal SNPs annotated as nonsynonymous, stop-
451 gain, and stop-loss ClinVar database annotations [59]. For HFI subsets, the functional impact of
452 missense germline variants within a cancer type's exome samples was determined using MetaSVM
453 [60] and annotations by ClinVar, when available. We considered a missense variant high-
454 functional impact if classified as Deleterious by MetaSVM or listed as Pathogenic/Likely-
455 Pathogenic in ClinVar. We used MetaSVM scores from the dbNSFP database which includes pre-
456 calculated function impact scores for 75,931,005 human non-synonymous single-nucleotide
457 variants [61]. Loss-of-function (LoF) variants, including frameshift indels, stop gain, and stop loss
458 variants, were also considered high-functional impact as well as variants annotated as high-
459 confidence LoF in gnomAD [62] or Pathogenic/Likely-Pathogenic in ClinVar. We only included
460 autosomal variants that met quality control measures described in the source paper by Huang *et al.*
461 [30].

462

463 ***Driver Gene Identification***

464

465 Potential driver gene identification was performed on the MutSigCV [63] server available at
466 https://www.genepattern.org/modules/docs/MutSigCV. Genes with FDR < 0.1 were deemed
467 statistically significant potential drivers.

468

469

470

471

472

473 ***Enrichment Analysis***

474

475 Enrichment analysis to identify pathways and biological processes was performed on g:Profiler
476 available at https://biit.cs.ut.ee/gprofiler/, with entities having g:SCS threshold < 0.05 considered
477 significant [64]. Visualization of cancer type-gene heatmap was done using ggplot2 [65].

478

479 ***Data Availability***

480

481 The results published here are in whole or part based upon data generated by the TCGA Research
482 Network: https://www.cancer.gov/tcga. Germline variants of TCGA cohorts were downloaded
483 from the Genomic Data Commons (GDC, https://gdc.cancer.gov/about-
484 data/publications/PanCanAtlas-Germline-AWG) of the National Cancer Institute (NCI) per Huang
485 *et al.* [30]. Clinical information was downloaded using RTCGA [66]. TCGA variant and meta-
486 data are available through the GDC portal at https://portal.gdc.cancer.gov/. Ancestry categories
487 are obtained from TCGAA [33] available at http://fcgportal.org/TCGAA/. We chose to not analyze
488 samples labeled "Native American [NA]" out of respect for Indigenous sovereignty. RTCGA
489 bioconductor package used to download clinical TCGA data is available at
490 https://rtcga.github.io/RTCGA/.

491

492 Genes with germline associations at the tissue-specific level were downloaded from COSMIC v90
493 [67]. Full lists are available in Supplementary Table 3.

494 ***Code Availability***

495

496 We used PLINK v1.90 [68] available at https://zzz.bwh.harvard.edu/plink/ to calculate IBS
497 matrices used to generate MDS plots. DBSCAN clusters were generated using the "dbscan"
498 package in R (https://cran.r-project.org/web/packages/dbscan/index.html) [69], and Hclust and K-
499 Means using "stats" (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html).

500

501 **References**

502

503 1. Roberts, D.E., *Fatal invention: How Science, Politics, and Big Business Re-create Race in the*
504 *Twenty-first Century*. 2011, New York: New Press.
505 2. Jablonski, N.G., *Skin color and race.* Am J Phys Anthropol, 2021. **175**(2): p. 437-447.
506 3. Marks, J., *Long shadow of Linnaeus's human taxonomy.* Nature, 2007. **447**(7140): p. 28-28.
507 4. Anemone, R.L., *Race and Human Diversity: A Biocultural Approach*. 2019, Oxford and New York:
508 Routledge.
509 5. Sax, B., *When Adam and Eve Were Monkeys: Anthropomorphism, zoomorphism, and other ways*
510 *of looking at animals*, in *The Routledge companion to animal-human history*, H. Kean and P.
511 Howell, Editors. 2018, Routledge/Taylor & Francis Group,: London ; New York.
512 6. HoSang, D.M., *On Racial Speculation and Racial Science: A Response to Shiao et al.* Sociological
513 Theory, 2014. **32**(3).
514 7. Gates, H.L. and A.S. Curran, *Who's Black and Why?: A Hidden Chapter from the Eighteenth-*
515 *century Invention of Race*. 2022, Cambridge, Massachusetts: The Belknap Press of Harvard
516 University Press.
517 8. Curran, A.S., *The Anatomy of Blackness: Science & Slavery in an Age of Enlightenment*. 2011,
518 Baltimore: Johns Hopkins University Press.
519 9. Hogarth, R.A., *Medicalizing Blackness: Making Racial Difference in the Atlantic World, 1780-*
520 *1840*. 2017, Chapel Hill: The University of North Carolina Press.
521 10. National Academies of Sciences, Engineering, Medicine, *Using Population Descriptors in*
522 *Genetics and Genomics Research: A New Framework for an Evolving Field*. 2023, Washington,
523 DC: The National Academies Press.
524 11. Reardon, J., *The Human Genome Diversity Project: A case study in Coproduction.* Social Studies
525 of Science, 2001. **31**(3): p. 357-388.
526 12. Jasanoff, S., *The Idiom of Co-Production*, in *States of Knowledge: The Co-Production of Science*
527 *and the Social Order*, S. Jasanoff, Editor. 2004, Routledge. p. 1-12.
528 13. Yudell, M., *Race Unmasked: Biology and Race in the Twentieth Century*. 2014, New York:
529 Columbia University Press.
530 14. Race, Ethnicity, and Genetics Working Group, *The use of racial, ethnic, and ancestral categories*
531 *in human genetics research.* Am J Hum Genet, 2005. **77**(4): p. 519-32.
532 15. Crawford, N.G., et al., *Loci associated with skin pigmentation identified in African populations.*
533 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. Tawk, R., et al., *Differences in Colorectal Cancer Outcomes by Race and Insurance.* Int J Environ Res Public Health, 2015. **13**(1): p. ijerph13010048.
41. Xu, D. and Y. Tian, *A Comprehensive Survey of Clustering Algorithms.* Annals of Data Science, 2015. **2**(2): p. 165-193.
42. Bailey, M.H., et al., *Comprehensive Characterization of Cancer Driver Genes and Mutations.* Cell, 2018. **173**(2): p. 371-385.e18.
43. Sanchez-Vega, F., et al., *Oncogenic Signaling Pathways in The Cancer Genome Atlas.* Cell, 2018. **173**(2): p. 321-337.e10.
44. Hanahan, D. and Robert A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.
45. McLean, S.-A., *Isolation by Distance and the Problem of the Twenty-First Century.* Human Biology, 2019. **91**(2): p. 81-94.
46. 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.
47. 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.
48. Oni-Orisan, A., et al., *Embracing Genetic Diversity to Improve Black Health.* New England Journal of Medicine, 2021. **384**(12): p. 1163-1167.
49. 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.
50. Zhou, X. and M. Stephens, *Genome-wide efficient mixed-model analysis for association studies.* Nat Genet, 2012. **44**(7): p. 821-4.
51. 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.
52. Mbatchou, J., et al., *Computationally efficient whole-genome regression for quantitative and binary traits.* Nat Genet, 2021. **53**(7): p. 1097-1103.
53. 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).
54. 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.
55. Yudell, M., et al., *Taking race out of human genetics.* Science, 2016. **351**(6273): p. 564-565.
56. Thorp, H.H., *Time to look in the mirror*, in *Science*. 2020.
57. Nobles, M.W., Chad; Wonkam, Ambroise; Wathuti, Elizabeth, *Science must overcome its racist legacy: Nature's guest editors speak*, in *Nature*. 2022.
58. 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.
59. Landrum, M.J., et al., *ClinVar: public archive of interpretations of clinically relevant variants.* Nucleic Acids Res, 2016. **44**(D1): p. D862-8.
60. Kim, S., et al., *Meta-analytic support vector machine for integrating multiple omics data.* BioData Min, 2017. **10**: p. 2.
61. 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.
62. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.
63. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-218.
64. 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.

65. Wickman, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016, New York, NY: Springer-Verlag.
66. Kosinski, M.B., Przemyslaw; Chodor, Witold. *The family of R packages containing TCGA data*. 2016; Available from: https://rtcga.github.io/RTCGA/index.html.
67. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer.* Nucleic Acids Res, 2019. **47**(D1): p. D941-D947.
68. 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.
69. Hahsler, M.P., Matthew; Arya, Sunil; Mount, David, *dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms.* 2022.

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)