

1 **HAVOC: Small-scale histomic mapping of biodiversity across entire tumor specimens using deep**
2 **neural networks**

3
4 **Anglin Dent^{1*}, Kevin Faust^{2-3*}, K. H. Brian Lam^{3,4}, Narges Alhangari¹, Alberto J. Leon³, Queenie**
5 **Tsang³, Zaid Saeed Kamil^{1,5}, Andrew Gao^{1,5}, Prodipto Pal^{1,5}, Stephanie Lheureux³, Amit Oza³,**
6 **Phedias Diamandis^{1,3,5,6}**

7
8 ¹Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON M5S 1A8,
9 Canada.

10 ²Department of Computer Science, University of Toronto, 40 St. George Street, Toronto, ON M5S 2E4,
11 Canada

12 ³Princess Margaret Cancer Centre, 101 College Street, Toronto, ON M5G 1L7, Canada.

13 ⁴Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of
14 California Los Angeles, Los Angeles, United States of America

15 ⁵Laboratory Medicine Program, Department of Pathology, University Health Network, 200 Elizabeth
16 Street, Toronto, ON M5G 2C4, Canada.

17 ⁶Department of Medical Biophysics, University of Toronto, 101 College St, Toronto, ON M5G 1L7,
18 Canada

19

20 ***equal contributions**

21

22 Please direct correspondence to:

23 **Phedias Diamandis, MD, PhD, FRCPC**

24 Neuropathologist, Department of Pathology, University Health Network

25 12-308, Toronto Medical Discovery Tower (TMDT),

26 101 College St, Toronto, M5G 1L7

27 p.diamandis@mail.utoronto.ca

28 416 340 4459

29

30

31

32

33 **Key words:** tumor heterogeneity, deep learning, computer vision, artificial intelligence, molecular
34 profiling

35 **Summary:**

36 **Intra-tumoral heterogeneity can wreak havoc on current precision medicine strategies due to**
37 **challenges in sufficient sampling of geographically separated areas of biodiversity distributed across**
38 **centimeter-scale tumor distances. In particular, modern tissue profiling approaches are still largely**
39 **designed to only interrogate small tumor fragments; which may constitute a minute and non-**
40 **representative fraction of the overall neoplasm. To address this gap, we developed a pipeline that**
41 **leverages deep learning to define topographic histomorphologic fingerprints of tissue and create**
42 **Histomic Atlases of Variation Of Cancers (HAVOC). Importantly, using a number of spatially-**
43 **resolved readouts, including mass-spectrometry-based proteomics and immunohistochemisy, we**
44 **demonstrate that these personalized atlases of histomic variation can define regional cancer**
45 **boundaries with distinct biological programs. Using larger tumor specimens, we show that HAVOC**
46 **can map spatial organization of cancer biodiversity spanning tissue coordinates separated by**
47 **multiple centimeters. By applying this tool to guide profiling of 19 distinct geographic partitions**
48 **from 6 high-grade gliomas, HAVOC revealed that distinct states of differentiation can often co-exist**
49 **and be regionally distributed across individual tumors. Finally, to highlight generalizability, we**
50 **further benchmark HAVOC on additional tumor types and experimental models of heterogeneity.**
51 **Together, we establish HAVOC as a versatile and accessible tool to generate small-scale maps of**
52 **tissue heterogeneity and guide regional deployment of molecular resources to relevant and**
53 **biodiverse tumor niches.**

54

55

56

57

58 Tumor heterogeneity underpins modern frameworks of tumor evolution and treatment resistance,
59 but resolving cancer biodiversity distributed across long distances in resection specimens has proven
60 challenging¹⁻³. Even small biopsies (~1 cm³) can contain ~10⁸ tumor cells; a number representing many
61 orders of magnitude more than what current single-cell profiling approaches can routinely process (~10³
62 cells)⁴⁻⁶. Similarly, asymmetric distributions of biological variation across specimens can result in non-
63 representative under-sampling and/or mixing of critical tumor subpopulations using bulk-based
64 expression profiling approaches^{7,8}. These complexities may be further amplified in large cohort studies
65 and analysis of recurrent tumor specimens where variations in cellular composition across samples can
66 complicate interpretations⁹⁻¹¹.

67 In geography, large-scale maps that provide high-resolution information of relatively small
68 topographic areas (e.g. cities, provinces) are often complemented with smaller-scale cartographic surveys
69 that document specific sets of relevant features over larger regions (e.g. countries and continents)^{12,13}. In
70 such disciplines, these small-scale atlases aid in guiding systems-level decisions and resource allocation
71 (e.g. nature conservation efforts). In the context of cancer, systematic approaches to establishing
72 coordinates of potential biodiversity across large tissue specimens could guide deployment of limited
73 molecular profiling resources to better capture tumor-level heterogeneity¹⁴ (**Fig 1a**).

74 To address this critical challenge, we developed HAVOC, a histology-based deep neural network
75 pipeline aimed at creating “Histomic Atlases of Variation Of Cancers” and providing spatially
76 contextualized estimates of biodiversity across virtually any scale (**Fig 1b**). This approach leverages both
77 classic and modern dogmas of histopathology; (i) that significant changes in molecular programs are often
78 accompanied with modifications in cellular morphologies, and (ii) that such cytoarchitectural variations
79 can be objectively defined by contemporary computer vision strategies¹⁵⁻¹⁸. Previous studies exploring
80 intra-tumoral heterogeneity using computer vision have largely used supervised approaches, in which

81 neural networks are trained using hematoxylin and eosin (H&E)-stained whole slide images (WSI) with
82 genetically-defined ground truth labels generated from bulk tumor tissue (e.g. TCGA)^{19–21}. Once trained
83 these models are then applied to predict the mutational, cell composition and transcriptional status of
84 individual image patches from larger tissue areas. While such approaches have shown capacity to predict
85 presence of immune infiltrates and a handful of specific mutations in a context-specific manner, more
86 intricate tumor cell-intrinsic biological programs (e.g. proliferation, hypoxia, DNA repair) have proven
87 more challenging to generalize across the majority of tumor types^{22–24}. Similarly, such approaches usually
88 require knowledge and design around specific mutations defined *a priori* (e.g. FGFR3 mutations in
89 bladder cancer) and may therefore be less suitable for screening for potential intra-tumoral heterogeneous
90 subclones that may be patient-specific and emerge downstream of these common initiating genetic
91 events^{25,26}. To address this, one recent study used deep learning to detect immune and stroma infiltrates
92 and leveraged these features as geographic guideposts to define immune “cold” and “hot” regions of non-
93 small cell lung carcinomas. Using whole-exome and RNA-sequencing on these deep-learning defined
94 regions, they showed that cancer subclones derived from immune cold regions shared closer proximity in
95 mutational space than subclones from immune hot regions¹⁴. Such molecularly-agnostic and prospective
96 mapping approaches of geospatial variability, are critical as they may provide more personalized and
97 precise insights into the emergence of treatment-resistant subclones and aggressive clinical
98 phenotypes^{27,28}.

99 Here, we show that by using unsupervised partitioning of patterns found on H&E-stained WSIs,
100 HAVOC can predict the precise coordinates and provide relative estimates of relevant morphologic and
101 molecular patterns of cancer diversity. Importantly, this approach does not require any *a priori* framework
102 of biodiversity, allowing for tumor heterogeneity to be explored free of any pre-defined constraints (e.g.
103 regional lymphocytic infiltrates) or narrow molecular definitions (e.g. specific mutations, aneuploidy) that

104 may not robustly generalize in molecularly heterogeneous cancers. Specifically, using high grade gliomas
105 as a proof-of-concept, an aggressive form of brain cancer that can show significant intra-tumoral variation,
106 HAVOC's predictions of biovariation showed strong concordance with both human- and molecularly-
107 defined estimates of heterogeneity. This regional analysis also specifically revealed that tumor
108 subpopulations with varying degrees of phenotypic differentiation and proliferative capacity can often co-
109 exist and be geographically separated in high grade gliomas; underscoring the need for more intelligent
110 sampling strategies. This unique spatially-defined glioma proteomic dataset generated in this study also
111 showed additional patient-specific regional differences that can be further explored in real-time in the
112 Brain Protein Atlas²⁹ (<https://www.brainproteinatlas.org/dash/apps/ad>). Importantly, we further highlight
113 the generalizability of HAVOC to other tumor types and cancer models and show it can be deployed
114 without the need for any additional context-specific training. In addition to the available code for local
115 implementation, we developed a server version of HAVOC that can be accessed via a web interface
116 (<https://www.codido.co>) without the need for any specialized hardware or software expertise. Routine
117 generation of such small-scale histomic atlases of biologic variations in cancer offers an opportunity to
118 better identify and document critically divergent cellular habitats across large geographic regions and may
119 aid in better managing tumor heterogeneity for both large cohort studies and personalized medicine efforts.

120

121 **Prediction and relative quantification of histomorphological heterogeneity by HAVOC**

122 To assess if delineation and quantification of spatial transitions in histomorphology could be
123 automated, we applied an unsupervised image clustering framework³⁰ to a digitized WSI cohort of brain
124 tumors comprising largely of diffuse gliomas (n=40) showing varying degrees of regional heterogeneity.
125 The advantage of this unsupervised deep learning strategy is that it helps overcome the significant case-
126 to-case morphologic variability that is seen in many malignant cancers such as high grade gliomas³¹.

127 Briefly, WSIs are individually partitioned into 0.066-0.27 mm² non-overlapping image patches (based on
128 specimen size and pattern of interest) and passed through a histology-optimized convolutional neural
129 network (CNN) previously trained on a diverse set of nearly 1 million pathologist-annotated image patches
130 extracted from over 1,000 brain tumors^{15,32}. Rather than carrying out classification, the 512-dimensional
131 deep learning feature vector (DLFV), generated in the penultimate layer of the CNN, is extracted and used
132 as a “histomic” feature set. These DLFV signatures are then utilized to carry out patch-level clustering
133 (k=2-9) and spatially map morphologic variation across entire WSIs (**Fig 1c-e, Supplemental Fig 1**). In
134 addition to lesion segmentation from normal brain tissue elements in early partitions, when present, this
135 approach also non-randomly segregated tumor areas with progressively more subtle regional morphologic
136 pattern differences in later subdivisions (e.g. variations in tumor cellularity and intra-tumoral edema,
137 $p=1.2e^{-11}$, Kolmogorov–Smirnov test) (**Fig 1f, Supplemental Fig 2**). Generation of feature activation
138 maps (FAMs) of salient individual deep learning features (DLFs), enriched in each compartment, provides
139 support and insight of human-perceivable histomorphologic patterns associated with each HAVOC
140 partition (**Fig 1g**). Importantly, Pearson correlation coefficients (r) of regional DLFVs correlated with
141 human expert estimates of histologic variation; allowing this metric to serve as a quantitative
142 approximation of the degree of variability across HAVOC-defined regions (**Fig 1h-i**, n=40 cases;
143 $p<0.0001$, Mann-Whitney U test). While we found that the optimal number of subdivisions varied
144 depending on the degree of tissue complexity found on individual slide, for most cases, the majority of
145 discernible patterns of histomorphologic variation ($r = \sim 0.74$) plateaued and reached saturation after 7-8
146 partitions; irrespective of the overall level of heterogeneity found on the WSI (**Supplemental Fig 3**).
147 Together, this data highlights how HAVOC can provide an automated, objective, and human expert-
148 concordant tool to estimate spatial histomorphologic variation across cancer tissue.

149

150 **Spatial morphologic fingerprints align with molecular patterns of heterogeneity**

151 Interestingly, in one of the glioblastoma cases of our initial cohort, we noted that HAVOC partitioning
152 captured a BRAFV600E-mutated tumor subclone showing an elevated Ki-67 (MIB1) proliferation index
153 on immunohistochemistry (**Fig 2a**). This supported the possibility that regional changes in morphologic
154 fingerprints may also predict relevant phenotype-level variability in tumor biology. To begin formally
155 testing if histomic heterogeneity, perceived by HAVOC, also correlated with global molecular differences,
156 we compared the r of DLFVs across the major histomorphologic hallmarks of diffuse gliomas including
157 regions of high cellularity (CT), infiltrating tumor (IT) and brain tissue at the leading tumor edge (LE) to
158 proteomic and transcriptional profiles generated from these areas in previous studies^{33,34} (**Supplemental**
159 **Fig 4**). Indeed, HAVOC-defined niche variations correlate with proteomic ($r^2=0.79$) and transcriptomic
160 ($r^2=0.89$) profiles with similar glioma histomorphologic regions (eg. multiple IT regions) having a higher
161 degree of DLFV and molecular similarity when compared to more distinct niches (eg. CT vs LE regions).

162 To more directly assess the predictive power of HAVOC for resolving distinct spatial expression
163 profiles, we next determined if regional historically-defined niches, exclusively within the cellular
164 glioma compartment, could predict phenotypic protein-level heterogeneity. We focused our regional
165 analysis on proteomic outputs due to known proteogenomic discordances in many cancers in which
166 genetic changes may not always translate to downstream cellular phenotypes³⁵⁻³⁷. Laser capture
167 microdissection (LCM) followed by liquid chromatography tandem mass spectrometry (LC-MS/MS)
168 analysis of cellular tumor regions highlighted significant regional variability and numerous differentially
169 encoded proteins (FDR = 0.1, n=3 separate microdissection replicates) (**Fig 2b-c, Supplemental Fig 5-6,**
170 **Supplementary Table 1-2**). Pathway analysis, using a previous defined set of 64 proteogenomically
171 concordance signatures³³, highlighted regional heterogeneity in proliferative (MYC-, $p=0.0006$), invasive
172 (KRAS-, $p=0.03$) and hypoxic ($p=0.0048$) programs³³. Importantly, the enrichment of these programs was

173 dampened when the entire tissue section was profiled; underscoring the value of regional profiling of
174 highly heterogeneous tumors using expression-based techniques (MYC- Red vs Bulk, $p=0.001$; Hypoxia-
175 Red vs Bulk, $p=0.008$) (**Fig 2d-e**). Interestingly, the two major HAVOC partitions from Patient IV of the
176 main study cohort showed a narrower profile of differences in functional programs (e.g. hypoxia),
177 highlighting the capacity of HAVOC to detect even subtle molecular differences across individual cases
178 driven by the microenvironment ($p=0.04$, t-test) (**Supplementary Fig 6b**).

179 Given the role of MYC in cell cycle progression, we validated these differences in the case
180 presented in Figure 2b using CNN-based estimates of Ki-67 staining. This confirmed dramatic spatial
181 variation in the proliferative capacity of these adjacent regions (**Fig 2f**, $p < 0.0001$). We further explored
182 the predictive potential of HAVOC for heterogeneously proliferating subclones in another independent
183 cohort of glioblastomas assembled to contain cases that displayed regional variations in Ki-67 ($n=5$,
184 **Supplementary Table 2**). In all these cases, HAVOC's heterogeneity maps captured H&E-based DLFV
185 signatures that aligned and correlated with differential rates of proliferation (t-test, $p<0.0001$,
186 **Supplemental Fig 7**).

187 To explore potential correlations between proliferation programs and morphology, we next
188 assessed if specific DLFs were being activated by discernible histomorphologic features within regions
189 displaying distinct Ki-67 indices in a representative case (**Supplemental Fig 8**). Indeed, FAMs of
190 regionally-enriched DLFs highlighted a transition in pattern with areas with high cellularity (DLF214;
191 activation on tumor nuclei) to regions displaying fibrillar cytoarchitecture (DLF134; activation on
192 cytoplasmic processes) in variably high to low Ki-67 positive regions, respectively (**Fig 2g**). Interestingly,
193 in this validation cohort, HAVOC-partitions (e.g. Orange, O) often formed “subclusters” (e.g. O1, O2,
194 O3), that while geographically separated, showed similar morphologic patterns (e.g. hypercellularity,
195 infiltrative phenotype) and maintained stable cluster-specific Ki-67 proliferation indices (**Fig 2h**). These

196 findings suggested that H&E-based DLFV signatures may generalize to multiple spatially-separated tumor
197 sub-regions and could help construct maps of biovariation even across larger specimens spanning multiple
198 histopathological slides.

199

200 **Mapping biovariation across centimeter scale distances and entire specimens with HAVOC**

201 Given the ability of histomic signatures to group geographically separated tissue regions with
202 similar biology on individual slides (Fig 2h), we next assessed the generalizability of this concept to the
203 organization of tumor heterogeneity within individual large tissue specimens spanning multiple slides. As
204 an example, we highlight a HAVOC map generated on a large recurrent isocitrate dehydrogenase (IDH)-
205 mutated 1p19q-codeleted anaplastic oligodendroglioma, CNS WHO grade 3, showing heterogeneous
206 radiographic signal and measuring 4.4 cm in maximum dimensions (**Fig 3a**). To map the histomorphologic
207 heterogeneity across this entire case, we first generated seven separate HAVOC partitions for each of the
208 12 H&E sections spanning the entire resected specimen (5.4 x 4.1 x 1.8 cm) (**Fig 3b**). All 84 HAVOC-
209 defined tumor regions were then histomorphologically organized across the entire specimen based on their
210 pairwise DLFV similarities (r) (**Fig 3c**). Even across multiple sections, regions clustering more closely
211 aligned with expert annotations rather than positional coordinates (**Fig 3d, Supplemental Fig 9**). This
212 arrangement was further objectively validated with substantially different estimated proliferation indices
213 (Ki-67) across the clusters of the geographically distributed partitions ($p < 0.005$; Mann-Whitney U Test,
214 **Fig 3d-e**). t-SNE mapping of all 10,973 0.27 mm² image patches³⁸ from this lesion also supported this
215 HAVOC arrangement with a gradient of morphological and biological patterns, transitioning from nodular
216 (blue), hypercellular (orange), moderately cellular (purple and green) and non-tumor/low cellularity (red)
217 tumor regions (**Supplemental Fig 10**).

218 To further examine if HAVOC could generalize to identify regional molecular profiles across
219 larger tumor distances, we carried out this multi-slide mapping workflow on paired slide from three
220 additional IDH-wildtype high grade gliomas (Histology: 2 WHO grade 4, 1 WHO grade 3) (**Fig 4a, d,**
221 **Supplemental Fig 11, Supplementary Table 1**). Overall regional DLFV correlation mappings showed
222 similar geographic relationships to global proteomic signatures (**Fig 4b,e**). In one case, HAVOC
223 appropriately defined a focal hyperdense outlier region from other tumor morphologies (**Fig 4a-c**). In one
224 of the other paired sets of samples, HAVOC partitions were reciprocally aligned with regions on different
225 slides (**Fig 4d-e**). The final case also highlighted distinct diffuse infiltrating biology exclusive to only one
226 of the two slides (Patient Va_RED), which was again appropriately segmented by HAVOC
227 (**Supplemental Fig 11**). Interestingly, while quantification of regional cellularity provided further support
228 for the proposed HAVOC groupings of the first 2 cases (**Figure 4**), the third case pairing had a more
229 uniform distribution of cellularity further supporting that HAVOC uses a diversity of cytoarchitectural
230 features to guide clustering (**Supplementary Fig 12**). All together, these experiments highlighted the
231 potential for HAVOC to appropriately map and align intra-tumor areas of heterogeneity across large tissue
232 distances.

233

234

235 **HAVOC reveals distinct geospatially separated differentiation states in high grade gliomas**

236 Molecular descriptions of tumor biology are often defined based on sampling and profiling of a
237 single tumor region³⁹. We therefore next wanted to examine if HAVOC was detecting critical cellular
238 states that may show reoccurring patterns of regional variation within individual tumors. We therefore
239 aggregated the profiles of the 19 regions spanning six IDH-wildtype high grade gliomas (5 WHO Grade
240 4, 1 WHO grade 3; concurrently profiled in this study) and carried out a group analysis (**Supplementary**

241 **Table 1**). Using the previously defined set of 64 proteogenomically concordant molecular programs,
242 unsupervised clustering of the 19 regions was partly driven by patient ID; speaking to the high inter-
243 tumoral heterogeneity found in IDH-wildtype high grade gliomas (**Supplementary Fig 13**). Importantly,
244 this analysis also revealed spatial intra-tumoral variations in molecular programs in almost all profiled
245 cases; including regional patterns of immune response, hypoxic response, proliferation and embryonic
246 differentiation states. UMAP dimensionality reduction revealed a significant influence of the later
247 differentiation signatures across the entire cohort (e.g. genes associated with an embryonic “poorly”
248 differentiated cell state⁴⁰) (**Fig 5a**). This was further supported with a strong inverse correlation with
249 regional programs associated with a mature astrocytic phenotype ($r=0.71$, $p < 2e^{-16}$) (**Fig 5b**). Supervised
250 regional analysis of this astrocytic-embryonic differentiated axis found that at least 5 out of the 6 profiled
251 cases showed significant spatial differences ($p < 0.005$, ANOVA) (**Fig 5c**). We found this high co-
252 occurrence of these contrasting patterns notable, as tumors, across multiple organ sites, are often still
253 approximated as being either “poorly” or “well” differentiated upon clinical presentation.

254 In the Cancer Genome Atlas (TCGA), this differentiation axis correlated with aggressiveness with
255 both lower WHO grades ($p < 7.4e^{-5}$, $n=446$) and IDH mutations ($p = 2.8e^{-11}$) of astrocytic (non 1p19q-
256 codeleted) tumors showing significantly higher differentiation signal at the bulk level (**Fig 5d,e**). We also
257 found support for both inter-and intra-tumoral variation along the astrocytic-to-embryonic axis in single
258 cell RNA data generated from both patient-derived glioblastoma tissue specimens and brain tumor stem
259 cell cultures (BTSCs)⁴¹ (**Fig 5f, Supplemental Fig 14**). Interestingly, while some samples analyzed
260 showed complex expression patterns with multiple co-existing states of differentiation (GBM_G620;
261 GBM_G983-C), others displayed fairly homogenous signatures (GBM_G1003-A); presumably related to
262 both the sampled region and the overall tumor biology. The notable spectrum of steady-state positions in
263 differentiation states in expanded BTSCs suggests that generation of patient models from single individual

264 tumor regions may only partially capture potential region-to-region variations in differentiation uncovered
265 in this analysis. Taken together, these results highlight how neural network-guided multi-regional
266 sampling and profiling of cancer tissue can reveal new insights of biovariation that may be hidden and not
267 immediately accessible from alternative single region profiling strategies.

268

269

270 **HAVOC biodiversity maps generalize to genomic differences and untrained tissue types**

271 As tumor heterogeneity is a challenge relevant to many cancer types⁴², we next assessed the
272 generalizability of HAVOC to other non-central nervous neoplasms. First, we retrieved an available WSI
273 of an experimental metastatic lung cancer mouse model in which two independent clones of a lung tumor
274 were injected into the tail vein and subsequently resulted in multiple spatially distinct liver metastases⁶
275 (**Fig 6a**). Importantly, using spatial DNA and RNA sequencing, two of the five metastatic deposits, in
276 addition the intervening liver tissue, were characterized in the originating study using copy number and
277 expression patterns differences (**Fig 6b, Supplemental Fig 15a**). Using HAVOC, we generated 11 total
278 partitions (to ensure saturation of the different histomorphological patterns) and compared them with the
279 characterized ground truth annotations. Indeed, image-based clustering segmented all five tumors into
280 fairly homogeneous lesions, in addition to defining various regional histological patterns of the liver tissue
281 (**Fig 6c-d, Supplemental Fig 15b-c**). Using the silhouette method, these five tumors formed two major
282 clusters ($k=2$ subclones) as the most parsimonious solution (**Supplemental Fig 15d**). Notably HAVOC
283 separated out both the two genomically-distinct subclones and peritumoral liver tissue with and without
284 inflammation, further supporting that distinct histomorphologic fingerprints can be leveraged to predict
285 regions of potential genomic, transcriptomic and cell composition heterogeneity (**Fig 6d**). Subsequent
286 FAMs exploring morphologic correlates highlighted distinct histomorphologic patterns within the tumor

287 subclones defined by the original study. In Clone A, FAMs highlighted a fairly advanced organization of
288 tumor cells with abundant nuclear palisades, while Clone B showed a more random arrangement with
289 large, atypical cells scattered throughout the lesion. Further, FAMs of the peritumoral liver tissue
290 highlighted a differential distribution of inflammatory cells across HAVOC-proposed groupings, in
291 agreement with the tumor, normal, and immune cell classes assigned from the original single-cell slide-
292 RNA-seq projections (**Fig 6e**).

293 While the other three remaining tumor regions were not genomically annotated in the original
294 study (potentially due to cell throughput limitations of the spatial DNA sequencing technique), they did
295 indeed show similar histologic patterns to the more atypical tumor, confirming the ability of HAVOC to
296 group metastases of similar histomorphologies, even across entire organs.

297 To further test the generalizability of the HAVOC workflow to other untrained tissue types, we
298 next applied it to “interesting” dermatopathology and pulmonary pathology cases showing a squamous
299 neuroendocrine “collision” tumor and divergent adenosquamous tumor differentiation respectively (**Fig**
300 **7**). On the skin biopsy, the position of the squamous cell carcinoma (*in situ*) is highlighted by the high
301 molecular weight keratin (CK34BE12). Within the dermis, a distinct infiltrative neuroendocrine neoplasm
302 composed of sheets, nests and cords of round basaloid cells that labeled with synaptophysin. Indeed,
303 HAVOC-proposed groupings of this specimen distinguished the distinct squamous and neuroendocrine
304 morphologies in alignment with the distinct immunohistochemical staining (**Fig 7a-b**). HAVOC
305 partitioning of the case of adenosquamous carcinoma also showed a high spatial concordance to TTF1
306 (adenocarcinoma) and p40 (squamous carcinoma) immunopositive tumoral components (**Fig 7c-d**,
307 **Supplementary Fig 16**). Altogether, these data demonstrate that HAVOC can serve as a tissue type- and
308 molecular-agnostic tool to map biodiversity in different cancers.

309

310 **Discussion**

311 Intra-tumoral heterogeneity has emerged as a key concept in current precision medicine efforts⁴². While
312 there have been technological breakthroughs to map such spatial biodiversity at the genomic,
313 transcriptomic and proteomic level, the relative discord between the upper limits of tissue profiling
314 throughput and tumor sizes creates an important bottleneck for comprehensive analysis of most tumor
315 specimens^{2,4,7}. Here, we highlight how a neural network-based tool (HAVOC) can detect and quantify
316 spatial distributions of cancer biodiversity across not only individual sections, but also on larger (entire)
317 tumor specimens in a hypothesis-agnostic manner. Importantly, we go to great lengths to prospectively
318 validate that these proposed histomorphologic patterns of spatial heterogeneity correlate with expert
319 human interpretations, key biomarkers of aggressiveness and global genetic, transcriptomic and protein-
320 level differences. Using this tool, we partitioned and profiled 19 distinct regions from 6 IDH-wildtype
321 high grade gliomas specimens and uncovered that different spatially confined states of differentiation can
322 co-exist within the same tumor. Interestingly, while the defined poorly differentiated state correlated with
323 many important indicators of clinical aggressiveness (WHO grade and IDH status), it does not appear to
324 correlate with proliferation or tumor initiating potential in previous studies, adding another layer of
325 complexity to existing models of tumor heterogeneity. Importantly, such findings also underscore the
326 potential of histomic mapping to guide and improve representative sampling and characterization of
327 heterogeneity across entire tumor resection specimens. While we focused our discussion and validation
328 efforts on spatial patterns of differentiation, there were many additional case-by-case geographic
329 differences identified that were outside the scope of this study (e.g. proliferation, hypoxia, immune
330 response). We therefore make this unique morphology-driven spatial analysis of the glioma proteome
331 available for further exploration through an inter-active data portal (Brain Protein Atlas²⁹;
332 <https://www.brainproteinatlas.org/dash/apps/ad>).

333 There are important distinctions of how HAVOC complements other computational approaches to
334 mapping intra-tumoral molecular heterogeneity directly from histology. Approaches that aim at directly
335 training and/or correlating deep learning outputs to specific molecular events are inherently empirical and
336 may be most effective at predicting fairly robust molecular signatures that are driven by heterogenous
337 cellular compositions (e.g. presence of lymphocytes, vessels) and/or a handful of transcriptional profiles
338 (e.g. cell division)^{22,23}. While these may be highly effective and cost-efficient in some clinical and research
339 contexts, such approaches may not have capacity to extend and generalize across all cancer types and
340 relevant molecular pathways of interest. This presents an important limitation and gap for targetable
341 biological programs that may not have strong “histomolecular” correlates, cancers with a high degree of
342 heterogeneity and for more modern personalized medicine strategies^{27,28}. A more hypothesis-free
343 approach has been the use of supervised deep learning approaches to identify specific patterns within
344 tumors (e.g. areas devoid of tumor infiltrating lymphocytes and stromal elements) that may signal
345 differential biology¹⁴. These “at risk” regions provide spatial targets for prospective and personalized
346 profiling of individual cancer specimens. We believe HAVOC extends this concept further by decoupling
347 the need for any pre-defined feature that may potentially bias or limit exploration into mapping of
348 biovariation with associated phenotypic changes. While, in certain circumstances, this may compromise
349 the sensitivity for detecting subtle spatial variations in important molecular programs, we believe it
350 provides a highly dynamic and generalizable solution for personalized discovery and characterization of
351 spatial changes in tumor biology. It therefore can be directly extended to not only a variety of tumor types,
352 but even to the longitudinal analysis of cancer in which molecular patterns may be influence by both tumor
353 progression and treatment-related effects⁹.

354 HAVOC does not require any dedicated tissue or significant computational resources, and contains
355 only a handful of tunable parameters (e.g. tile size, cluster numbers) that can be easily adjusted depending

356 on the specific context (e.g. smaller tumor deposits). These capabilities, when combined with the
357 processing of serial sections for molecular analysis, provide a computational approach for the
358 characterization of intra-tumoural heterogeneity at practically any scale and level of resource availability.
359 Moreover, as we demonstrated in this study, the large diversity of images used to train the default CNN
360 within HAVOC provides an adaptable unsupervised framework to detect tumor heterogeneity in a
361 relatively species-, tissue-, molecular platform-agnostic manner. We, therefore, envision HAVOC to serve
362 as a routine and powerful research tool for mapping intra-tumoral heterogeneity across both large-scale
363 and more intricate (n-of-1) tissue profiling studies. Given its ease of implementation and compatibility
364 with FFPE sections, we envision HAVOC becoming an essential tool to map heterogeneity on all clinical
365 and research tissue-based to help provide more systematic approaches to tissue selection for ongoing
366 personalized precision medicine efforts.

367 Because HAVOC is built to detect histomorphologic pattern differences, it can be theoretically
368 extended to non-neoplastic and immunohistochemistry-resolved tissue heterogeneity that may be niche-
369 and epigenetically patterned and independent of genetic (e.g. copy number) alternations. Moreover, when
370 coupled with salient feature activation mapping, it can complement spatial transcriptomic and proteomic
371 approaches by providing phenotypic correlates (e.g. edema, increased nuclear: cytoplasmic ratio) to
372 explain/predict interesting expression-level differences. To facilitate wide adoption, we have packaged
373 HAVOC in a number of ways to promote ease of use. For large-scale initiatives, we provide source code
374 to allow HAVOC to be deployed locally on large cohorts in an automated manner. For more translational
375 researcher and clinicians, we also host HAVOC in a cloud-based server (<https://www.codido.co>) that
376 allows analysis of digitized slides (.SVS) without any need for advanced software expertise or hardware.

377 There are some important caveats of HAVOC that should be considered when using this tool.
378 Firstly, its fundamental dependence on tissue patterns means that it can be influenced by non-biologically

379 factors (e.g. tissue folds/tears/focus/suboptimal uneven staining). In our experience, we found that these
380 artifacts can be easily identified and excluded from downstream analysis; both by careful post-hoc analysis
381 or by supervised classification of clusters to label tumor-specific partitions. Moreover, the sensitivity of
382 HAVOC for very subtle and intermixed patterns of biodiversity likely also needs to be assessed in a
383 context-specific manner. These biological factors may dictate the particular type of molecular profiling
384 technique best suited for HAVOC pairing. In summary, HAVOC provides a highly flexible,
385 generalizable, accessible, and scalable approach for mapping histomorphologic-correlated phenotypes and
386 could serve as an essential tool to explore and document biologic heterogeneity in human tissue and
387 disease.

388

389 **Author contributions**

390 A.D., K.F., S.L., A.O and P.D. conceived the idea and approach. K.F. developed the computational
391 workflow. A.D. and B.L. designed and carried out molecular validation experiments. N.A., P.P. A.G.,
392 Z.S.K., P.D. provided relevant cases and pathological annotations for the various validation studies. A.L.
393 and Q.T. performed bioinformatic analysis. A.D. K.F. and P.D. wrote the manuscript, with input from all
394 other authors.

395

396 **Acknowledgments**

397 The Diamandis Lab is supported by the Terry Fox New Investigator Award program, the Canadian
398 Institute of Health Research and the Brain Tumor Foundation of Canada. A.O. S.L., P.D, and P.P also
399 received research grant support from the Princess Margaret Cancer Foundation and the Ontario Institute
400 for Cancer Research.

401

402 **Competing Interests**

403 The authors declare no competing interests.

404

405 **References**

406

407 1. Aldape, K. *et al.* Challenges to curing primary brain tumours. *Nat. Rev. Clin. Oncol.* **16**, 509–520

408 (2019).

409 2. Dent, A. & Diamandis, P. Integrating computational pathology and proteomics to address tumor

410 heterogeneity. *J. Pathol.* (2022) doi:10.1002/PATH.5905.

411 3. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys.*

412 *Acta* **1805**, 105 (2010).

413 4. Del Monte, U. Does the cell number 10⁹ still really fit one gram of tumor tissue? *Cell Cycle* **8**,

414 505–506 (2009).

415 5. Couturier, C. P. *et al.* Single-cell RNA-seq reveals that glioblastoma recapitulates a normal

416 neurodevelopmental hierarchy. *Nat. Commun.* **11**, (2020).

417 6. Zhao, T. *et al.* Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nat.*

418 *2021 6017891* **601**, 85–91 (2021).

419 7. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary

420 dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 4009–14 (2013).

421 8. Lam, K. H. B., Valkanas, K., Djuric, U. & Diamandis, P. Unifying models of glioblastoma's

422 intra-tumoral heterogeneity. *Neuro-Oncology Adv.* (2020) doi:10.1093/noajnl/vdaa096.

423 9. Varn, F. S. *et al.* Glioma progression is shaped by genetic evolution and microenvironment

424 interactions. *Cell* **185**, 2184–2199.e16 (2022).

425 10. Tatari, N. *et al.* The proteomic landscape of glioblastoma recurrence reveals novel and targetable

- 426 immunoregulatory drivers. *Acta Neuropathol.* (2022) doi:10.1007/S00401-022-02506-4.
- 427 11. Lam, K. H. B. & Diamandis, P. Niche deconvolution of the glioblastoma proteome reveals a
428 distinct infiltrative phenotype within the proneural transcriptomic subgroup. *Sci. data* **9**, 596
429 (2022).
- 430 12. Jenkins, C. N., Pimm, S. L. & Joppa, L. N. Global patterns of terrestrial vertebrate diversity and
431 conservation. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E2603–E2610 (2013).
- 432 13. Durán, S. M. *et al.* Informing trait-based ecology by assessing remotely sensed functional
433 diversity across a broad tropical temperature gradient. *Sci. Adv.* **5**, (2019).
- 434 14. AbdulJabbar, K. *et al.* Geospatial immune variability illuminates differential evolution of lung
435 adenocarcinoma. *Nat. Med.* **26**, 1054–1062 (2020).
- 436 15. Faust, K. *et al.* Unsupervised Resolution of Histomorphologic Heterogeneity in Renal Cell
437 Carcinoma Using a Brain Tumor–Educated Neural Network. *JCO Clin. Cancer Informatics* **4**,
438 811–821 (2020).
- 439 16. Coudray, N. *et al.* Classification and mutation prediction from non-small cell lung cancer
440 histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
- 441 17. Bilal, M. *et al.* Development and validation of a weakly supervised deep learning framework to
442 predict the status of molecular pathways and key mutations in colorectal cancer from routine
443 histology images: a retrospective study. *Lancet Digit. Heal.* **3**, e763–e772 (2021).
- 444 18. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in
445 gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
- 446 19. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in
447 histopathology: enhancing cancer research and clinical oncology. *Nat. Cancer* **2022** *39* **3**, 1026–
448 1038 (2022).

- 449 20. Baxi, V., Edwards, R., Montalto, M. & Saha, S. Digital pathology and artificial intelligence in
450 translational medicine and clinical practice. *Mod. Pathol.* **35**, 23–32 (2022).
- 451 21. Hong, R., Liu, W., DeLair, D., Razavian, N. & Fenyö, D. Predicting endometrial cancer subtypes
452 and molecular features from histopathology images using multi-resolution deep learning models.
453 *Cell reports. Med.* **2**, (2021).
- 454 22. Schmauch, B. *et al.* A deep learning model to predict RNA-Seq expression of tumours from
455 whole slide images. *Nat. Commun.* **11**, (2020).
- 456 23. Levy-Jurgenson, A., Tekpli, X., Kristensen, V. N. & Yakhini, Z. Spatial transcriptomics inferred
457 from pathology whole-slide images links tumor heterogeneity to survival in breast and lung
458 cancer. *Sci. Reports 2020 101* **10**, 1–11 (2020).
- 459 24. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and
460 prognosis. *Nat. Cancer 2020 18* **1**, 800–810 (2020).
- 461 25. Loeffler, C. M. L. *et al.* Artificial Intelligence-based Detection of FGFR3 Mutational Status
462 Directly from Routine Histology in Bladder Cancer: A Possible Preselection for Molecular
463 Testing? *Eur. Urol. Focus* **8**, 472–479 (2022).
- 464 26. Farahmand, S. *et al.* Deep learning trained on hematoxylin and eosin tumor region of Interest
465 predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer. *Mod. Pathol.*
466 **35**, 44–51 (2022).
- 467 27. Lili, L. N., Matyunina, L. V., Walker, L. D., Daneker, G. W. & McDonald, J. F. Evidence for the
468 Importance of Personalized Molecular Profiling in Pancreatic Cancer. *Pancreas* **43**, 198 (2014).
- 469 28. Sicklick, J. K. *et al.* Molecular profiling of cancer patients enables personalized combination
470 therapy: the I-PREDICT study. *Nat. Med. 2019 255* **25**, 744–750 (2019).
- 471 29. Lam, K. H. B., Faust, K., Yin, R., Fiala, C. & Diamandis, P. The Brain Protein Atlas: a

- 472 conglomerate of proteomics datasets of human neural tissue. *Proteomics* 2200127 (2022)
473 doi:10.1002/PMIC.202200127.
- 474 30. Guérin, J., Gibaru, O., Thiery, S. & Nyiri, E. CNN features are also great at unsupervised
475 classification. 83–95 (2017).
- 476 31. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a
477 summary. *Neuro. Oncol.* **23**, 1231–1251 (2021).
- 478 32. Faust, K. *et al.* Intelligent feature engineering and ontological mapping of brain tumour
479 histomorphologies by deep learning. *Nat. Mach. Intell.* **1**, 316–321 (2019).
- 480 33. Lam, K. H. B. *et al.* Topographic mapping of the glioblastoma proteome reveals a triple-axis
481 model of intra-tumoral heterogeneity. *Nat. Commun.* **13**, (2022).
- 482 34. Puchalski, R. B. *et al.* An anatomic transcriptional atlas of human glioblastoma. *Science* (80-.).
483 **360**, 660–663 (2018).
- 484 35. Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous
485 Ovarian Cancer. *Cell* (2016) doi:10.1016/j.cell.2016.05.069.
- 486 36. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**,
487 382–387 (2014).
- 488 37. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer.
489 *Nature* **534**, 55–62 (2016).
- 490 38. Faust, K. *et al.* Visualizing histopathologic deep learning classification and anomaly detection
491 using nonlinear feature space dimensionality reduction. *BMC Bioinformatics* **19**, 173 (2018).
- 492 39. Wang, L. B. *et al.* Proteogenomic and metabolomic characterization of human glioblastoma.
493 *Cancer Cell* **39**, 509-528.e20 (2021).
- 494 40. Ben-Porath, I. *et al.* An embryonic stem cell-like gene expression signature in poorly

- 495 differentiated aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).
- 496 41. Richards, L. M. *et al.* Gradient of Developmental and Injury Response transcriptional states
497 defines functional vulnerabilities underpinning glioblastoma heterogeneity. *Nat. cancer* **2**, 157–
498 173 (2021).
- 499 42. Marusyk, A., Janiszewska, M. & Polyak, K. Intratumor heterogeneity: the Rosetta stone of
500 therapy resistance. *Cancer Cell* **37**, 471 (2020).
- 501 43. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image
502 Recognition. (2014).
- 503 44. Djuric, U. *et al.* Defining protein pattern differences among molecular subtypes of diffuse
504 gliomas using mass spectrometry. *Mol. Cell. Proteomics* mcp.RA119.001521 (2019)
505 doi:10.1074/mcp.RA119.001521.
- 506 45. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
507 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550
508 (2005).
- 509
- 510
- 511
- 512
- 513
- 514
- 515
- 516
- 517

518 **Methods**

519 **Ethics statement**

520 The University Health Network Research Ethics Board (UHN REB) approved the study REB #17-6193
521 as it has been found to comply with relevant research ethics guidelines, as well as the Ontario Personal
522 Health Information Protection Act (PHIPA), 2004. Patient consent was not directly obtained and instead
523 a consent waiver for this study was granted by UHN REB as the research was deemed to involve no more
524 than minimal risk as it used exclusively archival tissue specimens.

525

526 **Tissue cohort development and digital scanning**

527 All clinical cases included in our study cohort were retrieved from UHN archival tissue specimens. To
528 assess potential associations between DLFV correlation coefficients and human-perceived differences, an
529 initial brain tumor tissue cohort (n=40) largely comprising of diffuse glioma cases (IDH-wildtype and
530 IDH-mutants, WHO Grade 2-4) was developed. To compare proteome-level programs between distinct
531 partitions, cases were further examined to define regions relatively pure in tumor content (e.g. not regions
532 of low tumor purity) and with partitions sufficiently large enough for laser capture microdissection (LCM)
533 and LC-MS/MS analysis. In total, the LC-MS/MS cohort included seven high-grade gliomas³¹. Six cases
534 had glioblastoma histology and an IDH-wildtype immunohistochemical staining pattern. The final case
535 was an IDH-wildtype high grade glioma with a WHO grade 3 histology. Age ranges were provided for
536 anonymity and patient IDs were not known to anyone outside of research group (**Supplementary Table**
537 **1**). Additional independent confirmatory cases were also selected which displayed region-to-region
538 heterogeneity in their MIB1/Ki-67 proliferation indices (n=5 independent cases). The collision tumor in
539 the skin and the lung adenosquamous carcinoma were also retrieved from our local cancer center (UHN).
540 The H&E section of the metastatic lung cancer model was provided directly from the authors of the

541 relevant study⁶. The entire generated local cohort was digitally scanned as WSI with a compression quality
542 of 0.70 and a magnification of x20 on a Leica Aperio AT2 scanner.

543

544 **Implementation of the deep convolutional neural network**

545 We used a previously trained pathology-optimized version³² of the VGG19 CNN⁴³ to extract the deep
546 learning feature vectors used during HAVOC partitioning³⁰. Specifically, the original VGG19 model was
547 optimized using transfer learning on a set of 838,644 pathologist-annotated image patches spanning over
548 70 brain tissue and tumor types derived from over 1000 patients³². We previously showed that this training
549 diversity lead to node-weights within the network that align with human-discernible histopathological
550 feature representations that are applicable across multiple malignancies^{15,32}. Given the unsupervised
551 nature of HAVOC, the CNN was used without having to undergo any additional training or optimization
552 for this application. Previously learned class labels from the original fine-tuning of the VGG19 were
553 decoupled from this analysis and instead only the feature representations, in the form of a 512-dimensional
554 vector from the global average pooling layer of the CNN, prior to SoftMax reduction and classification,
555 were extracted and used for image-based clustering³².

556

557 **Regional partitioning and estimation of heterogeneity**

558 Tissue partitions by HAVOC were generated using an unsupervised image clustering framework as
559 previously described¹⁵. Briefly, WSIs for each clinical specimen of interest are first tiled into individual
560 0.066-0.27 mm² image patches. The particular size of the image patched was largely dictated by the size
561 of the available tissue (smaller tiles for smaller tissue) and histomorphologic pattern of interest (e.g.
562 cytoarchitecture vs. nuclear patterns benefiting from larger (low power) vs. smaller (high power
563 magnification) tiles respectively). Histomorphologic fingerprints for each tile are then generated by

564 averaging the Deep Learning Feature (DLF) values from the final global average pooling layer of the
565 VGG19 CNN. We refer to this 512-dimensional feature representation matrix as the Deep Learning
566 Feature Vector (DLFV). As this network was tuned on a diverse set of histological images, many
567 individual DLFs align with human-identifiable histologic features, including fibrosis, epithelium, and
568 mucoid patterns thus allowing it to group images patches, with relatively similar morphologies, into
569 meaningful clusters³². To identify spatial transitions in histomorphological patterns for each case, the
570 DLFVs generated from each tile are clustered, using Ward hierarchical clustering, as previously
571 described³². Images that group together by specific clustering threshold (default k=9) are deemed to show
572 a relatively similar histomorphologic signature and grouped together to form a HAVOC partition. In our
573 experience, these often tend to show spatial relationships concordant with expert histopathologic review.
574 The relative distances (similarity) of each HAVOC partition (spatial region) are quantified using the
575 Pearson correlation coefficient of each region's average DLFVs. In addition to quantitatively correlating
576 with histomorphologic differences perceived by humans, we found that the overall correlation coefficient
577 is less affected by artifacts that often skew results when raw DLFV were used to cluster regions spanning
578 multiple slides. Therefore, while we find clustering of raw tile DLFVs from individual slides effective,
579 we observe more robust results with correlation coefficients are used to compare HAVOC partitions from
580 independent slides.

581

582 **Region-specific deep learning feature selection and mapping**

583 To understand which morphologic features were potentially driving, or at the very least associated with,
584 the unsupervised slide subgroupings, we first identified the most significantly enriched DLFs in each of
585 the HAVOC-proposed groupings. To visually determine what morphology such DLFs were detecting, we
586 then assess image tiles at the extremes of the DLFs of interest across the entire slide (both the highest-

587 scoring and the lowest scoring). Interpretations of potential morphological differences between the tiles
588 at the two extremes were provided by two or more blinded pathologists. This qualitative expert assessment
589 was then complimented and validated by generating isolated feature activation maps (FAMs) of the
590 candidate DLFs of interest to evaluate if the activated tile coordinates matched the location of the
591 histologic feature suggested by expert pathologists. Python code used to generate the FAMs can be found
592 at BitBucket (<https://bitbucket.org/diamandislabii/faust-feature-vectors-2019>).

593

594

595 **Pathologist evaluation of HAVOC-defined heterogeneity**

596 Both computational silhouette and human evaluation approaches were employed to evaluate the
597 pathologic correlations of HAVOC-proposed regions of heterogeneity and determine the optimal number
598 of WSI partitions when appropriate. For most cases, we found that the vast degree of pathologist-
599 discernible patterns of histomorphologic variations reached saturation after 7-8 WSI partitions irrespective
600 of the overall level of complexity and tissue heterogeneity. While 7 partitions worked well for mapping
601 most histomorphologic patterns of variation in our clinical cohort, the clustering framework is easily
602 tunable and can be extended to addition partitions using a stepwise k-means clustering approach to explore
603 additional and more subtle regions of heterogeneity in complex tissue specimens (**Supplemental Fig 1,**
604 **15 & 16**). To capture finer microscopic details, relevant in some circumstances, the size of the image
605 patches can also be tuned to the appropriate application. For examination of relatively large regions of
606 cellular tumor, image patches with a width of 512 and 1024 pixels works extremely well (**Fig 2**). For more
607 focal deposits (**Fig 6**), the tile size can be reduced to 256 or 512 pixels in length to reduce the intra-patch
608 pattern variation.

609

610 **Laser capture microdissection (LCM) and LC-MS/MS proteomic profiling**

611 Tissue sections were stained prior to LCM to improve contrast and provide references across the slides to
612 guide precise microdissection as previous described^{33,44}. Specifically, formalin-fixed paraffin-embedded
613 (FFPE) tissue blocks were sectioned (at 10 μ m thickness) and mounted onto Leica PEN slides (Cat No.
614 11505189). Slides were subsequently deparaffinized with 100% xylene (2x), 100% ethanol, 95% ethanol,
615 70% ethanol, and 50% ethanol (3 minutes each). These slides were then stained with hematoxylin (1
616 minute), rinsed in de-ionized water (1 minute), and stained in 1% eosin Y (Fisher scientific). Slides were
617 finally air dried prior to laser capture microdissection.

618 Regions of interest from these sections were micro-dissected using a Leica LMD 70000 (Leica
619 Microsystems, Inc., Bannockburn, IL). HAVOC-generated color tiled maps were used as a reference to
620 guide parameters for dissection. Samples were collected in an Eppendorf tube and subsequently stored at
621 room temperature until further sample preparation began.

622 Proteomic extraction was performed with the addition of 50 μ L of 1% Rapigest, 200 μ L of a
623 dithiothreitol, ammonium bicarbonate (50 mM), and tris-HCl solution to each sample. Samples were
624 subsequently sonicated using a Bioruptor Plus on high with 30 second intervals for 1 hour. Solutions were
625 heated to 95 degrees Celcius for 45 minutes, followed by 80 degrees Celsius for 90 minutes with a
626 ThermoMixer. 20 μ L of iodoacetamide was then added to each solution in the absence of light for
627 alkylation. 1 μ g of trypsin/Lys-C mix was then added to each sample and reacted overnight at 37 degrees
628 Celsius. The solutions were subsequently acidified with trifluoroacetic acid at a final concentration of 1%
629 ahead of stagetip cleanup.

630 In preparation for mass spectrometry analysis, samples were desalted using Omix C18 tips
631 following manufacturer protocol. Elution of peptides was completed with 3 μ L (0.1% formic acid, 65%
632 acetonitrile) and dilution of peptides was completed with 57 μ L (0.1% formic acid in MS water). 18 μ L

633 of solution (2.5 ug of peptides) was loaded with an autosampler for mass spectrometry as previously
634 described³³. Briefly, peptide elutions from an EASY-Spray column ES803A occurred at a rate of 300
635 nL/min with increasing concentration of 0.1% formic acid in acetonitrile over a one hour gradient. This
636 setup was coupled to a Q Exactive HF-X with a spray voltage of 2 kV with a 60 minute data-dependent
637 acquisition method. The full MS1 scan was completed from 400-1500 m/z at a resolution of 70, 000 in
638 profile mode, with the top 28 ions selected for further fragmentation with HCD cell. Fragment detection
639 occurred with an Orbitrap using centroid mode set at a resolution of 17,500. The following MS parameters
640 were used: MS1 Automatic Gain Control (AGC) target was set at 3×10^6 with maximum injection time
641 of 100 ms, MS2 AGC set at 5×10^4 with maximum injection time of 50 ms, isolation window of 1.6 Da,
642 underfill ratio 2%, intensity threshold 2×10^4 , normalized collision energy (NCE) of 27, charge exclusion
643 was set to fragment 2+, 3+ and 4+ charge state ions only, peptide match was set to preferred and dynamic
644 exclusion set to 42 (for 90 min method).

645

646 **Entire tumor histomic profiling and organization**

647 HAVOC evaluation of larger and even entire tumor specimens was completed by histomic profiling of
648 each H&E-stained section for each of the corresponding tumor block. Unsupervised analysis of each WSI
649 was completed individually as described in previous sections and then merged in a separate and final step.
650 For the example of the 12-slide tumor specimen shown, the number of clusters for each WSI was set to 7;
651 within the range that results in saturation of perceivable histomorphologic patterns at a tile size of 0.27
652 mm². The DLFVs of each cluster from each of the related WSIs of the specimen were then subsequently
653 organized using pairwise Pearson Correlation coefficients. We provide a separate script for this final step
654 in the repository. Multi-slide Pearson correlation matrixes of HAVOC proposed groupings were generated
655 using Matplotlib and Seaborn Python data visualization libraries. Evaluation of the fidelity of the multi-

656 slide regional alignment was assessed by clustering of expert annotations of each HAVOC-partition,
657 immunohistochemical staining patterns (MIB1/Ki-67) and LC-MS/MS analysis for the 3 patients with
658 paired slides.

659

660 **Statistical Analysis**

661 MaxQuant Andromeda (version 1.5.5.1) search engine was used to process mass spectrometry raw data
662 files against the Human Swissprot protein database (July, 2019 version). Proteins were filtered to include
663 only those appearing in at least 60% within a sample. Raw protein values were Log₂ transformed, with
664 non-valid values imputed (downshift=0.3, width=1.8). Analysis of proteomic data was performed using
665 biostatistical platforms Perseus (www.coxdocs.org) and ssGSEA⁴⁵ ([https://gsea-
666 msigdb.github.io/ssGSEA-gpmodule/v10/index.html](https://gsea-msigdb.github.io/ssGSEA-gpmodule/v10/index.html)). Single sample gene set enrichment analysis was
667 used to define pathways enriched in each HAVOC proposed region. One-way ANOVA testing and
668 Tukey's post hoc test were completed in R (v4.0.4) to identify differences in enriched pathways across
669 HAVOC proposed groupings. Student's t-tests were used to test the difference in Ki-67 proliferation
670 indices across HAVOC proposed groupings. Additional statistical tests used for individual analyses are
671 mentioned in the appropriate text and figure coordinates.

672

673 **Gene set enrichment analysis**

674 Gene sets enrichment analysis was used to understand the biological significance of the regionally distinct
675 molecular profiles both on individual samples (**Fig 2**) and at a cohort level (**Fig 5**). Three gene expression
676 datasets were analyzed: (i) Proteomics dataset generated in this study. For this, proteins with >25% of
677 missing values were removed, resulting in a dataset of LFQ intensity values with 1,920 proteins across 60
678 injected aliquots, which typically included 3 technical replicates per tissue specimen. (ii) RNA-seq data

679 and associated clinical information from the merged cohort of LGG and GBM (Ceccarelli et al, *Cell*, 2016)
680 retrieved from cBioportal (http://www.cbioportal.org/study/summary?id=lgggbm_tcga_pub).
681 Oligodendrogliomas were excluded from the analysis resulting in a cohort 446 diffuse “astrocytic” tumors.
682 (iii) The final cohort represented normalized scRNA data previously published by Richards et al. (*Nature*
683 *Cancer*, 2020). This dataset was downloaded from the Broad Institute’s Single Cell Portal
684 (https://singlecell.broadinstitute.org/single_cell/study/SCP503). To reduce signature enrichment bias due
685 to poor sequencing coverage, samples with less than 10,000 UMIs were excluded from the analysis, and
686 the resulting dataset contained 28 glioma stem cells samples and 6 patient derived GBM samples.

687 The gene expression profiling was focused on a list of 64 genesets previously described by our
688 group (Lam et al, *Nature Communications*, 2022) which were selected from the MSigDB-7.2 database
689 (<https://www.gsea-msigdb.org/gsea/msigdb>) on the basis of being informative in glioblastoma and
690 showing a high degree of proteo-transcriptomic correlation. Here, we used a revised list of 64 genesets
691 that differs from the previously published as follows: (i) The pathway
692 “REACTOME_HSP90_CHAPERONE_CYCLE_FOR_STEROID_...” was dropped during the version
693 upgrade to MSigDB-7.4, and (ii) the signature LEIN_ASTROCYTOMA_MARKERS was included as a
694 result of screening for signatures informative of “astrocytic” differentiation. Single sample geneset
695 enrichment analysis (ssGSEA) was performed with the Bioconductor package GSVA v1.44.2.

696 The Astro-ES axis is a composite score designed to cover the spectrum of cellular differentiation
697 that is found in diffuse glioma, where high values correspond to a “well differentiated” astrocytic
698 phenotype and lower values with an embryonic “poorly differentiated” state. This score is calculated by
699 zero-centering the ssGSEA scores of the LEIN_ASTROCYTOMA_MARKERS and
700 BENPORATH_ES_1 genesets, and then subtracting the latter from the former.

701

702 **Estimation of regional cellular density in HAVOC-defined regions**

703 We estimated cell density in profiled regions to complement the interpretation of HAVOC defined regions.
704 To do this WSI data files (.svs format) were loaded on QuPath v0.3.2. For each region of interest, 5 circular
705 sampling areas spanning 100-125 μm^2 and histologically representative of the region of interest, were
706 delineated, and followed by running Estimate Stain Vectors and Cell Detection Tool with default
707 parameters. The density of each sampling area was computed as follows: number of detections x 100 /
708 area, and finally, the regional cellular density is the average of the cellular density of the 5 sampling areas.
709 Finally, the regional cellular density was categorized according to the following arbitrarily selected
710 reference framework: Ultra-High (>1.2), High (8-1.2), Mid (6-8) and Low (<6).

711

712 **Data availability**

713 The mass spectrometry proteomics data of all HAVOC derived regions presented in this manuscript have
714 been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset
715 identifier PXD037548 (username: reviewer_pxd037548@ebi.ac.uk; password: fy4uPFAi). The data can
716 also be examined directly through our inter-active data portal (Brain Protein Atlas²⁹;
717 <https://www.brainproteinatlas.org/dash/apps/ad>). Labeling legend for mass spectrometry proteomics data
718 can be found at <https://bitbucket.org/diamandislabii/havoc>.

719

720 As described, some of the data used in this publication derived from The Cancer Genome Atlas
721 Program (TCGA) and deposited at the Data Coordinating Center (DCC) for public access
722 [<http://cancergenome.nih.gov/>]. The RNA-Seq IvyGAP data used are publicly available at Gene
723 Expression Omnibus through GEO series accession number GSE107560. The single-cell are publicly
724 available through the Broad Institute Single-Cell Portal

725 (https://singlecell.broadinstitute.org/single_cell/study/SCP503) and CReSCENT60
726 (<https://crescent.cloud>; study ID CRES-P23). Additional proteomic data from different glioblastoma
727 regions were also derived from Lam et al³³ are also publicly available through the ProteomeXchange
728 Consortium via the PRIDE partner repository with the dataset identifier PXD019381. The H&E slide and
729 ground truth annotations for the metastatic lung carcinoma mouse model was provided directly from the
730 authors and the relevant study⁶.

731

732 **Code availability**

733 Code for the 512 dimensional feature vector extractor, feature activation mapping (FAM) and the original
734 trained VGG19 model used in HAVOC is available on Bitbucket
735 (<https://bitbucket.org/diamandislabii/faust-feature-vectors-2019>) and
736 (<https://doi.org/10.5281/zenodo.3234829>). Python source code and an interactive Colab notebook for
737 running HAVOC and integrating multiple slides into a single analysis are available at
738 <https://bitbucket.org/diamandislabii/havoc> and
739 <https://colab.research.google.com/drive/1Gx7gXTBIBF5iNY0REL7QktM-oAMv4W6S?usp=sharing>.
740 HAVOC can also be tested and run directly within a web browser by uploading a .SVS digital WSI on
741 <https://www.codido.co>

742

Figure 1

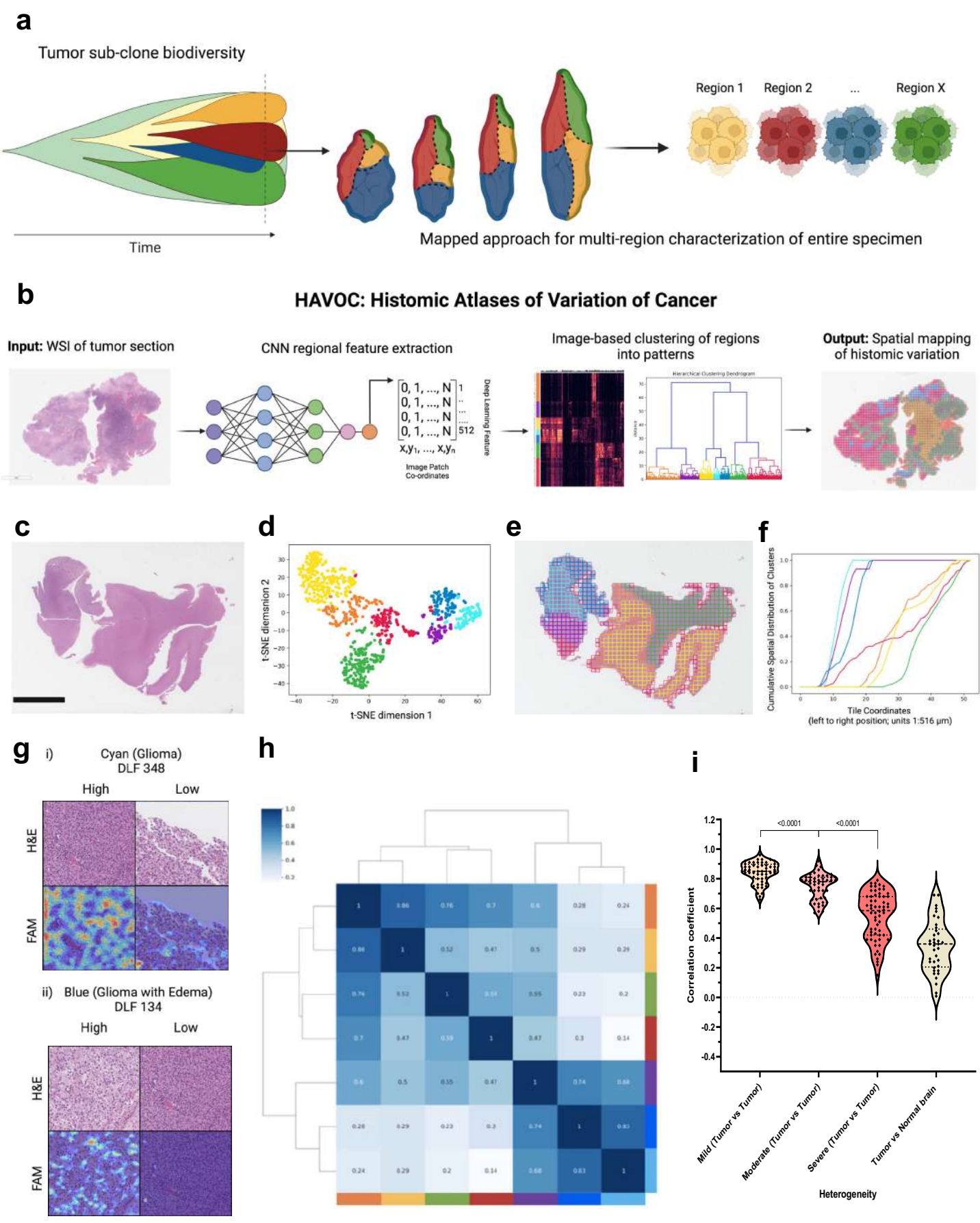


Figure Legends

Figure 1 | Mapping biodiversity in cancer tissue with HAVOC (a) Cartoon depicting regional evolution of intra-tumoral subclones and a map-guided approach to sampling. (b) HAVOC workflow summary. A pre-trained convolutional neural network (CNN) is used as a morphology feature extractor for input images. The generated deep learning feature vectors (DLFVs) of individual tiles are then used to carry out image-based clustering and map spatial coordinates to distinct histomic fingerprints back to the original whole slide image (WSI). (c-e) Representative mapping of a diffuse glioma. The relative spectrum of morphologic patterns of image tiles can be explored by dimensionality reduction (e.g. t-SNE) and clustering. Both highlight distinct tumor (blue, purple cyan) and non-tumoral (yellow, green, red) regions. Scale bar: 6 mm. (f) Horizontal (vertical left to right raster) cumulative distribution plot of the overall fraction of tiles from each HAVOC-defined cluster in panel d across entire WSI highlighting non-random spatial distributions of HAVOC-defined histomorphologies (Kolmogorov–Smirnov statistic of horizontal distribution of cellular tumor_{Cyan} vs cellular tumor with edema_{blue} = 0.69, $p=1.2e^{-11}$) (g) Mapping of individual DLFs over-represented in these HAVOC-defined tumor regions highlight interpretable morphological patterns of these defined glioma niches (e.g. tumor nuclei and edema respectively). (h) Pairwise Pearson correlation coefficients (r) of the DFLVs of HAVOC-defined partitions in panel (e) highlighting inverse correlation with the degree morphological heterogeneity (and t-SNE distances in panel (d)) across this representative case (i) Violin plots showing concordance between HAVOC r values and semi-quantitative assessments of regional heterogeneity (e.g. mild, moderate, or severe) by experts. DLFV r of lesional vs non-lesional regions included as reference.

Figure 2

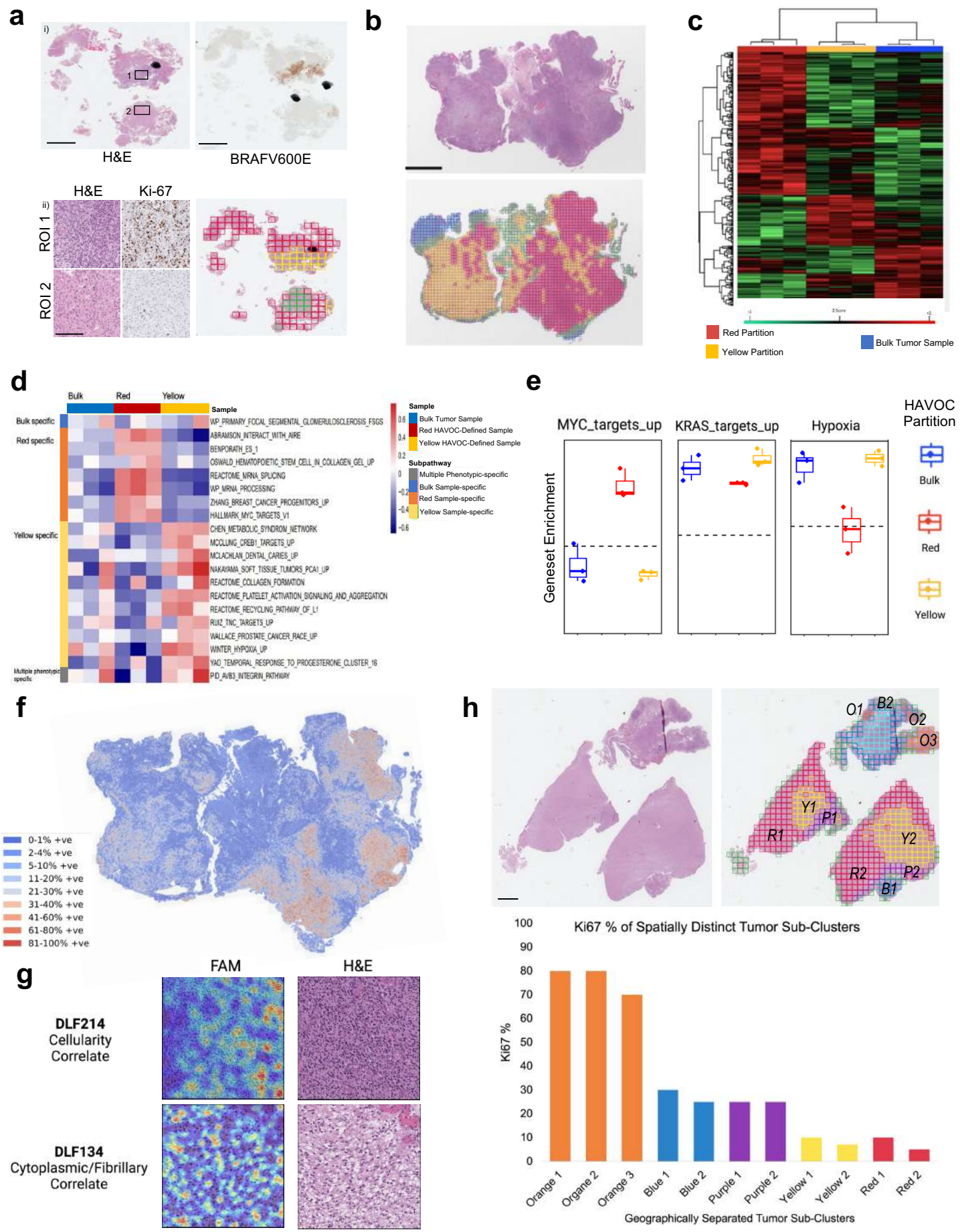
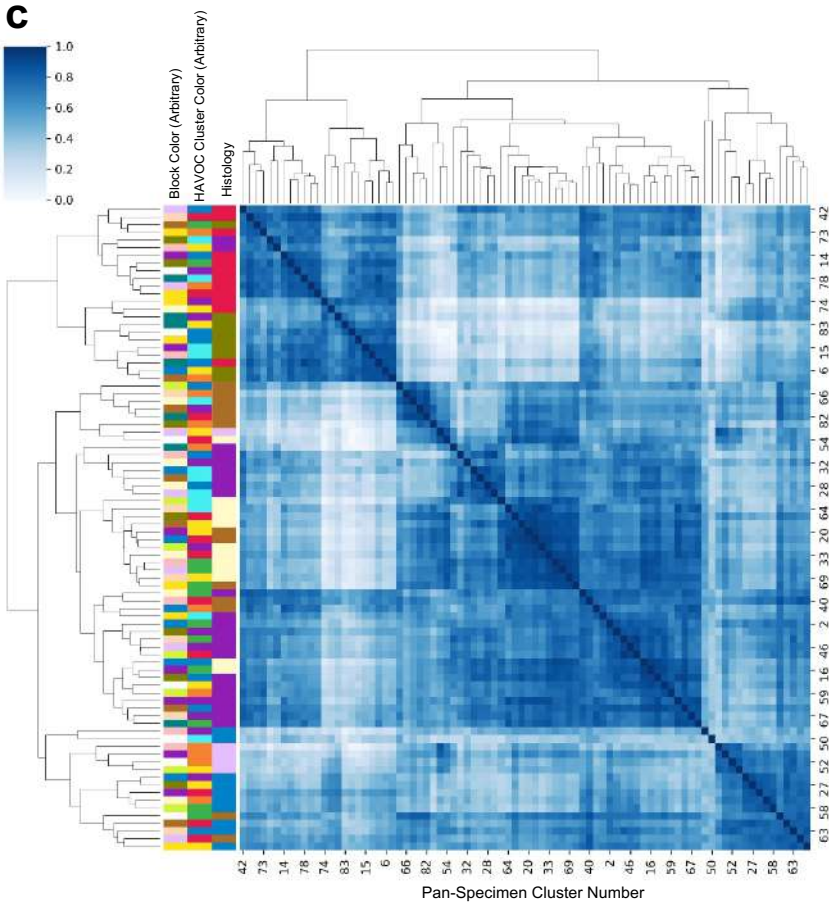
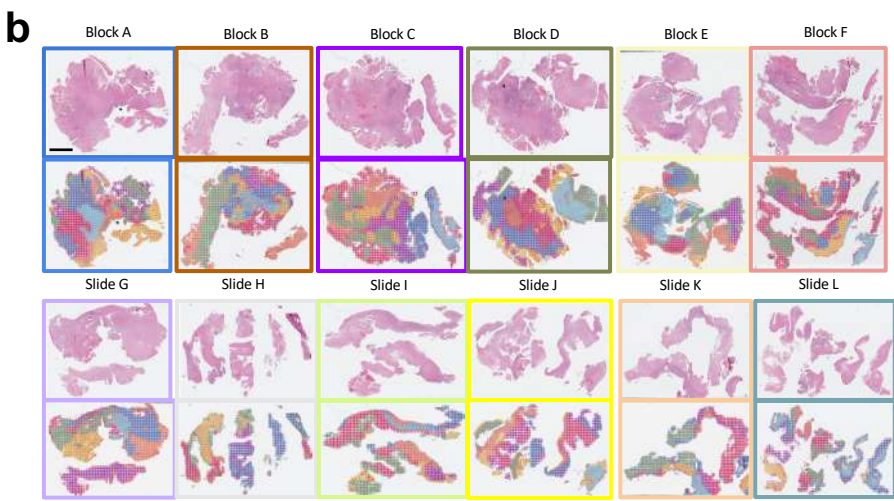
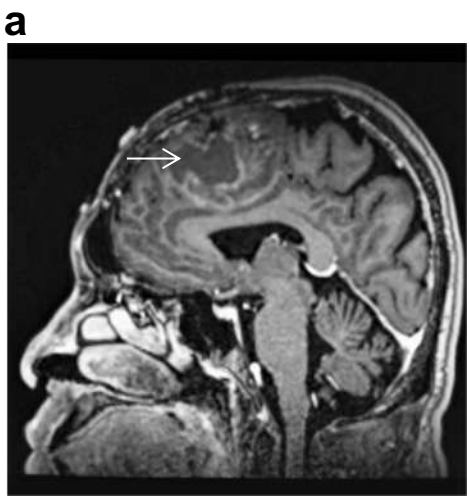


Figure 2 | HAVOC-defined partitions align with regional biodiversity. (a) Histopathology images of an IDH-wildtype glioblastoma demonstrating intra-tumoral heterogeneity with a BRAFV600E-mutated hyper-proliferative subclone resolved by immunohistochemistry. This genetic and biologically-defined subclone was resolved by a specific HAVOC partition (yellow). Scale bar: 2 mm. (b) WSI image of a glioblastoma mapped with HAVOC. Scale bar: 4 mm. (c) LC-MS/MS profiling of HAVOC-defined partitions from panel (b) shows distinct molecular profiles from each other and the overall (bulk) specimen. (d-e) Single-sample Gene Set Enrichment Analysis highlights heterogeneity across various programs and relevant glioma axes including MYC (proliferation), KRAS (invasion), and hypoxia. Bulk signatures provided for reference. (f) Spatial proliferation index differences (as assessed via CNN-based quantification of Ki-67) spatially align with HAVOC partition. (g) Feature activation mapping define morphologic correlates of the profiled HAVOC partitions. Partitions showing proliferative and invasive biology show high cellularity and cytoplasmic/fibrillary patterns respectively. (h) Representative H&E and HAVOV map of an IDH-wildtype glioblastoma section with geographically separated subclusters (e.g. Orange O1, O2, O3) showing similarly grouped histomic (HAVOC) signatures. Despite distinct spatial coordinates, subclusters belonging to the same HAVOC partitions displayed similar biology (e.g. Ki-67 proliferation indices). Scale bar: 2 mm.

Figure 3



- Leading Edge of Tumor
- Nodule
- Moderate Cellularity
- Normal White Matter
- High Cellularity
- Hemorrhage
- Artifacts

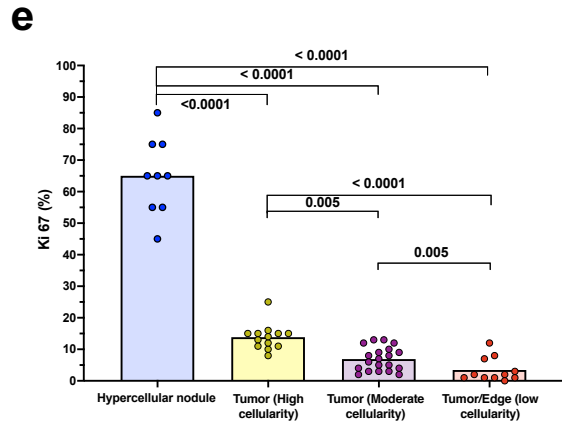
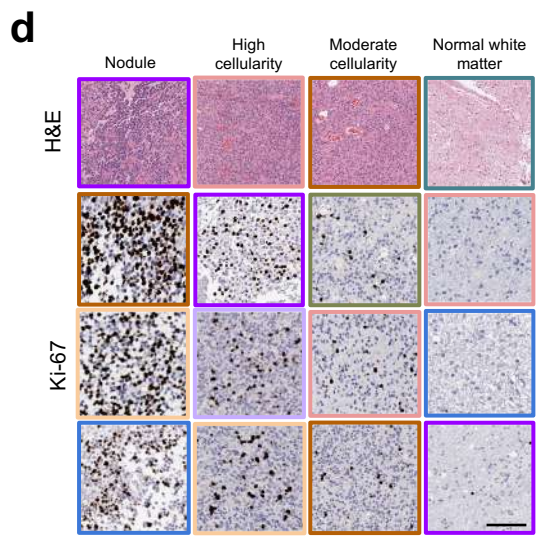


Figure 3 | Small scale mapping of biovariation across entire tumor specimens using HAVOC
(a) MRI image of a recurrent IDH-mutated, 1p19q co-deleted oligodendroglioma measuring 5.4 x 4.1 x 1.8 cm in dimensions showing a heterogeneous pattern with multiple cysts and variable contrast enhancement. (b) 12 sequential H&E sections of the entire tumor with accompanying HAVOC heterogeneity maps. (c) Pairwise Pearson correlation matrix arranging all 84 clusters shown in panel (b). Slide colours corresponds to the same colours shown in former panel. There is a strong association of the clusters with expert-annotated morphologic patterns. (d) Representative images of Ki-67 (MIB1) from the different clustered regions. (e) Histogram of estimated Ki-67 (MIB1) across different regions.

Figure 4

Patient II

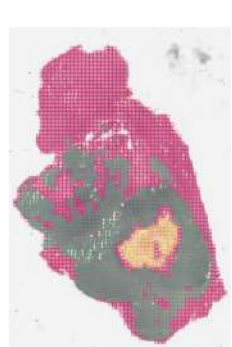
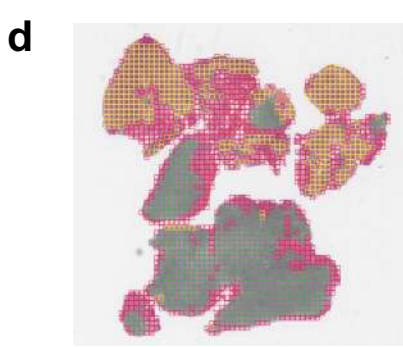
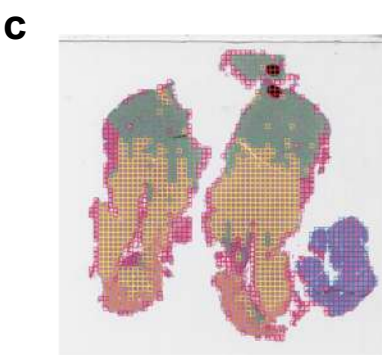
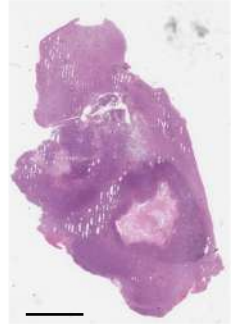
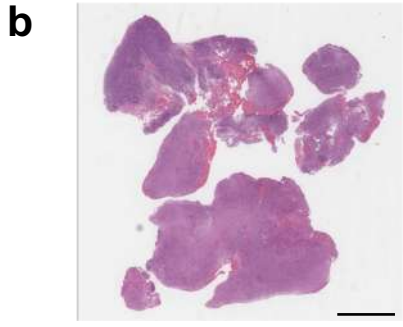
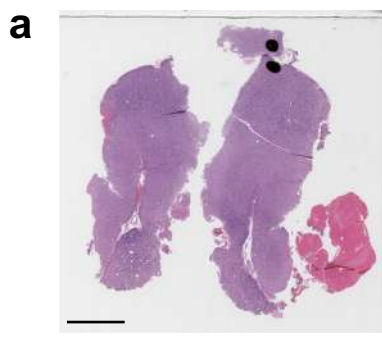
Patient III

Ila

I Ib

IIIa

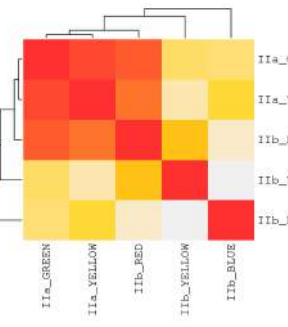
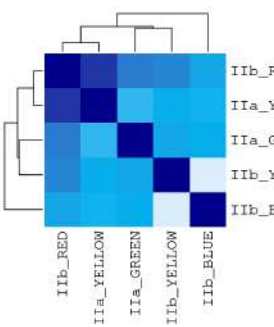
IIIb



e

Regional DLFV Signatures

Regional Protein Signatures



f

Regional DLFV Signatures

Regional Protein Signatures

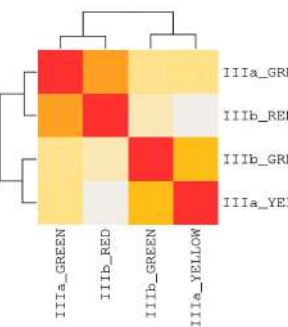
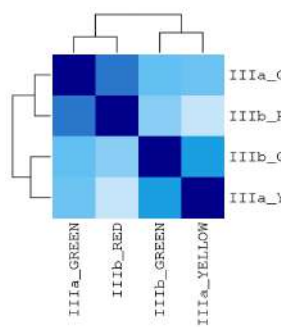


Figure 4 | HAVOC mapping across independent slides align with overall molecular correlations. (a) HAVOC partitions across paired slides from an individual IDH-wildtype high grade glioma (b) Integrated DLFV correlation matrices across both slides define a distinct focal hyperdense outlier region (Sample Iib blue cluster). These multi-slide maps are in agreement with global patterns of proteomic variations derived from each of the major HAVOC defined partitions. (c) Another example of HAVOC multi-slide regional partitions that identified reciprocally aligned regions across slide pairs. (d) HAVOC-defined inter-slide niche similarities and agreement with proteomic variations.

Figure 5

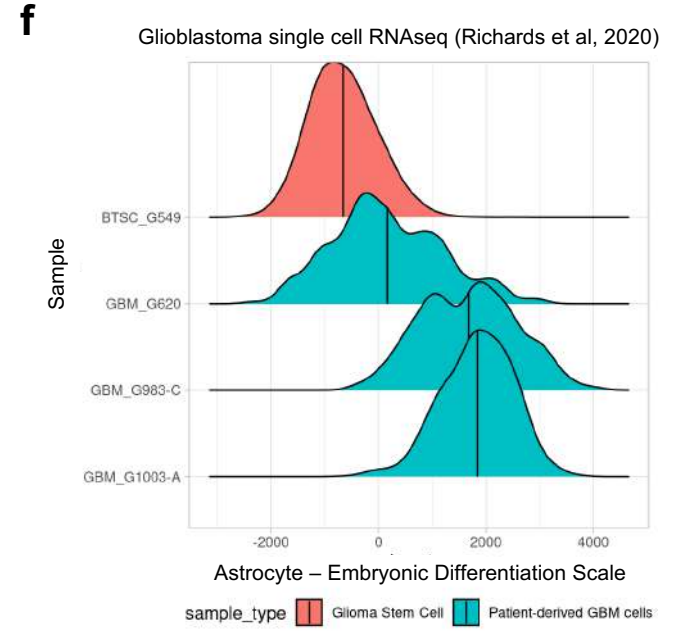
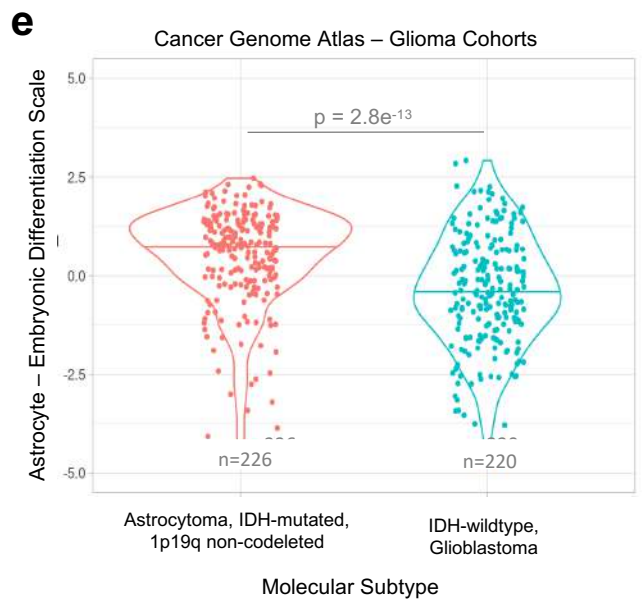
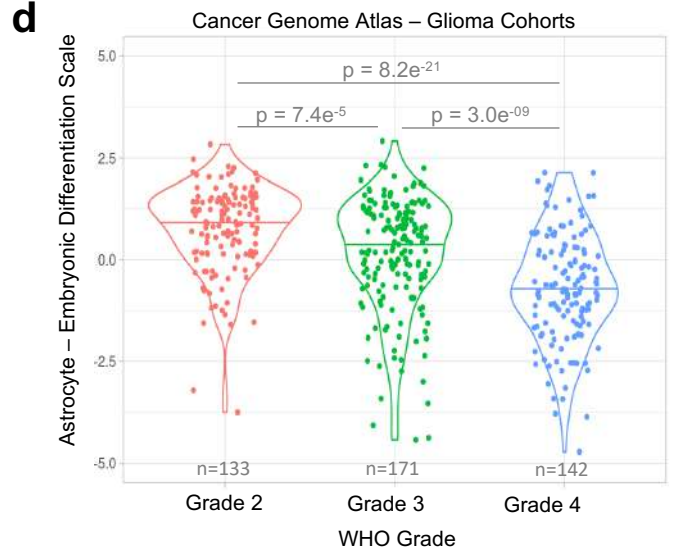
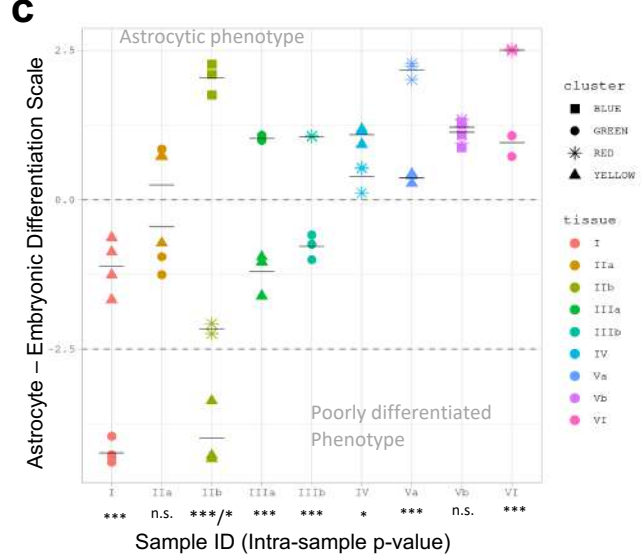
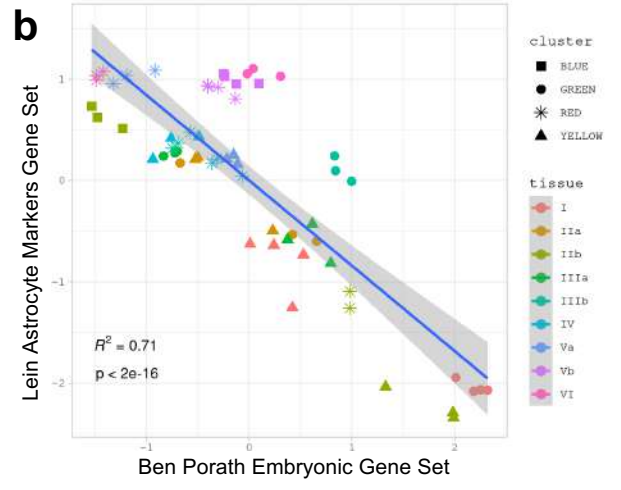
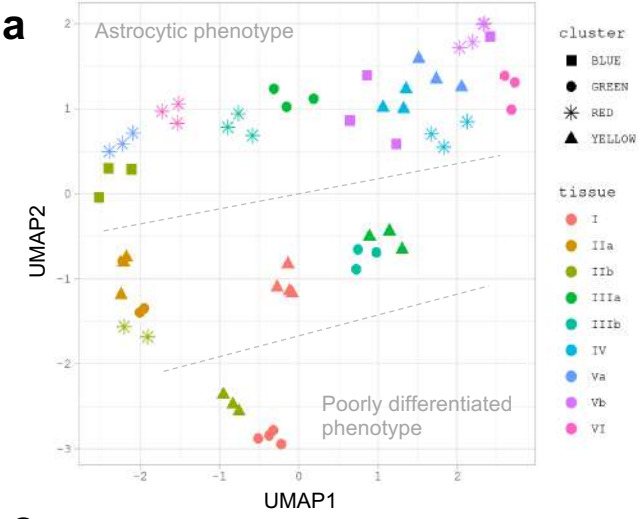
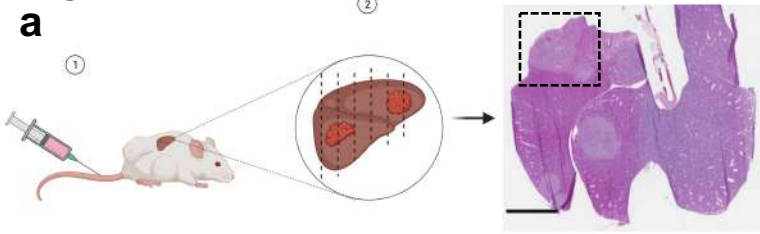


Figure 5| HAVOC reveals spatially organized patterns of molecular heterogeneity in high grade gliomas. (a) Unsupervised analysis of 19 HAVOC defined tumor regions across 6 high grade gliomas by performing UMAP dimensional reduction of the ssGSEA scores of 64 proteogenomically concordant genesets (b) Inverse relationship between the ssGSEA scores of the embryonic stem cell state (Ben Porath ES_1) and the Astrocytic differentiation geneset (Lein Astrocyte Markers Gene Set). p-value generated by linear fit model. (c) Regional differences in state of differentiation (Astro-ES axis) across each of the spatially resolved and profiled slides. The level of statistical significance of the differences between the regions of each tissue slide was assessed by ANOVA; p-values are indicated as follows: *** < 0.001, ** < 0.01, * < 0.05, and n.s. = not significant. (d) Distribution of the Astro-ES axis in astrocytic tumors (non-1p19q codeleted) from TCGA-GBM and TCGA-LGG cohorts stratified by WHO grade and (e) IDH status. (f) Varying distributions along the Astro-ES axis at the single-cell level in patient-derived GBM cells (n=3) and glioma stem cells (n=1); dataset previously published by Richards et al.

Figure 6



b i) H&E Whole Slide Image ii) Identification of Metastatic Clones

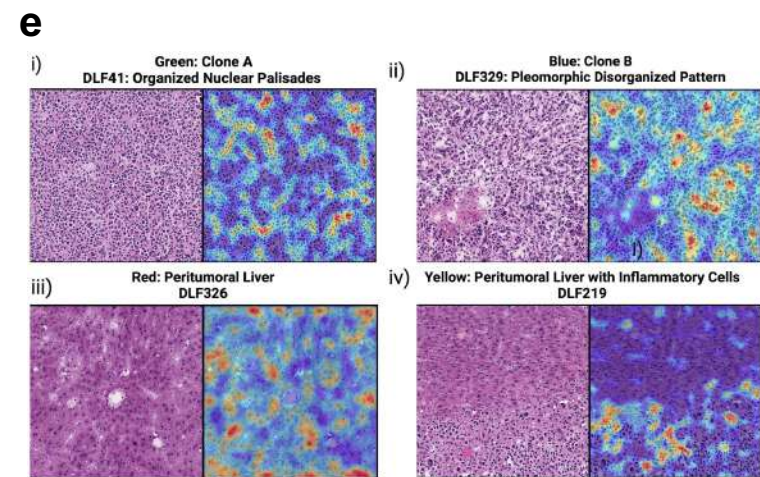
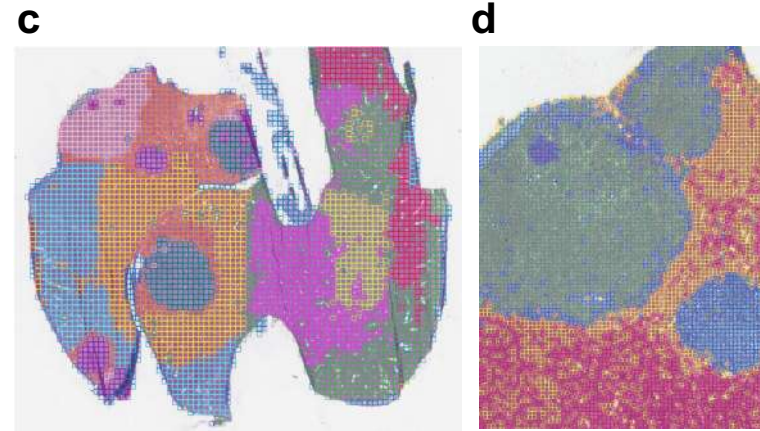
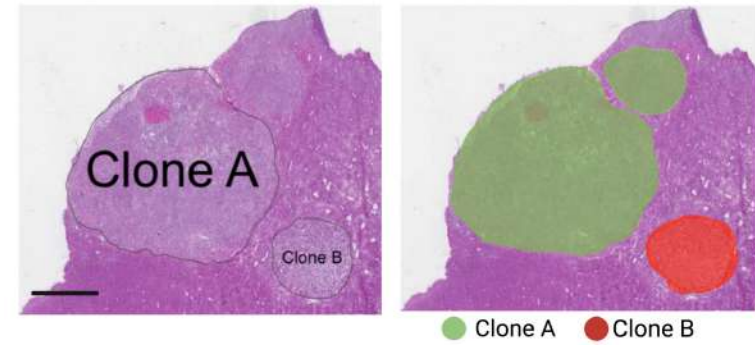


Figure 6| HAVOC defines genetically distinct metastatic clones (a) Schematic and H&E-stained section of a mouse liver modeling polyclonal $Kras^{G12D/+}Trp53^{-/-}$ lung cancer metastases from *Zhao et al.* Spatial profiling in the boxed region allowed benchmarking of HAVOC in this model. Scale bar: 5 mm. (b) Slide-DNA-seq and Slide-RNAseq of focused region provided ground truth of normal liver, tumor clone “A” and “B”. Dotted lines in the right panel indicate tumour boundaries. Scale bar: 2 mm. (c) HAVOC mapping of entire H&E sections with 11 partitions to ensure saturation of the different histomorphological patterns. HAVOC segmented all five tumors into stable groupings, identified in pink, teal, and purple. (d) HAVOC partitions ($k=4$; tile width: 128 pixel) of the focused region mapped by slide-DNA seq (original study) spatially divided tumor into 2 homogeneous subclones (green vs blue). Surrounding liver tissue was also separated into regions of peritumoral liver and immune infiltrates (red vs yellow) that match single-cell RNA seq projection (see supplemental Fig 13a for ground truth from original study). (e) FAMs of selected differentially activated DLFs for each tumor regions (i-ii) and DLFs enriched in peritumoral region highlighting liver regions with and without inflammatory cells (iii-iv).

Figure 7

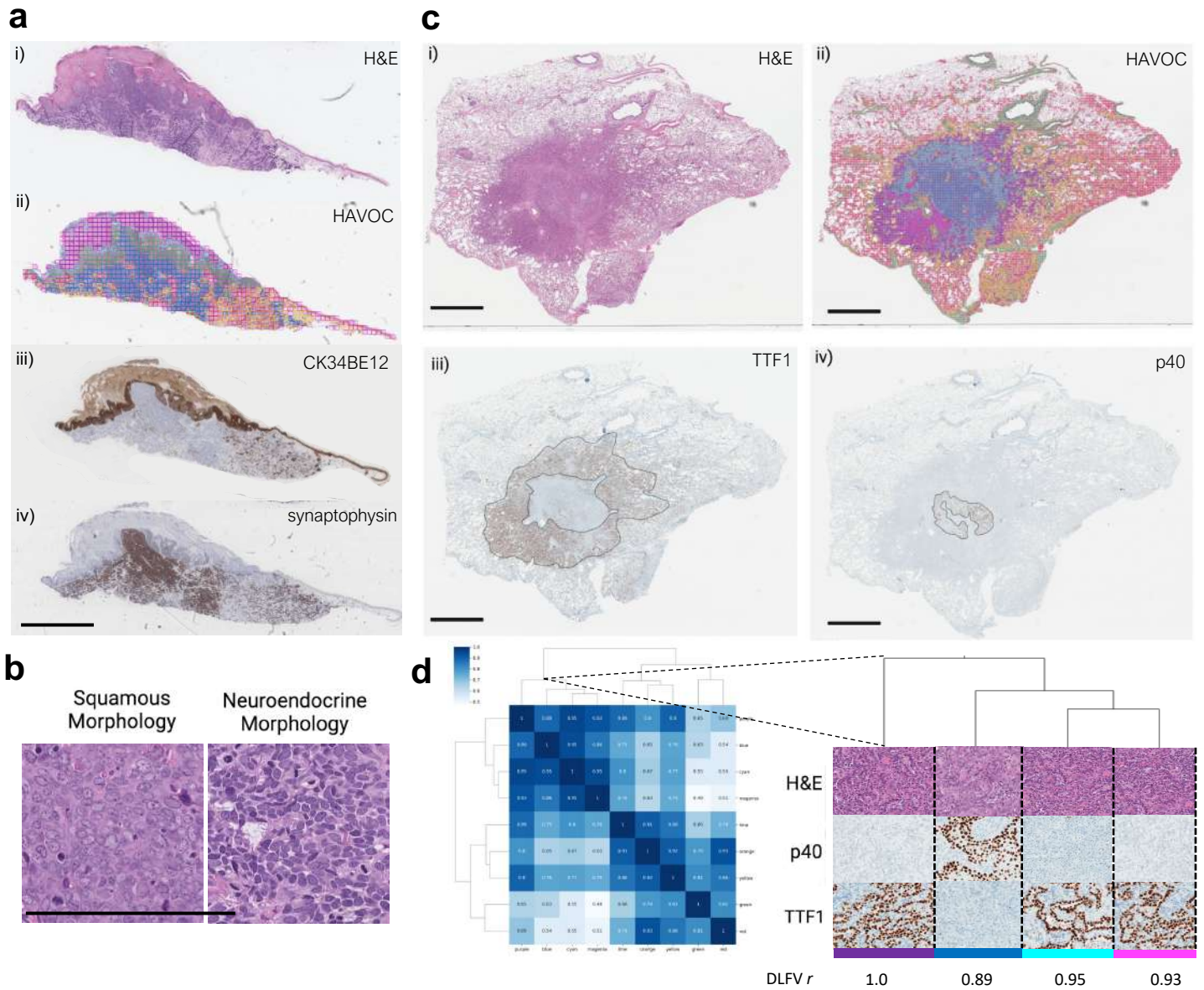


Figure 7 | HAVOC generalizes to untrained tissue types (a) (i) Dermatopathological specimen of a “collision” tumor comprised of distinct regions of squamous cell and neuroendocrine carcinoma. (ii) HAVOC partitions resolved the distinct tumor types that matched the immunohistochemical ground truth (iii) CK34BE12 and (iv) synaptophysin. Scale bar: 2 mm. (b) Representative high power magnification micrographs of HAVOC-defined distinct tumoral sub-clones in dermatopathological specimen highlighting the squamous (green partition) and neuroendocrine (blue partition) components. Scale bar: 200 μ m. (c) (i) H&E-stained lung resection with a neoplasm showing divergent adenosquamous differentiation. Tumor sub-components are highlighted with distinct (ii) p40 (squamous) and (iii) TTF1 (adenocarcinoma) immunohistochemical staining. (iv) HAVOC partitions of original WSI show subregions that align with the distinct tumoral patterns defined by the squamous/adenocarcinoma markers. Scale bar: 6 mm. (d) Hierarchical clustering of HAVOC-defined regions and representative (i) H&E, (ii) p40, and (iii) TTF1 images reveals two distinct tumor sub-patterns.