



Intratumoral heterogeneity in microsatellite instability status at single-cell resolution

Title	Intratumoral heterogeneity in microsatellite instability status at single-cell resolution
Author(s)	Anthony, Harrison;Seoighe, Cathal
Publication Date	2026-02-05
Publisher	Cell Press
Repository DOI	https://doi.org/10.1016/j.isci.2026.114860

Journal Pre-proof



Intratumoral heterogeneity in microsatellite instability status at single cell resolution

Harrison Anthony, Cathal Seoighe

PII: S2589-0042(26)00235-X

DOI: <https://doi.org/10.1016/j.isci.2026.114860>

Reference: ISCI 114860

To appear in: *iScience*

Received Date: 17 July 2025

Revised Date: 4 November 2025

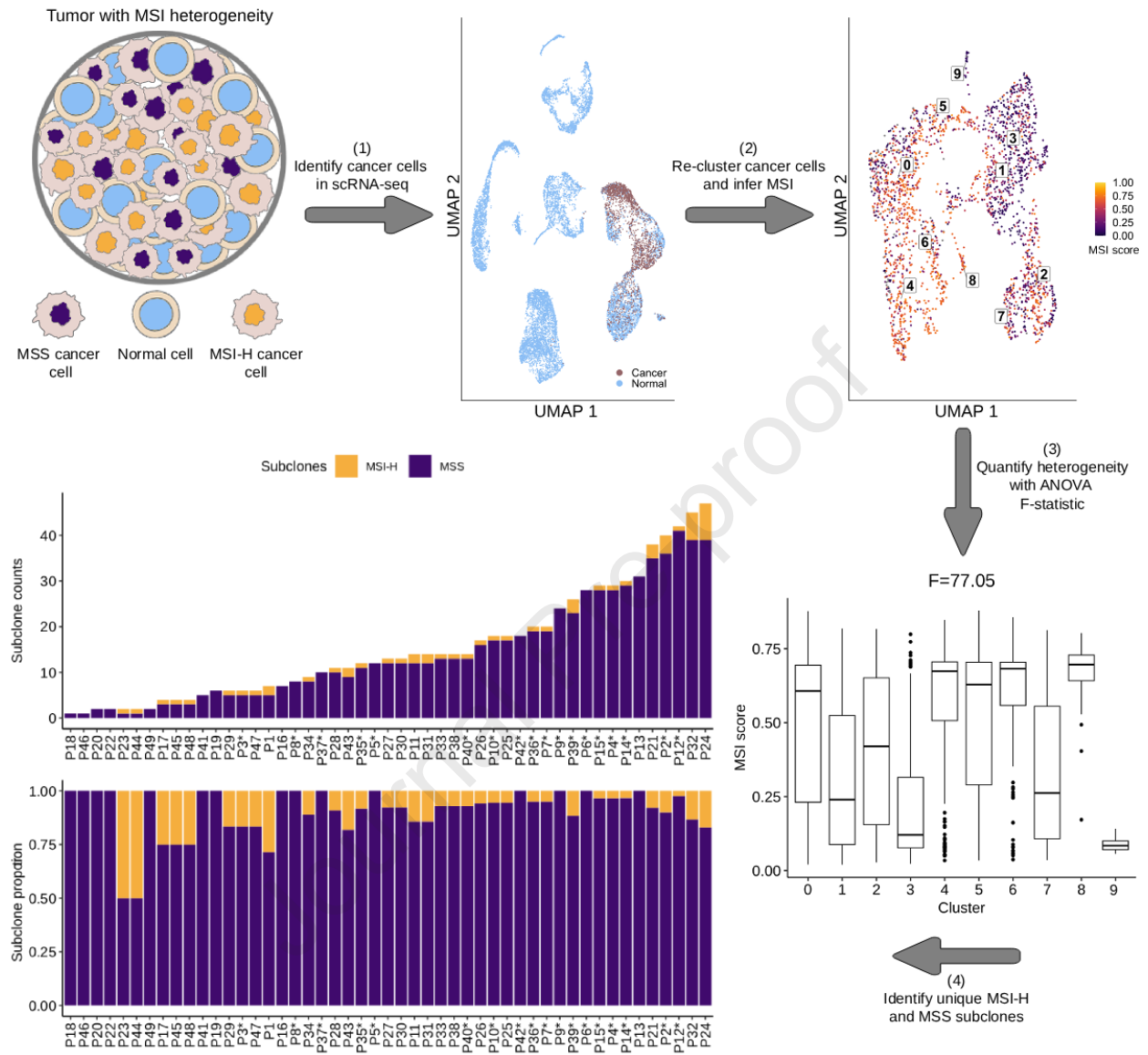
Accepted Date: 28 January 2026

Please cite this article as: Anthony, H., Seoighe, C., Intratumoral heterogeneity in microsatellite instability status at single cell resolution, *iScience* (2026), doi: <https://doi.org/10.1016/j.isci.2026.114860>.

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: <https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article>. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2026 The Author(s). Published by Elsevier Inc.

Single-cell resolution of heterogeneity in MSI



1 Title: Intratumoral heterogeneity in microsatellite instability status at single cell
2 resolution

3 Authors and affiliations: Harrison Anthony^{1,2} and Cathal Seoighe^{*1,2}

4 ¹ School of Mathematical and Statistical Sciences, University of Galway, Galway, Ireland

5 ² The Research Ireland Centre for Research Training in Genomics Data Science, Galway,
6 Ireland

7
8 *Lead contact: Cathal Seoighe (cathal.seoighe@universityofgalway.ie), School of
9 Mathematical and Statistical Sciences, University Road, University of Galway, Ireland.

10

11

12

13

14

15

16

17

18

19

20

21 **SUMMARY**

22 Intratumoral heterogeneity complicates the interpretation of single-test biomarkers.
23 Microsatellite instability (MSI) is one such biomarker, which is used to guide immune
24 checkpoint inhibitor treatment by classifying samples as having high microsatellite instability
25 (MSI-H) or as microsatellite stable (MSS). However, it is unknown whether MSI itself is a
26 heterogeneous phenomenon. To test this, we curated data from several single-cell RNA
27 sequencing studies with clinical MSI status and developed a computational pipeline that
28 quantifies intratumoral heterogeneity in MSI. Out of 49 individuals, 15 showed evidence of
29 divergence in MSI status between clusters of cancer cells and most had distinct MSI-H and MSS
30 subclones. These results question the use of MSI as a binary biomarker, and we hypothesize
31 that accounting for heterogeneity could improve its use as a predictive biomarker. Further
32 studies are required to determine the frequency of MSI heterogeneity at the population level
33 and if it can have clinical implications.

34 **KEYWORDS**

35 Intratumoral heterogeneity, Microsatellite instability, Single-cell analysis, Subclonal diversity,
36 Computational pipeline, Snakemake, Predictive biomarkers, Colorectal cancer

37 **INTRODUCTION**

38 Subclonal diversity within a tumor is a critical consideration in cancer research and treatment.
39 The overall diversity found in a single neoplasm is called intratumoral heterogeneity (ITH).
40 While ITH was first conceptualized to be genetic in nature¹, it is now used to describe genetic,
41 epigenetic, and phenotypic differences between subclones². The diversity within a tumor is
42 important because ITH has been linked to poor patient outcomes, therapy resistance, and
43 relapse^{3,4}. Furthermore, biomarkers that rely on single-sample tests can be susceptible to
44 sampling bias when ITH is present^{5,6}. While its origins are still debated⁷, one well known driver
45 of ITH is genome instability⁸.

46 Genome instability is a hallmark of cancer, characterized by a higher rate of accumulation of
47 mutations during replication, typically due to deficiencies in DNA repair genes⁹. The two most
48 common forms of genomic instability are at the chromosomal level, where instability is

49 characterized by aneuploidy and chromosomal aberrations¹⁰, and at the microsatellite level,
50 where short tandem repeats expand and contract in a mutator phenotype manner¹¹. The latter,
51 referred to as microsatellite instability (MSI), is hypothesized to be the result of a deficient
52 mismatch repair (dMMR) pathway and is commonly used as a biomarker to help guide immune
53 checkpoint inhibitor (ICI) treatment. This is done by classifying cancers as either having high
54 microsatellite instability (MSI-H) or as being microsatellite stable (MSS)¹². The classification is
55 normally carried out using a single-sample test that compares five microsatellite markers
56 between a tumor and paired-normal sample^{13,14}. While the interplay between chromosomal
57 instability and ITH is well defined and explored¹⁵⁻¹⁷, the relationship between MSI and ITH is
58 less clear.

59 Up to this point, most research on MSI and ITH has been framed around how MSI can impact
60 and shape the variation present within a tumor. Most studies in this area have focused on
61 specific mutations^{18,19} and the immune cell types present in the tumor microenvironment
62 (TME)^{20,21}, with the latter being crucial to precision medicine efforts. Researchers have shown
63 that MSI-H cancers have a “hot” microenvironment with an abundance of tumor infiltrating
64 lymphocytes²², and that they respond well to anti-PD-1 therapy which prevents T-cell
65 exhaustion²³. Despite the successes of using MSI status to guide anti-PD-1 therapy, there are
66 still challenges to its adoption as a predictive biomarker.

67 Some issues remain with the use of MSI status as a predictive biomarker as researchers have
68 reported cases of low treatment response rates and intrinsic treatment resistance when using
69 MSI status to guide anti-PD-1 therapy^{22,24,25}. One possible explanation for this is ITH as it has
70 been linked to therapy resistance^{3,4} and is known to complicate the interpretation of clinical
71 biomarkers^{5,6}. While researchers have studied cell types present in the TME, little is known
72 about whether MSI itself is a heterogeneous phenomenon. Although some case studies exist
73 that documented cases of ITH in MSI status²⁶⁻³¹, the question of whether MSI itself is
74 frequently a heterogeneous phenomenon, with some subclones displaying MSI while others do
75 not, has yet to be examined in detail. This warrants further investigation as it may ultimately,
76 lead to improved biomarker performance.

77 The current literature suggests that subclonality of MSI status is relatively rare or entirely
78 absent^{32,33}, but that is not always the case. There are many examples of individuals not only
79 with discordant MSI statuses between the primary tumor site and metastases^{26–29} but also
80 between multiple sites in the primary tumor^{30,31}. While these are small case studies, they
81 provide anecdotal evidence for cancers comprising MSI-H and MSS subclones. However, there
82 has, as yet, been no attempt to evaluate the frequency with which this occurs. Furthermore, a
83 detailed examination of heterogeneity requires an assessment of MSI at the single-cell level
84 with next-generation sequencing, not with the traditional methods of PCR and IHC used in
85 these case studies, as these are limited to detecting clear spatial heterogeneity.

86 Here we aimed to address these gaps through an analysis of published single-cell datasets that
87 include paired clinical MSI status. To do this, we developed a custom Snakemake³⁴ pipeline that
88 identifies MSI-H cells and uses novel methods to assess levels of heterogeneity and have made
89 this pipeline available as an open-source, scalable resource to the scientific community. We
90 evaluated the pipeline by mixing varying numbers of MSI-H and MSS cells from different
91 samples. Applying this framework, we show evidence of heterogeneity in MSI status at the
92 single-cell level and estimate its prevalence in the curated data. We also examine the nature of
93 MSI heterogeneity through a detailed investigation of single-cell data from two individuals –
94 one classified as MSI-H and the other MSS through PCR/IHC tests.

95 **RESULTS**

96 **Computational pipeline distinguishes MSI-H and MSS individuals and captures ITH**

97 To determine whether MSIsensor-RNA could distinguish between MSI-H and MSS individuals,
98 we ran it on the aggregate expression of all cells and again on only the cancer cells for each
99 individual. As expected, MSI-H individuals generally had higher MSI scores than MSS individuals,
100 and MSIsensor-RNA was able to broadly distinguish between the two groups (**Error! Reference**
101 **source not found.**A,B and Figure S1A,B). These results were seen for both aggregated
102 expression of all cells and only cancer cells, but subsetting down to only cancer cells yielded
103 lower MSI scores. There were also disagreements between PCR/IHC MSI status and MSIsensor-

104 RNA score with several MSS individuals having relatively high MSI scores and several MSI-H
105 individuals having low MSI scores (**Error! Reference source not found.A,B**).

106 Next, we simulated different levels of heterogeneity to determine how well our pipeline
107 captured ITH in MSI status. For this purpose, we simulated different levels of heterogeneity,
108 ranging progressively from pure MSS cells to pure MSI-H cells by mixing together samples from
109 two individuals comprised of homogeneously MSI-H cancer cells and homogeneously MSS
110 cancer cells (Tables S1 and S2). As expected, the more homogeneous samples (MSS, mix M1,
111 mix M9 and MSI-H in **Error! Reference source not found.C**) had low F-statistic values while
112 mixtures with more equal proportions of MSI-H and MSS cells (mixes M3-M5 in **Error!**
113 **Reference source not found.C**) had high F-statistic values. Increasing the proportion of MSI-H
114 cells until mix M7 resulted in an overall reduction in the number of MSS subclones identified
115 (**Error! Reference source not found.D** and Tables S1 and S2), and there was one MSI-H subclone
116 that was consistently detected after the proportion of MSI-H cells was 0.1 (mixes M1-M9).
117 Together, these results show that the F-statistic is sensitive to ITH, and that the number of
118 subclones can be consistently identified across replicates, providing useful context to the
119 heterogeneity.

120 **MSI-H and MSS individuals have evidence of ITH in MSI status**

121 In order to assess heterogeneity in MSI status, we first calculated F-statistics (see Materials and
122 Methods) based on clusters of cancer cells and identified subclones based on CNV patterns. We
123 found that MSI-H and MSS individuals both had evidence of heterogeneity in MSI. In total, 15 of
124 49 individuals showed evidence of divergence in MSI status between distinct clusters of cancer
125 cells ($F > 25$; **Error! Reference source not found.** and **Error! Reference source not found.**).

126 Several individuals had very large estimates of heterogeneity based on F-statistics (75.20 to
127 116.10) with most of these individuals being originally deemed to be MSS, and one originally
128 deemed to be MSI-H from a PCR or IHC test. In contrast, the lowest F-statistics (1.30 to 1.68)
129 were found in MSI-H and MSS individuals, and the ANOVA tests were not statistically significant
130 in either case ($P > 0.05$; Table S3). This was also seen in most other individuals with fewer than
131 three cancer cell clusters (**Error! Reference source not found.** and Table S3).

132 In general, MSI-H and MSS individuals had similar distributions of F-statistic values but with
133 several outliers having large F-statistic values among the MSS individuals (Figure 2A,B).
134 Interestingly, nearly every individual in the analysis had both MSI-H and MSS subclones, and a
135 larger proportion of MSS subclones (**Error! Reference source not found.**A,B and Table 2). The
136 exceptions were two individuals who each had a subclone proportion of 0.5 (**Error! Reference
137 source not found.**B), but they had very few cancer cells and too few cancer clusters to calculate
138 an F-statistic for comparison (Table S3). Those with the most MSI-H subclones, six and eight,
139 were originally determined to be MSI-H, but one MSS individual also had four MSI-H subclones.
140 Independent of MSI status, the distribution in number of clusters across individuals was
141 relatively even, and most individuals had two or fewer samples used in the clustering process
142 (**Error! Reference source not found.**C,D).

143 **Single-cell level resolution of heterogeneity in one MSI-H and one MSS individual**

144 We selected two individuals (P24 and CRC2786) with relatively high F-statistics and many MSI-H
145 subclones to illustrate the heterogeneity in MSI that is evident from single cell RNA-Seq data
146 (**Error! Reference source not found.** and **Error! Reference source not found.**). The MSI-H
147 individual, P24, had good overlap in cells classified as cancer with scATOMIC (**Error! Reference
148 source not found.**A) and those with high MSI scores determined by MSIsensor-RNA (**Error!
149 Reference source not found.**B). The re-clustered cancer cells appear to cluster by MSI score;
150 notably, clusters two and three (**Error! Reference source not found.**C). Those larger differences
151 in the clusters were also seen in the pseudobulk analysis of the re-clustered cancer cells (**Error!
152 Reference source not found.**D).

153 The MSS individual, CRC2786, also had good overlap between the cells determined to be
154 cancerous and those with a high MSI score; however, there was less separation of cancer and
155 normal cells in this individual (**Error! Reference source not found.**A,B). Similarly, in the re-
156 clustered cancer cells, cells with higher and lower MSI scores were somewhat more
157 intermingled than for the MSI-H individual, although clusters three and four seem to be
158 predominantly MSS and MSI-H, respectively (**Error! Reference source not found.**C). This result

159 is recapitulated in the pseudobulk analysis with cluster three showing a low MSI score and
160 cluster four a much higher one (**Error! Reference source not found.D**).

161 **Significant differences in MSI score and gene expression between clusters of cancer cells**

162 We examined MSI ITH in CRC2786 and P24 further by assessing differences between clusters of
163 cancer cells. We found both individuals to have many clusters with significantly different MSI
164 scores (**Error! Reference source not found.**), and several genes showed differences in
165 expression between clusters and between cells identified as MSI-H and MSS (Figures S2 and S3).
166 Both individuals had clusters with high and low MSI scores (**Error! Reference source not
167 found.A,B**). These differences were found to be statistically significant ($P < 0.05$ using a Tukey
168 HSD test; Tables S4 and S5). We found that 35 cluster pairs for CRC2786 had significantly
169 different MSI scores and 17 cluster pairs for P24 (**Error! Reference source not found.C,D**) were
170 significantly different.

171 Within the clusters of cancer cells gene expression was also significantly different between
172 clusters (Figures S2A and S3A; Tables S6 and S7), and between the MSI-H and MSS cells in those
173 clusters (Figures S2B and S3B; Tables S8 and S9). The top five differentially expressed genes for
174 each cluster of cancer cells for each individual were retained for analysis as well as the top 50
175 differentially expressed genes between MSI-H and MSS cells. Individual CRC2786 had three
176 genes: *MALAT1*, *EEF1A1*, and *SH3BGRL3* in common between those differentially expressed
177 between clusters and between cells with different MSI status; Figure S2A,B). P24, on the other
178 hand had one gene, *PCLAF* that was differentially expressed between clusters and between
179 MSI-H and MSS cells. When comparing the differential expression analyses for both individuals,
180 we found three genes: *BMX*, *LRMP*, and *SH2D6* that were differentially expressed between
181 clusters and two genes (*TYMS*, *OXCT1*) in common that differentiated MSI-H and MSS cells.

182 **DISCUSSION**

183 In our study, we showed that MSI status can be heterogeneous at the single-cell level and
184 provide a pipeline to measure that heterogeneity with the clustering of cancer cells and CNV
185 based subclone analysis. These results contrast with the assumption that is commonly made,

186 both in research and in clinical practice, that MSI is dichotomous. While this assumption has
187 proven useful, enabling MSI-H to be used as a biomarker for (ICI) treatment³⁵, overall responder
188 rate has been reported to be as low as 31%²². We hypothesize that this could be explained, at
189 least in part, by the heterogeneity in MSI-H individuals, which a binary classification fails to take
190 into account.

191 Single-sample tests (like the ones used to assign MSI status) are susceptible to under-sampling
192 bias when spatial ITH is present³. This would require multi-sample, multi-regional tests to
193 improve classification as MSI-H and MSS cells could be concentrated in different regions of the
194 tumor. As we did not have access to spatial data we could not resolve whether the
195 heterogeneity we identified was organized spatially. However, there is some evidence that this
196 could be the case. One previous study found that MSI-H colorectal cancers had spatial ITH of
197 specific frameshift mutations in several mononucleotide repeats, but this regional ITH was
198 determined to be the result of disease progression and not related to an MSI-H phenotype³⁶.
199 Similarly, another study also found discrepancies between IHC and PCR tests when sampling
200 different regions of MSI-H cancers, suggesting that a multi-sample biopsy would be more
201 appropriate when testing for MSI²⁶.

202 Both studies, however, used PCR or IHC tests and likely missed the level of heterogeneity we
203 discovered by analyzing single-cell data. A future study with spatially resolved single-cell
204 transcriptomics³⁷ would provide an opportunity to determine if ITH in MSI is structured
205 regionally and whether this has any clinical relevance. For example, if we found that MSI-H and
206 MSS cells were regionally clustered and separated from one another, this would suggest that
207 the single-region, single-sample biopsy technique currently in use to assess MSI clinically is
208 inadequate

209 One of our primary findings, that both MSI-H and MSS individuals had a mixture of MSI-H and
210 MSS cells is supported by the findings of another study³⁸. Similarly to our study, they also found
211 that MSI-H and MSS individuals had evidence of ITH in MSI using single-cell sequencing data;
212 however, their methodology, which involved clustering cells based on gene-set enrichment of
213 MSI-H and MSS signatures, did not identify any MSS individuals with only MSS cells. Our

214 pipeline was able to find examples of MSS individuals comprising MSS cells only, which would
215 make sense given that MSI is a relatively rare trait, and it would be unlikely to be present in
216 every MSS individual in a study cohort. This is due to the main difference between our
217 methods, as we test individual cells for microsatellite instability, whereas Zhao et al.³⁸ labelled
218 cells as MSI-H at the pseudo-bulk level with gene-set enrichment guided cell clustering. Our
219 study is also unique as we aimed to quantify ITH and provide our pipeline in an open access
220 format.

221 Our finding that nearly every MSI-H and MSS individual had MSI-H and MSS subclones has not
222 yet been reported in other studies; however, two case studies that infer subclonality of dMMR
223 status from discordant IHC test results have been reported^{27,39}. Even though dMMR and MSI-H
224 technically refer to different phenomena, MSI is considered to be the byproduct of dMMR and
225 both are predictive of (ICI) treatment efficacy. Combined with our findings, these case studies
226 provide insights that could help explain reports of 30% or more of MSI-H cancers having
227 primary resistance to single-agent ICI's^{24,40}.

228 A treatment regime for an MSI-H cancer would potentially miss one or more MSS subclones,
229 leaving behind a population of cells that would not respond in the same way to
230 immunotherapy. This is because a tumor with coexisting MSI-H and MSS subclones would have
231 a different type of TME shaped by immune cells and the PD-1/PD-L1 pathway. The key
232 difference is that a tumor consisting mostly of MSI-H subclones will have a higher neoantigen
233 load due to an abundance of frameshift mutations^{41,42}, and consequently, a "hotter" TME
234 characterized by many tumor infiltrating lymphocytes²². In turn, this leads to an upregulation of
235 PD-L1 in tumor and stromal cells in the TME causing T-cell dysfunction²³, and allows cancer cells
236 to escape immunosurveillance⁴³. Investigation of the impact of MSI heterogeneity on treatment
237 and the TME would benefit from longitudinal data consisting of tumor samples before and at
238 several time points after treatment from individuals with and without heterogeneity in MSI.
239 This type of study would be warranted by our results as we provide a plausible mechanism for
240 treatment resistance which is not currently given adequate consideration⁴⁰.

241 Our computational pipeline is the first to identify and quantify heterogeneity in MSI status at
242 the single-cell level. We built the pipeline around MSIsensor-RNA and scATOMIC, two pan-
243 cancer, machine learning based approaches. The combination of these programs may give rise
244 to some potential issues. Naturally, as both approaches are trained on gene expression data,
245 there will be overlap in genes used to train both classifiers and consequently overlap in cell type
246 prediction. Yet, we found different genes to be differentially expressed between cancer cell
247 clusters and MSI-H and MSS cells. This is likely because there is no overlap in training data
248 between the two tools. One other factor to consider is that we found a loss of microsatellite
249 instability signal in MSI-H individuals after subsetting down to the cancer cells. Despite being
250 necessary at the single-cell level to only label cells as MSI-H if they were also determined to be
251 cancerous by scATOMIC, there were likely instances where MSIsensor-RNA correctly identified
252 MSI-H cells and scATOMIC did not. Going forward, it would be beneficial for a benchmarking
253 study to be done to determine if MSIsensor-RNA could also better identify cancer cells in MSI-H
254 individuals. Another factor to consider is that there can be an overlap between the genes used
255 in clustering of cells and the genes used to generate an MSI score. Whether one or more of the
256 100 genes used in the MSIsensor-RNA baseline are included in the 2,000 most variable genes
257 used in clustering steps of pipeline will change from individual to individual. While not included
258 in this study, we have checked clustering of cancer cells with and without the 100 genes used
259 by MSIsensor-RNA and found it did not appear to affect the clustering results.

260 While other MSI detection tools exist, we chose to use MSIsensor-RNA which infers MSI status
261 based on gene expression as we were using single-cell RNA sequencing data. MSI is typically
262 detected in NGS data by comparing the distribution of indels in microsatellites between a
263 paired-normal and tumor sample. However, the tool we used, MSIsensor-RNA, does not
264 directly detect MSI with microsatellites but instead, uses machine learning models trained on
265 gene expression patterns from MSI-H and MSS individuals. This technique is better suited to
266 detect dMMR, which is traditionally measured with IHC staining of genes involved in the
267 mismatch repair pathway. Furthermore, we have shown in a previous study that RNA-based
268 detection methods have a lower performance than DNA-based detection methods on bulk
269 sequencing data⁴⁴. However, the authors of MSIsensor-RNA report high performance on single-

270 cell RNA sequencing data⁴⁵, and we found that it could broadly distinguish between the
271 individuals deemed MSI-H and MSS with PCR/IHC tests in our dataset (Figure 1A,B and Figure
272 S1A,B). Based on these factors, it would be worthwhile to reproduce our results with data
273 generated from other single-cell sequencing technologies such as whole-genome amplification
274 and sequencing which would permit the use of more well-known and established NGS tools
275 that measure differences in microsatellite repeats, such as MANTIS and MSIsensor^{46,47}.

276

277 Altogether, we found that heterogeneity in microsatellite instability is more common than
278 previously reported, and we found it both in MSI-H and MSS individuals. These results could
279 help to explain why there are reports of treatment resistance and low response rates in MSI-H
280 cancers treated with ICI therapy; however, our study only analyzed single-cell RNA sequencing
281 data from 49 individuals that underwent 3' and 5' single-cell RNA sequencing. Further studies
282 are warranted to determine the frequency of heterogeneity in this biomarker at the population
283 level and whether the presence of MSI-H and MSS subclones can have clinical impacts,
284 including the capacity for rapid evolution of resistance to treatments for which MSI-H is used as
285 a biomarker.

286

287 **Limitations of the study**

288 The primary limitation of this study is the relatively small number of individuals that had
289 publicly available single-cell sequencing data with paired clinical MSI status. Although our study
290 consisted of 134 single-cell RNA sequencing samples, these were from only 49 distinct
291 individuals. This limited our ability to assess the frequency of heterogeneity in this biomarker in
292 the general population. In order to better gauge the frequency of heterogeneity in MSI status at
293 the population level, the results found in our study would need to be replicated in a large
294 cohort-based study. Additionally, our study did not have sufficient clinical metadata to establish
295 whether ITH in MSI has clinical implications, and we therefore did not aim to address this
296 subject. Although some individuals in our study did receive ICI treatment, there was no
297 information on when the sample was collected or when treatment had been administered.

298 Future work in this area would need to include such metadata in order to unravel whether
299 individuals respond differently to treatment if they have a tumor with MSI-H and MSS
300 subclones.

301 **RESOURCE AVAILABILITY**

302 **Lead contact**

303 Requests for further information and should be directed to the lead contact, Cathal Seoighe
304 (cathal.seoighe@universityofgalway.ie).

305 **Materials availability**

306 This study did not generate new unique reagents.

307 **Data and code availability**

- 308 • Data: This paper analyzes existing, publicly available data, accessible from either the
309 European Genome-Phenome Archive, the Sequence Read Archive, or from the Gene
310 Expression Omnibus (Key Resources Table) .
- 311 • Code: All original code and results have been deposited to Zenodo (Key Resources Table)
312 A distributable version of the computational pipeline used in this study, SINGLE-MSI, is
313 also available via Zenodo (Key Resources Table). We have written the entire workflow in
314 Snakemake to ensure reproducibility and scalability.
- 315 • Other: Any additional information required to analyze the data reported in this paper is
316 available from the lead contact upon request.

317

318 **ACKNOWLEDGEMENTS**

319 We would like to thank the patients and researchers who made this study possible with the
320 sharing of their data. This includes patients from the CRC-SG1, KUL3 and KUL5 cohorts in
321 Joanito et al., patients involved in PICC study (NCT03926338) from Li et al., and the 6 individuals
322 from Yunnan Cancer Hospital from Wu et al. We would also like to thank Micheál Ó Dálaigh for
323 useful conversations on navigating single-cell cancer data and Anna Großbach for advice on
324 figure design. This research was funded by Research Ireland through the Research Ireland

325 Centre for Research Training in Genomics Data Science under Grant number 18/CRT/6214.
326 Lastly, we would like to thank Dr. Eleanor Jayawant for making the original cell figures used in
327 the first panel of our graphical abstract available through the following license: CC-BY
328 4.0 Unported <https://creativecommons.org/licenses/by/4.0/>.

329 **DECLARATION OF INTERESTS**

330 The authors declare no competing interests.

331

332 **AUTHOR CONTRIBUTIONS**

333 Conceptualization, H.A. and C.S.; methodology, H.A and C.S.; Investigation, H.A.; writing—
334 original draft, H.A.; writing—review & editing, H.A. and C.S.; funding acquisition, C.S.; resources,
335 C.S.; supervision, C.S.

336 **SUPPLEMENTAL INFORMATION**

337 Document S1. Figures S1-S3, Tables S1, S3-S9

338 Table S2. Mixing experiment raw results, related to Figure 1.

339

340 **FIGURE TITLES AND LEGENDS**

341 Figure 1. MSIsensor-RNA and simulation results.

342 Box plots showing the distribution of (A) MSI score for individuals calculated using the
343 aggregate expression of all cells, and (B) MSI score for individuals calculated using the
344 aggregate expression of only cancer cells. Also shown are the mean values of (C) the F-statistic
345 and (D) the number of subclones for the different cell mixes shown on the x-axes (with
346 increasing proportions of MSI-H cells ranging from 0.1 in mix M1 to 0.9 in mix M9). The error
347 bars in (C) and (D) correspond to plus/minus twice the standard error around the mean. The
348 MSS and MSI-H samples in panels C and D are the obtained values for all cells in those samples
349 and do not represent an average. See Figure S1 showing ROC and precision-recall curves for
350 MSIsensor-RNA as well as Tables S1 and S2 which contain summary statistics and the raw
351 results for the mixing experiments, respectively.

352 Figure 2. Distributions of summary statistics in single-cell RNA sequencing data.

353 Box plots showing the distribution of (A) F-statistics grouped by PCR/IHC MSI status, (B) the
354 proportion of MSI-H to MSS cells grouped by PCR/IHC MSI status. Also shown are histograms
355 displaying the frequency of (C) the number of cancer cell clusters and (D) the number of tumor
356 samples for all individuals. See Table S3 which includes the ANOVA test results for each
357 individual.

358 Figure 3. MSI-H and MSS subclone compositions.

359 Stacked bar plots of (A) the number of subclones for each individual in the analysis and (B) the
360 proportion of subclone types for each individual. Individuals that had a PCR/IHC test result of
361 MSS are indicated with an asterisk.

362 Figure 4. Clustering of cells for MSI-H individual.

363 UMAP plots for MSI-H individual P24 showing (A) tumor versus normal cell classification, (B)
364 MSI scores for each cell, (C) MSI scores for re-clustered cancer cells, and (D) MSI score for
365 aggregated pseudobulk expression of each cancer cell cluster. Any grey colors indicate an NA
366 value. See also Tables S4, S6, and S8 which contain the results of Tukey HSD analysis between
367 all cancer cell clusters, differential gene expression analysis between all cancer cell clusters, and
368 differential gene expression analysis between MSI-H and MSS cells, respectively.

369

370 Figure 5. Clustering of cells for MSS individual.

371 UMAP plots for MSS individual CRC2786 showing (A) tumor versus normal cell classification, (B)
372 MSI scores for each cell, (C) MSI scores for re-clustered cancer cells, and (D) MSI score for
373 aggregated pseudobulk expression of each cancer cell cluster. Any grey colors indicate an NA
374 value. See also Tables S5, S7, and S9 which contain the results of Tukey HSD analysis between
375 all cancer cell clusters, differential gene expression analysis between all cancer cell clusters, and
376 differential gene expression analysis between MSI-H and MSS cells, respectively.

377

378 Figure 6. Distribution of MSI scores and difference in means for cancer cell clusters.

379 Box plots showing the distribution of MSI scores for each cluster of cancer cells in (A) individual
380 CRC2786 and (B) individual P24. Also shown are the 95% confidence intervals for the difference
381 in mean MSI scores between each cluster pair for (C) individual CRC2786 and (D) individual P24.

382 TABLES AND TEXT BOXES

383 Table 1. Individual summary statistics and subclone information

Individual	Clusters	F	Samples	MSI-H cells	MSS cells	MSS	MSI-H	PCR/IHC
CRC2783	4	17.15	3	40	211	5	2	MSI-H
CRC2786	10	75.67	4	101	2182	36	4	MSS
CRC2787	2	1.62	2	4	121	5	1	MSS
CRC2794	6	10.23	4	3	939	28	1	MSS
CRC2795	6	8.17	4	0	367	7	0	MSS
CRC2801	7	4.8	5	1	1019	7	0	MSS
CRC2803	8	16.04	4	5	619	19	1	MSS
CRC2810	3	1.68	4	1	155	7	0	MSS
CRC2811	8	19.63	4	1	1255	7	0	MSS
CRC2816	7	8.21	4	2	585	17	1	MSS
CRC2817	7	10.62	10	56	442	12	2	MSI-H
CRC2821	11	100.85	9	240	9698	41	1	MSS
CRC2829	9	10.04	2	0	1147	7	0	Unknown
CRC2841	8	93.35	11	5	1392	29	1	MSS
CRC2899	9	20.28	4	3	1444	28	1	MSS
P11	3	6.94	1	0	119	7	0	MSI-H
P12	3	3.02	1	11	92	3	1	MSI-H
P14	1	NA	1	1	20	7	0	MSI-H
P15	3	2.25	1	0	139	7	0	MSI-H
P17	1	NA	1	0	36	7	0	MSI-H
P18	10	30.89	1	58	1633	35	3	MSI-H
P19	2	1.3	1	0	41	7	0	MSI-H
P21	1	NA	1	3	27	1	1	MSI-H
P23	10	29.07	1	200	1785	39	8	MSI-H
P24	7	75.2	1	22	732	17	1	MSI-H
P25	7	34.94	2	13	630	16	1	MSI-H
P26	6	5.09	1	12	427	12	1	MSI-H
P27	5	11.26	2	4	385	10	1	MSI-H
P28	5	51.15	1	4	192	5	1	MSI-H
P29	5	10.99	1	3	380	12	1	MSI-H
P30	6	26.69	1	62	434	12	2	MSI-H
P31	10	18.71	2	149	1833	39	6	MSI-H
P32	7	19.41	2	3	370	13	1	MSI-H
P33	6	15.71	1	24	301	8	1	MSI-H
SC024	5	21.77	2	4	338	11	1	MSS
SC027	8	20.28	2	3	769	19	1	MSS

SC029	4	5.78	2	0	300	7	0	MSS
SC035	6	47.22	2	37	415	13	1	MSI-H
SC040	9	116.1	4	82	1067	23	3	MSS
SC041	5	10.01	2	32	279	13	1	MSS
SC042	3	31.24	2	0	141	7	0	Unknown
SC043	7	6.75	2	0	813	7	0	MSS
SC044	8	31.9	3	58	412	9	2	MSI-H
SRR23490337	1	NA	1	3	31	1	1	MSI-H
SRR23490338	2	7.66	1	11	70	3	1	MSI-H
SRR23490339	1	NA	1	0	18	7	0	MSI-H
SRR23490340	4	27.42	1	24	190	5	1	MSI-H
SRR23490341	3	48.02	1	14	74	3	1	MSI-H
SRR23490342	1	NA	1	0	48	7	0	MSI-H

384 This table contains the summary statistics for each individual included in the analysis. The
385 clusters column refers to the number of unique cancer clusters, and F is the ANOVA F-statistic
386 used to measure heterogeneity in those clusters. The samples column describes the number of
387 samples each individual had for the analysis, and the MSI-H and MSS cells column is the number
388 of cell type for that individual. The MSS and MSI-H columns refer to the number of
389 microsatellite stable and microsatellite instability high subclones for each individual. The
390 original IHC/PCR status for each individual is included and was established in previous studies
391 (Table 2). Any NA values represent F-statistics that could not be calculated due to fewer than 2
392 clusters of cancer cells being present.

393 Table 2. Single-cell sequencing datasets.

Dataset ID	Cancer type	Individuals	Samples	Sequencing
EGAD00001008555	Colorectal/metastatic	15	77	Illumina HiSeq 4000
EGAD00001008584	Colorectal/metastatic	3	6	Illumina HiSeq 4000
EGAD00001008585	Colorectal/metastatic	6	18	Illumina NextSeq 500/NovaSeq 6000
GSE205506	Colorectal	19	27	Illumina NovaSeq 6000/DNBSEQ- T

PRJNA932556	Colorectal	6	6	HiSeq X Ten
-------------	------------	---	---	-------------

394

395 A table detailing each dataset used in the study. The dataset column specifies study cohorts,
396 Genome Expression Omnibus ID (GSE prefix), or SRA project code (PRJNA prefix).

397 **REFERENCES**

- 398 1. Nowell PC. The clonal evolution of tumor cell populations. *Science (80-)*.
399 1976;194(4260):23-28. doi:10.1126/science.959840
- 400 2. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: A looking glass for
401 cancer? *Nat Rev Cancer*. 2012;12(5):323-334. doi:10.1038/nrc3261
- 402 3. Marusyk A, Janiszewska M, Polyak K. Intratumor Heterogeneity: The Rosetta Stone of
403 Therapy Resistance. *Cancer Cell*. 2020;37(4):471-484. doi:10.1016/j.ccell.2020.03.007
- 404 4. Qazi MA, Vora P, Venugopal C, et al. Intratumoral heterogeneity: Pathways to treatment
405 resistance and relapse in human glioblastoma. *Ann Oncol*. 2017;28(7):1448-1456.
406 doi:10.1093/annonc/mdx169
- 407 5. Gilson P, Merlin JL, Harlé A. Deciphering Tumour Heterogeneity: From Tissue to Liquid
408 Biopsy. *Cancers (Basel)*. 2022;14(6). doi:10.3390/cancers14061384
- 409 6. McGranahan N, Swanton C. Biological and Therapeutic Impact of Intratumor
410 Heterogeneity in Cancer Evolution. *Cancer Cell*. 2015;28(1):141.
411 doi:10.1016/j.ccell.2015.06.007
- 412 7. Sun R, Hu Z, Curtis C. Big bang tumor growth and clonal evolution. *Cold Spring Harb
413 Perspect Med*. 2018;8(5):1-14. doi:10.1101/cshperspect.a028381
- 414 8. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic
415 heterogeneity in cancer evolution. *Nature*. 2013;501(7467):338-345.
416 doi:10.1038/nature12625
- 417 9. Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability an evolving hallmark of
418 cancer. *Nat Rev Mol Cell Biol*. 2010;11(3):220-228. doi:10.1038/nrm2858
- 419 10. Thompson SL, Bakhoun SF, Compton DA. Mechanisms of Chromosomal Instability. *Curr
420 Biol*. 2010;20(6):R285-R295. doi:10.1016/j.cub.2010.01.034
- 421 11. Yamamoto H, Watanabe Y, Maehata T, Imai K, Itoh F. Microsatellite instability in cancer:
422 a novel landscape for diagnostic and therapeutic approach. *Arch Toxicol*.
423 2020;94(10):3349-3357. doi:10.1007/s00204-020-02833-z
- 424 12. Lee V, Murphy A, Le DT, Diaz LA. Mismatch Repair Deficiency and Response to Immune
425 Checkpoint Blockade. *Oncologist*. 2016;21(10):1200-1211.
426 doi:10.1634/theoncologist.2016-0046

- 427 13. Murphy KM, Zhang S, Geiger T, et al. Comparison of the microsatellite instability analysis
428 system and the Bethesda panel for the determination of microsatellite instability in
429 colorectal cancers. *J Mol Diagnostics*. 2006;8(3):305-311.
430 doi:10.2353/jmoldx.2006.050092
- 431 14. Berg KD, Glaser CL, Thompson RE, Hamilton SR, Griffin CA, Eshleman JR. Detection of
432 microsatellite instability by fluorescence multiplex polymerase chain reaction. *J Mol*
433 *Diagnostics*. 2000;2(1):20-28. doi:10.1016/S1525-1578(10)60611-3
- 434 15. Bakhoun SF, Landau DA. Chromosomal instability as a driver of tumor heterogeneity and
435 evolution. *Cold Spring Harb Perspect Med*. 2017;7(6):1-14.
436 doi:10.1101/cshperspect.a029611
- 437 16. Furuya T, Uchiyama T, Murakami T, et al. Relationship between chromosomal instability
438 and intratumoral regional DNA ploidy heterogeneity in primary gastric cancers. *Clin*
439 *Cancer Res*. 2000;6(7):2815-2820.
- 440 17. van den Bosch T, Derks S, Miedema DM. Chromosomal Instability, Selection and
441 Competition: Factors That Shape the Level of Karyotype Intra-Tumor Heterogeneity.
442 *Cancers (Basel)*. 2022;14(20):1-17. doi:10.3390/cancers14204986
- 443 18. Choi EJ, Kim MS, Song SY, Yoo NJ, Lee SH. Intratumoral Heterogeneity of Frameshift
444 Mutations in MECOM Gene is Frequent in Colorectal Cancers with High Microsatellite
445 Instability. *Pathol Oncol Res*. 2017;23(1):145-149. doi:10.1007/S12253-016-0112-
446 3/FIGURES/2
- 447 19. Jo YS, Kim MS, Yoo NJ, Lee SH. Somatic Mutations and Intratumoral Heterogeneity of
448 MYH11 Gene in Gastric and Colorectal Cancers. *Appl Immunohistochem Mol Morphol*.
449 2018;26(8):562-566. doi:10.1097/PAI.0000000000000484
- 450 20. Jung M, Lee JA, Yoo SY, Bae JM, Kang GH, Kim JH. Intratumoral spatial heterogeneity of
451 tumor-infiltrating lymphocytes is a significant factor for precisely stratifying prognostic
452 immune subgroups of microsatellite instability-high colorectal carcinomas. *Mod Pathol*.
453 2022;35(12):2011-2022. doi:10.1038/s41379-022-01137-0
- 454 21. Wu W, Liu Y, Zeng S, Han Y, Shen H. Intratumor heterogeneity: the hidden barrier to
455 immunotherapy against MSI tumors from the perspective of IFN- γ signaling and tumor-
456 infiltrating lymphocytes. *J Hematol Oncol*. 2021;14(1):1-28. doi:10.1186/s13045-021-
457 01166-3
- 458 22. Wang R, Lian J, Wang X, et al. Intrinsic resistance and efficacy of immunotherapy in
459 microsatellite instability-high colorectal cancer: A systematic review and meta-analysis.
460 *Biomol Biomed*. 2023;23(2):198. doi:10.17305/BJBMS.2022.8286
- 461 23. Chen L, Han X. Anti-PD-1/PD-L1 therapy of human cancer: Past, present, and future. *J Clin*
462 *Invest*. 2015;125(9):3384-3391. doi:10.1172/JCI80011

- 463 24. Heregger R, Huemer F, Steiner M, Gonzalez-Martinez A, Greil R, Weiss L. Unraveling
464 Resistance to Immunotherapy in MSI-High Colorectal Cancer. *Cancers (Basel)*.
465 2023;15(20):1-18. doi:10.3390/cancers15205090
- 466 25. Battaglin F, Naseem M, Lenz HJ, Salem ME. Microsatellite Instability in Colorectal Cancer:
467 Overview of Its Clinical Significance and Novel Perspectives. *Clin Adv Hematol Oncol*.
468 2018;16(11):735. Accessed April 2, 2025.
469 <https://pmc.ncbi.nlm.nih.gov/articles/PMC7493692/>
- 470 26. Chapusot C, Martin L, Bouvier AM, et al. Microsatellite instability and intratumoural
471 heterogeneity in 100 right-sided sporadic colon carcinomas. *Br J Cancer*. 2002;87(4):400-
472 404. doi:10.1038/sj.bjc.6600474
- 473 27. Evrard C, Messina S, Sefrioui D, et al. Heterogeneity of Mismatch Repair Status and
474 Microsatellite Instability between Primary Tumour and Metastasis and Its Implications
475 for Immunotherapy in Colorectal Cancers. *Int J Mol Sci*. 2022;23(8).
476 doi:10.3390/ijms23084427
- 477 28. Huang Q, Yu T, Li L, et al. Intraindividual Tumor Heterogeneity of Mismatch Repair Status
478 in Metastatic Colorectal Cancer. *Appl Immunohistochem Mol Morphol*. 2023;31(2):84-93.
479 doi:10.1097/PAI.0000000000001089
- 480 29. Luchini C, Mafficini A, Chatterjee D, et al. Histo-molecular characterization of pancreatic
481 cancer with microsatellite instability: intra-tumor heterogeneity, B2M inactivation, and
482 the importance of metastatic sites. *Virchows Arch*. 2022;480(6):1261-1268.
483 doi:10.1007/s00428-021-03205-3
- 484 30. Riedinger CJ, Esnakula A, Haight PJ, et al. Characterization of mismatch-
485 repair/microsatellite instability-discordant endometrial cancers. *Cancer*.
486 2024;130(3):385-399. doi:10.1002/cncr.35030
- 487 31. Tachon G, Frouin E, Karayan-Tapon L, et al. Heterogeneity of mismatch repair defect in
488 colorectal cancer and its implications in clinical practice. *Eur J Cancer*. 2018;95:112-116.
489 doi:10.1016/j.ejca.2018.01.087
- 490 32. Evrard C, Tachon G, Randrian V, Karayan-tapon L, Tougeron D. Microsatellite Instability:
491 Diagnosis, Heterogeneity, Discordance, and Clinical Impact in Colorectal Cancer Camille.
492 *Cancers (Basel)*. 2019;(Figure 1):1-25.
- 493 33. Georgiades IB, Curtis LJ, Morris RM, Bird CC, Wyllie AH. Heterogeneity studies identify a
494 subset of sporadic colorectal cancers without evidence for chromosomal or
495 microsatellite instability. *Oncogene*. 1999;18(56):7933-7940. doi:10.1038/sj.onc.1203368
- 496 34. Mölder F, Jablonski KP, Letcher B, et al. Sustainable data analysis with Snakemake.
497 *F1000Research*. 2021;10:33. doi:10.12688/f1000research.29032.2
- 498 35. Zhao P, Li L, Jiang X, Li Q. Mismatch repair deficiency/microsatellite instability-high as a

- 499 predictor for anti-PD-1/PD-L1 immunotherapy efficacy. *J Hematol Oncol.* 2019;12(1):1-
500 14. doi:10.1186/s13045-019-0738-1
- 501 36. Choi YJ, Kim MS, An CH, Yoo NJ, Lee SH. Regional Bias of Intratumoral Genetic
502 Heterogeneity of Nucleotide Repeats in Colon Cancers with Microsatellite Instability.
503 *Pathol Oncol Res.* 2014;20(4):965-971. doi:10.1007/s12253-014-9781-y
- 504 37. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to
505 elucidate intercellular tissue dynamics. *Nat Rev Genet.* 2021;22(10):627-644.
506 doi:10.1038/s41576-021-00370-8
- 507 38. Zhao F, Wang S, Bai Y, et al. Cellular MSI-H score: a robust predictive biomarker for
508 immunotherapy response and survival in gastrointestinal cancer. *Am J Cancer Res.*
509 2024;14(11):5551-5567. doi:10.62347/AIWP6518
- 510 39. Amemiya K, Hirotsu Y, Nagakubo Y, et al. Simple IHC reveals complex MMR alternations
511 than PCR assays: Validation by LCM and next-generation sequencing. *Cancer Med.*
512 2022;11(23):4479-4490. doi:10.1002/cam4.4832
- 513 40. Zhang Q, Li J, Shen L, Li Y, Wang X. Opportunities and challenges of immunotherapy for
514 dMMR/ MSI-H colorectal cancer. *Cancer Biol Med.* 2023;20(10):1-7.
515 doi:10.20892/j.issn.2095-3941.2023.0240
- 516 41. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and
517 endometrial cancer genomes. *Cell.* 2013;155(4):858. doi:10.1016/j.cell.2013.10.015
- 518 42. Turajlic S, Litchfield K, Xu H, et al. Insertion-and-deletion-derived tumour-specific
519 neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.*
520 2017;18(8):1009-1021. doi:10.1016/S1470-2045(17)30516-8
- 521 43. Cha JH, Chan LC, Li CW, Hsu JL, Hung MC. Mechanisms Controlling PD-L1 Expression in
522 Cancer. *Mol Cell.* 2019;76(3):359-370. doi:10.1016/j.molcel.2019.09.030
- 523 44. Anthony H, Seoighe C. Performance assessment of computational tools to detect
524 microsatellite instability. *Brief Bioinform.* 2024;25(5). doi:10.1093/bib/bbae390
- 525 45. Jia P, Yang X, Yang X, Wang T, Xu Y, Ye K. MSIsensor-RNA: Microsatellite Instability
526 Detection for Bulk and Single-cell Gene Expression Data. *Genomics Proteomics*
527 *Bioinformatics.* Published online 2024. doi:10.1093/gpbjnl/qzae004
- 528 46. Kautto EA, Bonneville R, Miya J, et al. Performance evaluation for rapid detection of pan-
529 cancer microsatellite instability with MANTIS. *Oncotarget.* 2017;8(5):7452-7463.
530 doi:10.18632/oncotarget.13918
- 531 47. Niu B, Ye K, Zhang Q, et al. MSIsensor: Microsatellite instability detection using paired
532 tumor-normal sequence data. *Bioinformatics.* 2014;30(7):1015-1016.
533 doi:10.1093/bioinformatics/btt755

- 534 48. Ashktorab H, Ahuja S, Kannan L, et al. A meta-analysis of MSI frequency and race in
535 colorectal cancer. *Oncotarget*. 2016;7(23):34546-34557. doi:10.18632/oncotarget.8945
- 536 49. Ashktorab H, Smoot DT, Carethers JM, et al. High incidence of microsatellite instability in
537 colorectal cancer from African Americans. *Clin Cancer Res*. 2003;9(3):1112-1117.
538 Accessed January 17, 2026.
539 [https://aacrjournals.org/clincancerres/article/9/3/1112/289190/High-Incidence-of-](https://aacrjournals.org/clincancerres/article/9/3/1112/289190/High-Incidence-of-Microsatellite-Instability-in)
540 [Microsatellite-Instability-in](https://aacrjournals.org/clincancerres/article/9/3/1112/289190/High-Incidence-of-Microsatellite-Instability-in)
- 541 50. Gutierrez C, Ogino S, Meyerhardt JA, Iorgulescu JB. The Prevalence and Prognosis of
542 Microsatellite Instability-High/Mismatch Repair-Deficient Colorectal Adenocarcinomas in
543 the United States. *JCO Precis Oncol*. 2023;7(7):e2200179. doi:10.1200/po.22.00179
- 544 51. Quaas A, Biesma HD, Wagner AD, et al. Microsatellite instability and sex differences in
545 resectable gastric cancer – A pooled analysis of three European cohorts. *Eur J Cancer*.
546 2022;173:95-104. doi:10.1016/j.ejca.2022.06.025
- 547 52. Joanito I, Wirapati P, Zhao N, et al. Single-cell and bulk transcriptome sequencing
548 identifies two epithelial tumor cell states and refines the consensus molecular
549 classification of colorectal cancer. *Nat Genet*. 2022;54(7):963-975. doi:10.1038/s41588-
550 022-01100-4
- 551 53. Li J, Wu C, Hu H, et al. Remodeling of the immune and stromal cell compartment by PD-1
552 blockade in mismatch repair-deficient colorectal cancer. *Cancer Cell*.
553 doi:10.1016/j.ccell.2023.04.011
- 554 54. Wu T, Zhang X, Liu X, et al. Single-cell sequencing reveals the immune microenvironment
555 landscape related to anti-PD-1 resistance in metastatic colorectal cancer with high
556 microsatellite instability. *BMC Med*. 2023;21(1):1-18. doi:10.1186/s12916-023-02866-y
- 557 55. Tickle T, Tirosh I, Georgescu C, Brown M, Haas B. inferCNV of the Trinity CTAT Project.
558 Published online 2019. <https://github.com/broadinstitute/inferCNV>
- 559 56. Nofech-Mozes I, Soave D, Awadalla P, Abelson S. Pan-cancer classification of single cells
560 in the tumour microenvironment. *Nat Commun*. 2023;14(1):1-14. doi:10.1038/s41467-
561 023-37353-8
- 562 57. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*.
563 2008;28(5):1-26. doi:10.18637/jss.v028.i05
- 564 58. Wickham H. Ggplot2. *Wiley Interdiscip Rev Comput Stat*. 2011;3(2):180-185.
565 doi:10.1002/wics.147
- 566 59. Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and
567 scalable single-cell analysis. *Nat Biotechnol*. 2024;42(2):293-304. doi:10.1038/s41587-
568 023-01767-y
- 569 60. Yates J, Kraft A, Boeva V. Filtering cells with high mitochondrial content depletes viable

570 metabolically altered malignant cell populations in cancer single-cell studies. *Genome*
571 *Biol.* 2025;26(1):1-26. doi:10.1186/s13059-025-03559-w

572 61. Christopher M, John R. Package “MLeval” Machine Learning Model Evaluation. Published
573 online 2022. <https://cran.r-project.org/package=MLeval>

574 62. R Studio Team. A language and environment for statistical computing. *R Found Stat*
575 *Comput.* 2021;3:<https://www.R-project.org>. <http://www.r-project.org>

576 STAR★METHODS

577 KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Previously published single-cell RNA sequencing data	Joanito et al. ⁵²	EGA ID'S: EGAD00001008555, EGAD00001008584, EGAD00001008585
Previously published single-cell RNA sequencing data	Li et al. ⁵³	GEO: GSE205506
Previously published single-cell RNA sequencing data	Wu et al. ⁵⁴	SRA: PRJNA932556
SINGLE-MSI Pipeline manuscript results	This paper	Zenodo: https://doi.org/10.5281/zenodo.18249691
Software and algorithms		
Cell Ranger version 7.2.0	10X Genomics	https://www.10xgenomics.com/support/software/cell-ranger/downloads/
Conda version 24.1.2	Anaconda	https://anaconda.org/anaconda/conda
InferCNV version 1.20.0	Tickle et al. ⁵⁵	https://anaconda.org/channels/bioconda/packages/bioconductor-infercnv/files
MSIsensor-RNA version 0.1.6	Jia et al. ⁴⁵	https://anaconda.org/channels/bioconda/packages/msisensor-rna/overview

scATOMIC version 2	Nofech-Mozes et al. ⁵⁶	https://github.com/abelson-lab/scATOMIC
SINGLE-MSI Pipeline	This paper	Zenodo: https://doi.org/10.5281/zenodo.18250137
Snakemake version 8.27.1	Mölder et al. ³⁴	https://snakemake.github.io/
R version 4.3.3	R CoreTeam	https://www.r-project.org/
R package caret version 7.0-1	Kuhn ⁵⁷	https://cran.r-project.org/web/packages/caret/index.html
R package ggplot2 version 3.5.1	Wickham ⁵⁸	https://cran.r-project.org/web/packages/ggplot2/index.html
R package MLeval version 0.3	CRAN	https://cran.r-project.org/web/packages/MLeval/index.html
R package Seurat version 5.1.0	Hao et al. ⁵⁹	https://cran.r-project.org/web/packages/Seurat/index.html
R package stats version 1.3.0	R Studio Team	https://cran.r-project.org/doc/manuals/r-patched/packages/stats/refman/stats.html

578

579 Experimental model and study participant details

580 We used single-cell RNA sequencing data that was generated as part of three previous
581 studies^{52–54} (**Error! Reference source not found.**, Key Resources Table). Raw FASTQ files were
582 downloaded from either the European Genome-Phenome Archive or from the Sequence Read
583 Archive (Key Resources Table). All other data was downloaded in matrix format from the Gene
584 Expression Omnibus (Key Resources Table). The data consists of 134 samples from 49
585 individuals with metastatic or non-metastatic colorectal cancer. Individuals were grouped into
586 MSI-H and MSS categories based on the original PCR/IHC clinical status reported in previous
587 studies. In total there were 29 deemed MSI-H, 18 MSS, and two did not have a reported MSI
588 status (**Error! Reference source not found.**). Each sample was created with either Single Cell 3'
589 v2, 3' v3, or 5' Reagent Kit from 10X Genomics and was sequenced either on an Illumina
590 NextSeq 500, NovaSeq 6000, BGISEQ DNBSEQ-T7, or HiSeq X Ten machine. Complete

591 sequencing and library preparation information can be found by referencing the Dataset ID in
592 **Error! Reference source not found..**

593 Individuals from datasets EGAD00001008555, EGAD00001008584, EGAD00001008585 had
594 multi-regional samples from the same tumor and multi-site samples from metastatic tissue and
595 lymph nodes. Although variation in the multi-site samples could be considered intra-individual
596 heterogeneity rather than ITH, we kept them in the analysis to retain as many cancer cells and
597 as much heterogeneity as possible. The other two datasets GSE205506 and PRJNA932556
598 include individuals that had treatment for MSI (anti-PD-1 and celecoxib). We excluded the
599 following samples because we did not identify any cancer cells: XHC080-SI-GA-B11, XHC082-SI-
600 GA-C1, XHC127-SI-GA-F10, EXT129, EXT051, and EXT097.

601
602 While demographic metadata (age, gender, etc.) was available for most samples, we did not
603 factor this into our analysis and acknowledge it as a limitation. Our study was computational
604 and did not employ a traditional experimental design that would account for these confounding
605 factors. While there is evidence that age, race, and gender affect prognosis and frequency of
606 MSI-H cancer⁴⁸⁻⁵¹ our study was explorative in nature and did not aim to describe the
607 relationship between these types of variables and the detection of ITH in MSI. As there were
608 very few publicly available single-cell RNA sequencing datasets with paired clinical MSI status at
609 the time we conducted this study, little would have been gleaned from incorporating
610 demographic metadata. We suggested that studies with larger sample sizes would be needed in
611 order to determine the frequency of ITH in MSI at the population level, and those studies would
612 also be needed to investigate such questions related to demographic information.

613 All ethical approval information for each dataset can be found in the original publications (Table
614 2, Key Resources Table).

615

616 **METHOD DETAILS**

617 **Data processing**

618 We aligned FASTQ files to the GRCh38 human reference genome and converted them to a gene
619 count matrix using the 10x Genomics Cell Ranger v7.2.0 software suite. From there all matrix

620 files were processed using the R package Seurat⁵⁹ following the Seurat best practices tutorials
621 (<https://satijalab.org/seurat/>). Briefly, Seurat objects were created and only genes detected in a
622 minimum of 3 cells were used for downstream analysis. We further filtered out cells with fewer
623 than 100 features, more than 3000 features, and if a cell had more than 35 percent of all genes
624 labeled as mitochondrial. While we followed the Seurat best practices closely, the default
625 settings aim to maximize immune cell type identification and filtered out the majority of cancer
626 cells. The filter settings described here were designed to maximize the number of tumor cells
627 retained while still removing cells that were poor-quality or likely necrotic. These filter settings,
628 specifically mitochondrial gene percentage, are supported by a recent study that showed that
629 filter settings that are too strict remove viable cancer cells from single-cell sequencing data⁶⁰.
630 After filtering, all gene count matrices then were normalized using the LogNormalize option
631 with the scale setting set to 10,000. The 2,000 most variable genes were found using the “vst”
632 selection method and were used to cluster together groups of cells with the RunPCA function.
633 The first 15 PCA dimensions were used to run the following functions: FindNeighbors,
634 FindClusters (resolution set to 0.5), and RunUMAP. If an individual had multiple samples, they
635 were integrated together by using the IntegrateLayers function, with the method set to
636 CCAIntegration and k.weight set to 50. Each integrated sample then underwent re-clustering
637 with the settings previously mentioned. The integration step after subsetting down to only
638 cancer cells used the same settings except the FindClusters resolution was set to 0.8, and only
639 the first 10 principal components were used.

640 **Cell classification and measuring ITH**

641 After each sample is processed, all cells are classified as either cancer or normal and then MSI-H
642 or MSS. These classification steps are built upon two machine learning based programs trained
643 on large pan-cancer datasets. The first, scATOMIC⁵⁶, was used to distinguish tumor cells from
644 normal ones, and the second, MSIsensor-RNA⁴⁵ determined MSI status. Both tools were run
645 with default settings, but to get an MSI score for each cell, we had to transform the prebuilt
646 MSIsensor-RNA baseline file. This was done by filtering both the count matrix and baseline to
647 only include gene names common to both. The filtered baseline and count matrix files were

648 then used to get MSI scores for all cells within a sample. From there cells were classified as MSI-
649 H if they also were labeled as cancer by scATOMIC and if they had an MSI score of .75 or more
650 (75% probability the cell is MSI-H).

651 Levels of ITH were assessed with two methods. First, we measured ITH by testing for
652 differences in mean MSI score between cancer cell clusters with a one-way ANOVA test. The
653 ANOVA F-statistic was used to describe levels of heterogeneity in the biomarker, with a large
654 value of the F-statistic indicating greater heterogeneity in MSI. Secondly, we identified
655 subclones within each individual by comparing CNVs between MSI-H, MSS, and normal cells.
656 This was done by passing the relevant cell classification for each unique barcode to InferCNV⁵⁵
657 (DOI: 10.18129/B9.bioc.infercnv). InferCNV was used with default settings except in the case of
658 CRC2821, which had many more cancer cells than the other samples. We increased the k_nn
659 setting from the default of 20 to 50 to take into account the larger dataset. Lastly, we ran
660 differential expression using a Wilcoxon Rank Sum Test (the default for Seurat) between
661 clusters of cancer cells and between MSI-H and MSS cancer cells for each individual.

662 We verified how well our pipeline captured heterogeneity in MSI status by mixing together
663 randomly sampled tumor and normal cells in varying proportions using a custom R script. We
664 simulated varying levels of heterogeneity by mixing the cells of one sample that had
665 homogenous MSI-H cancer cells (GSM6213995 from individual P33) and one sample with
666 homogenous MSS cancer cells (XHC118-SI-GA-F1 from individual CRC2811). In total, we had
667 eleven different mixes, with the proportion of MSI-H cells ranging from 0 to 1 (in increments of
668 0.1) and the remainder being MSS (Table S1). The results of these mixing experiments were
669 replicated 100 times, except for the pure MSS and MSI-H cases, for which all cancer cells were
670 included.

671 Although MSIsensor-RNA has been shown to classify single-cell RNA sequencing samples
672 accurately⁴⁵, we checked its ability to distinguish between MSI-H and MSS samples in our
673 datasets at the individual level. This was done by scoring individuals with MSIsensor-RNA using
674 all available cells and again with just the cancer cells. We used the AggregateExpression

675 function in Seurat to create the two different scenarios, and measured MSIsensor-RNA
676 performance with ROC-AUC using the MLeval and caret packages in R^{57,61}.

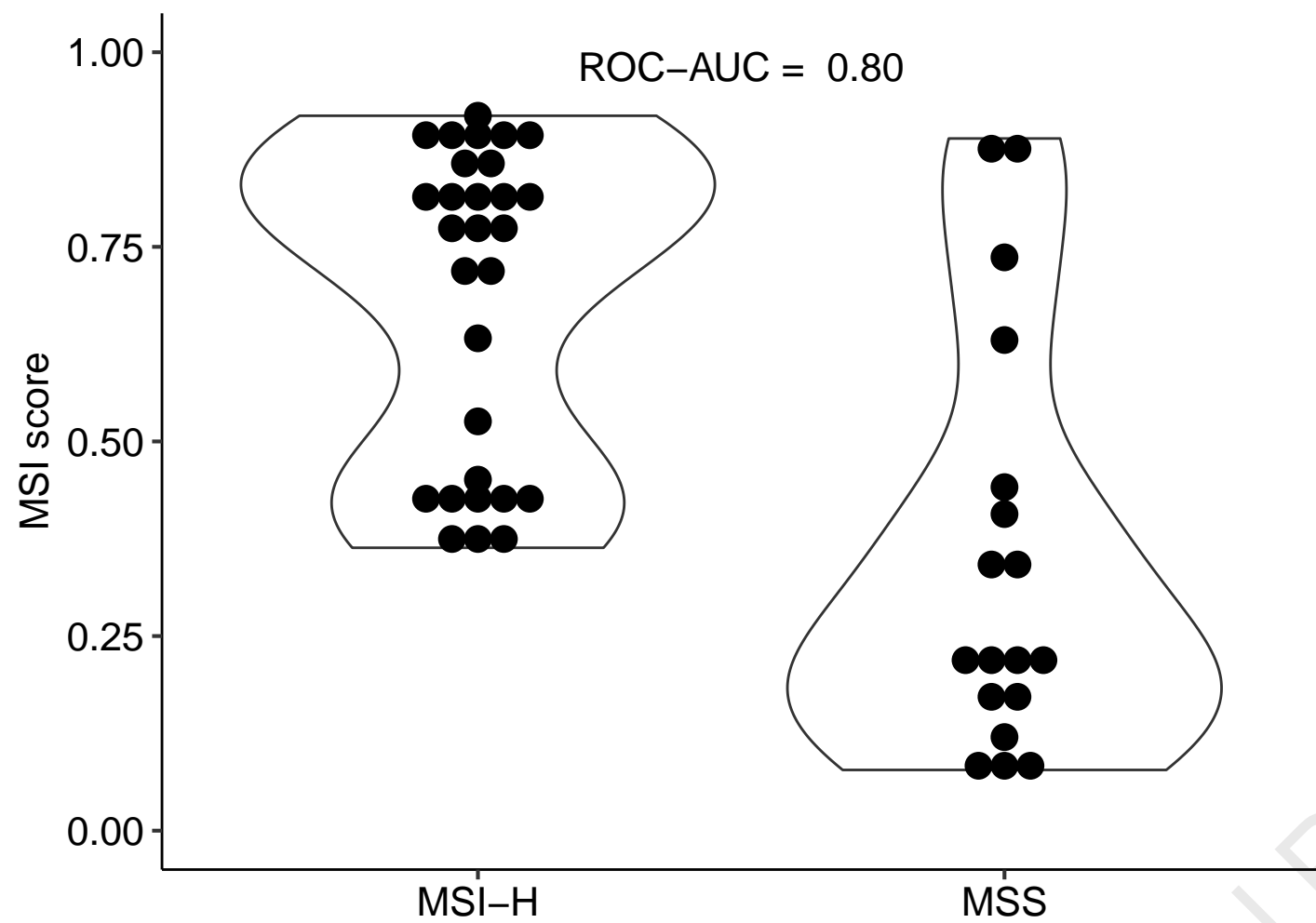
677 **QUANTIFICATION AND STATISTICAL ANALYSIS**

678 All statistical analyses were carried out in R (R version 4.3.3; <https://www.R-project.org/>)⁶² and
679 all plots were created with ggplot2⁵⁸. Two statistical tests were performed as part of our
680 computational pipeline. The first is a one-way ANOVA test that we used to measure ITH by
681 comparing the difference in means between clusters of cancer cells. This was done with the aov
682 function and was followed with Tukey's Honestly Significant Difference test using the TukeyHSD
683 function, both of which are from the stats package⁶².

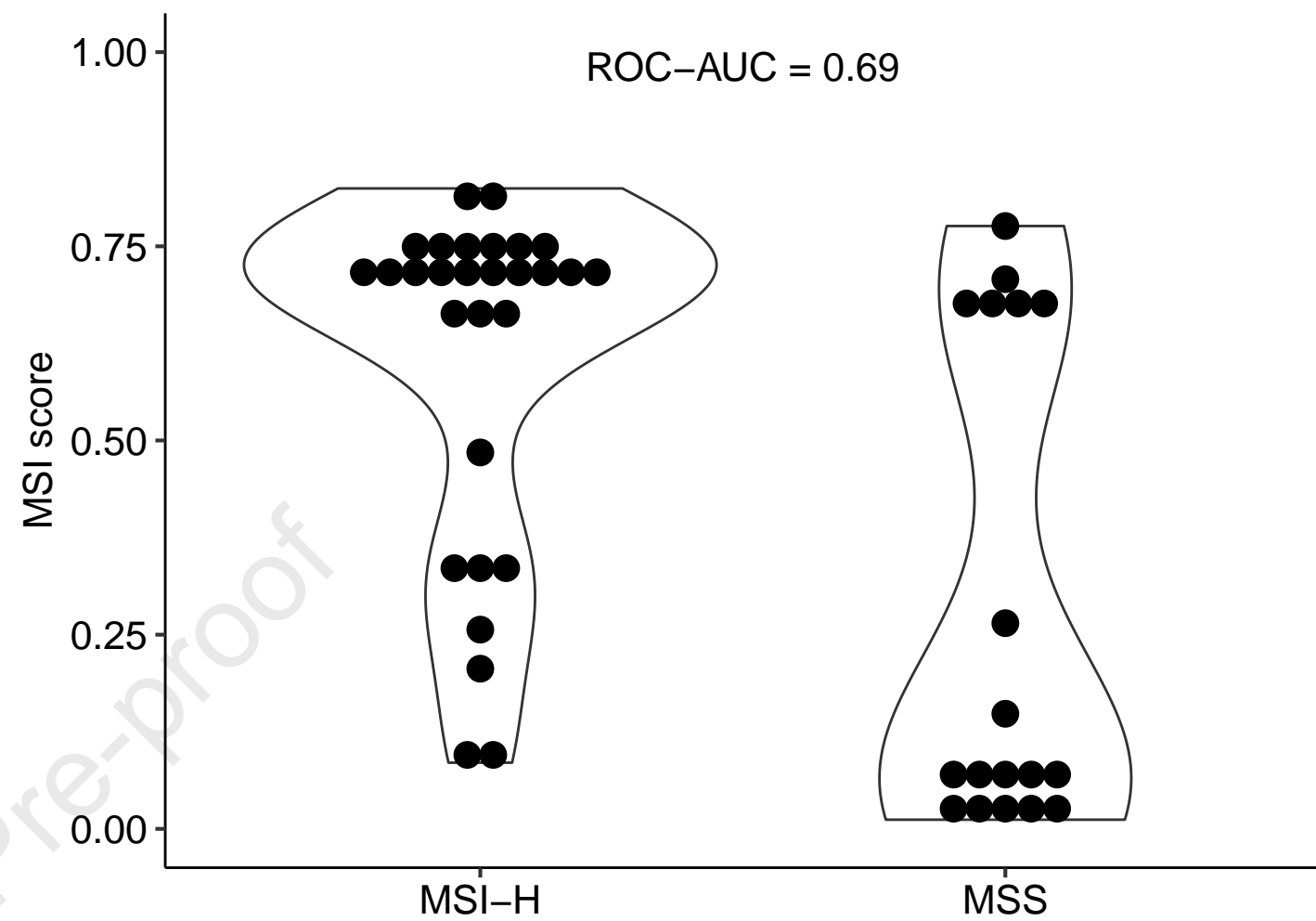
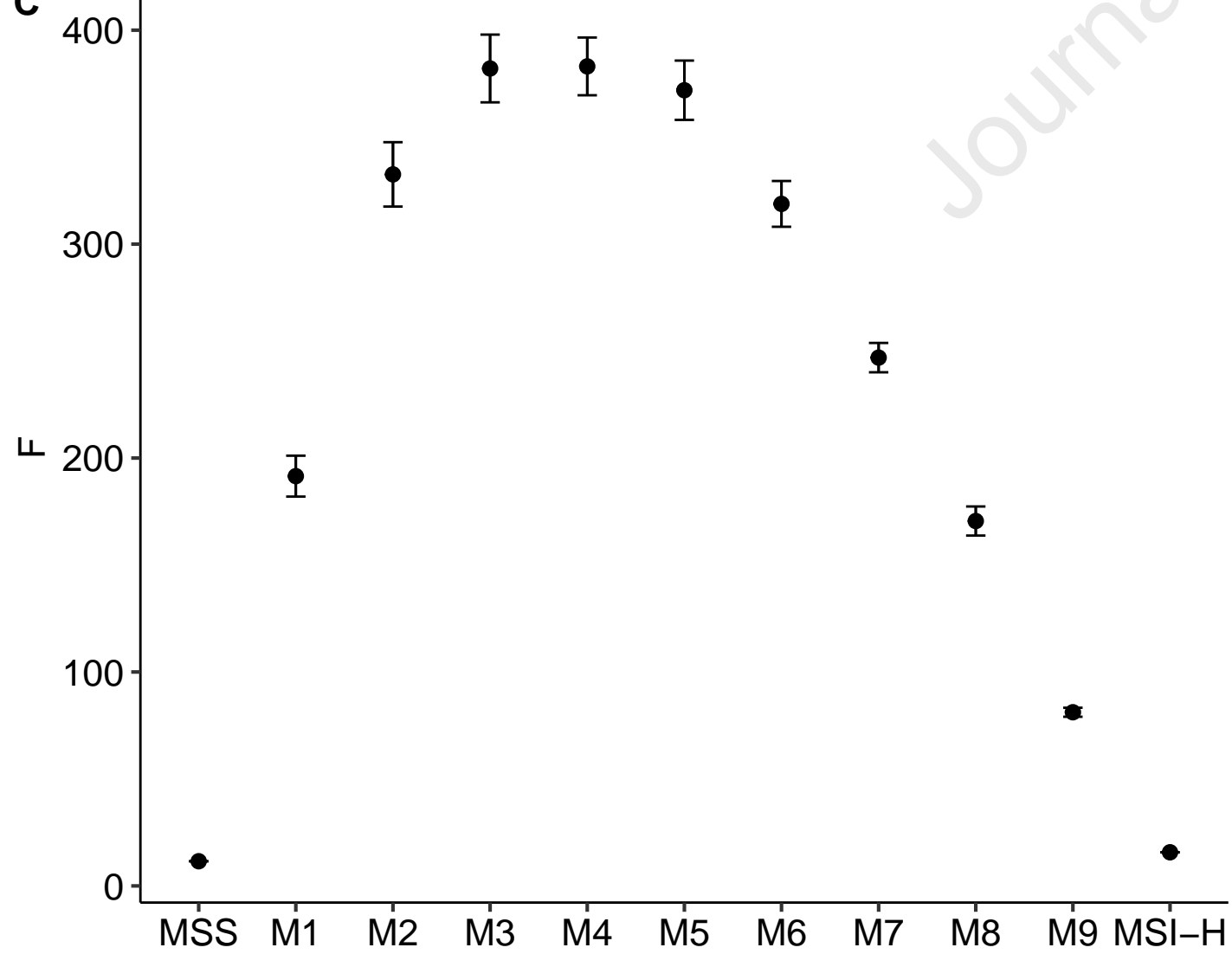
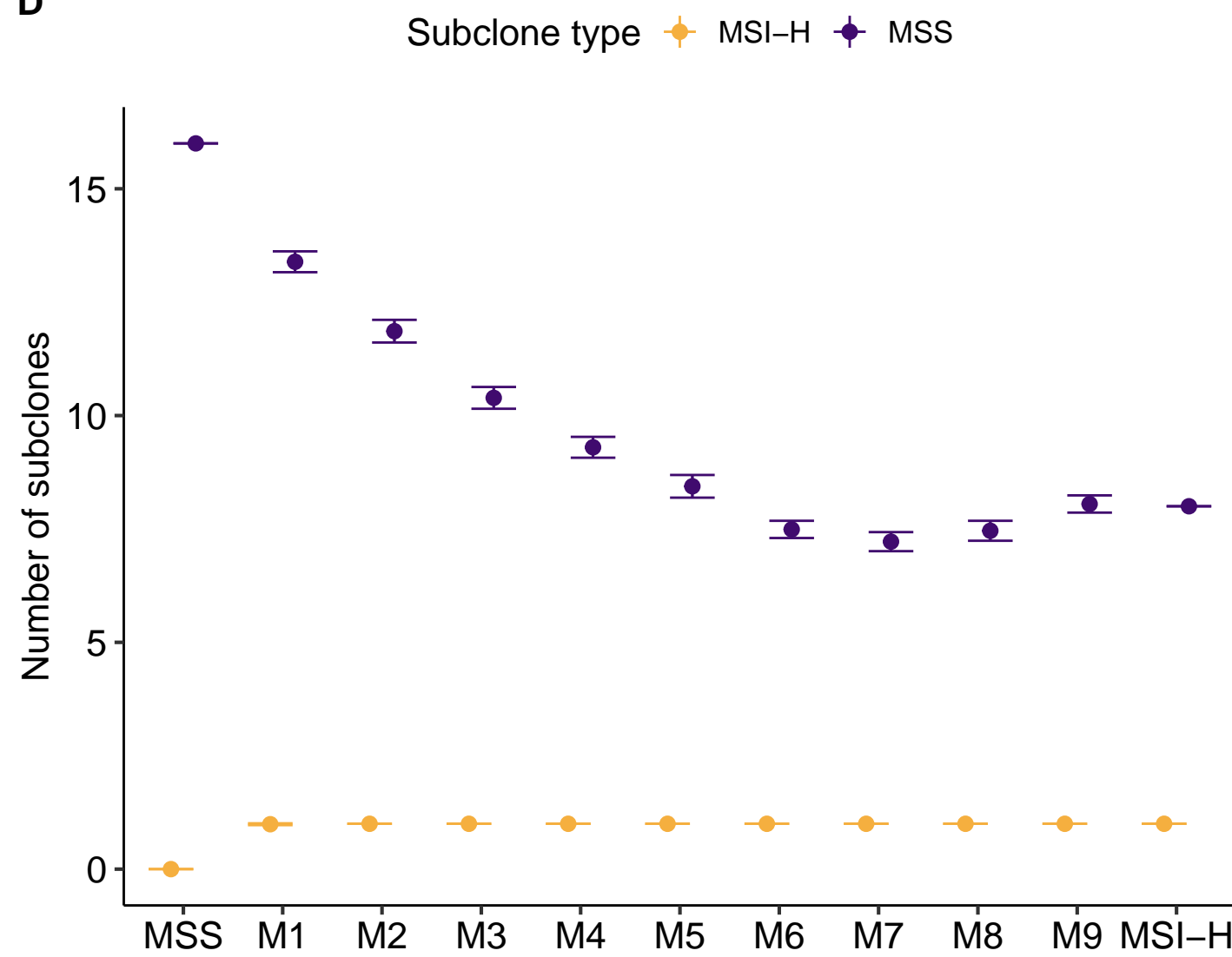
A

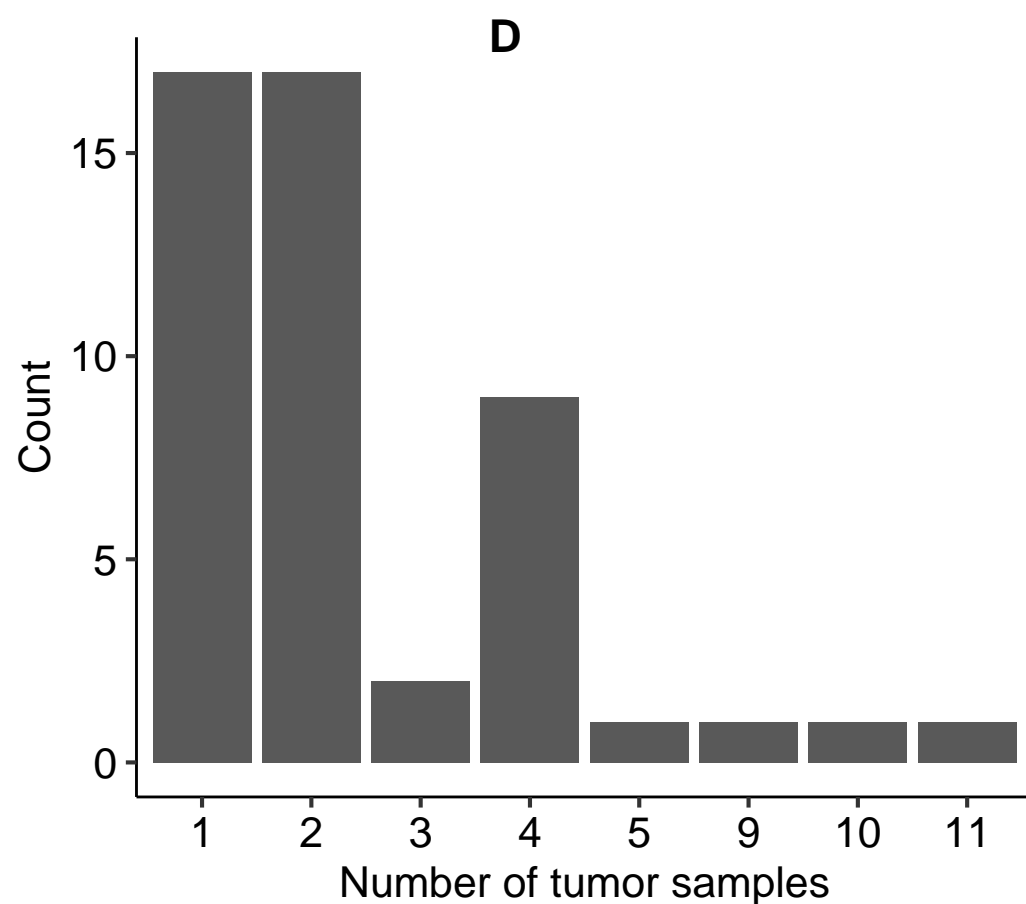
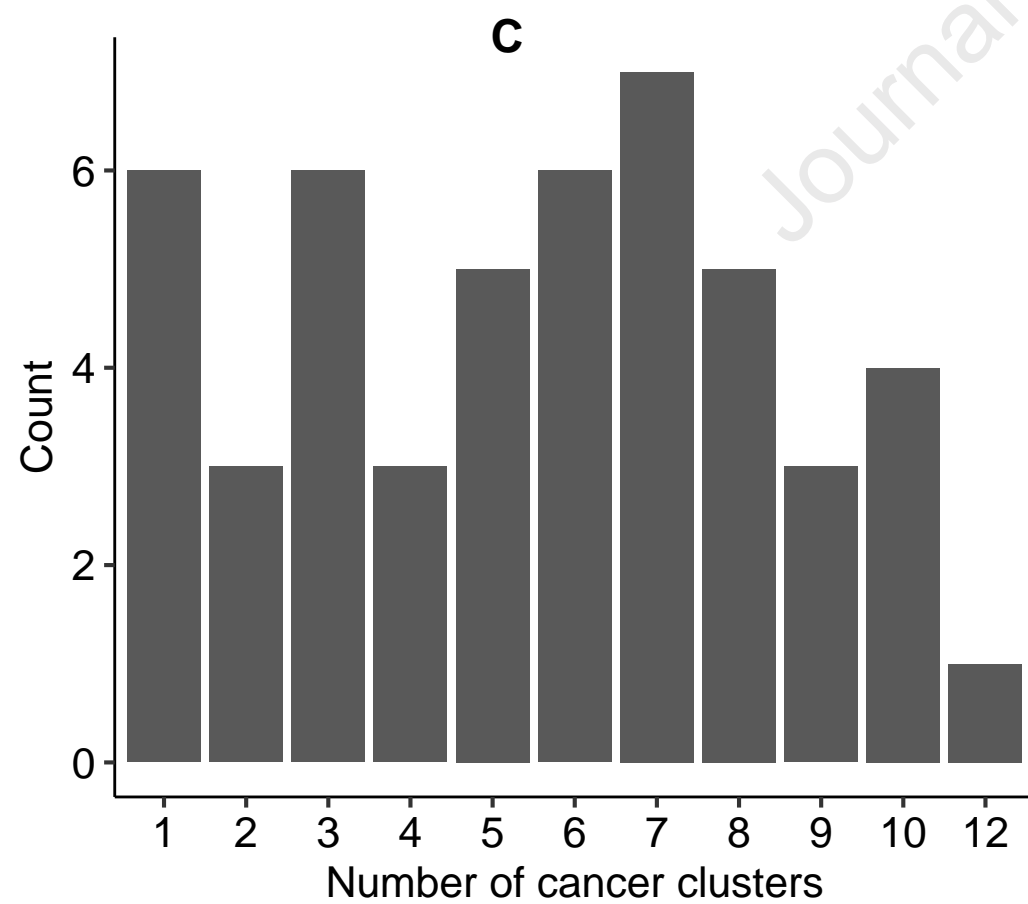
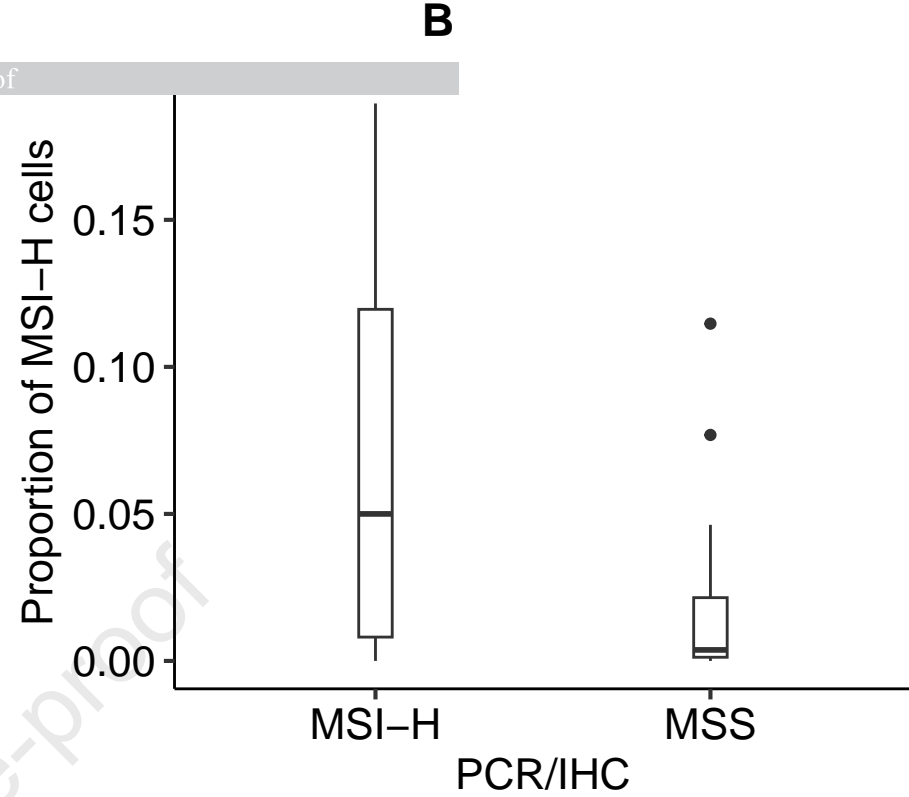
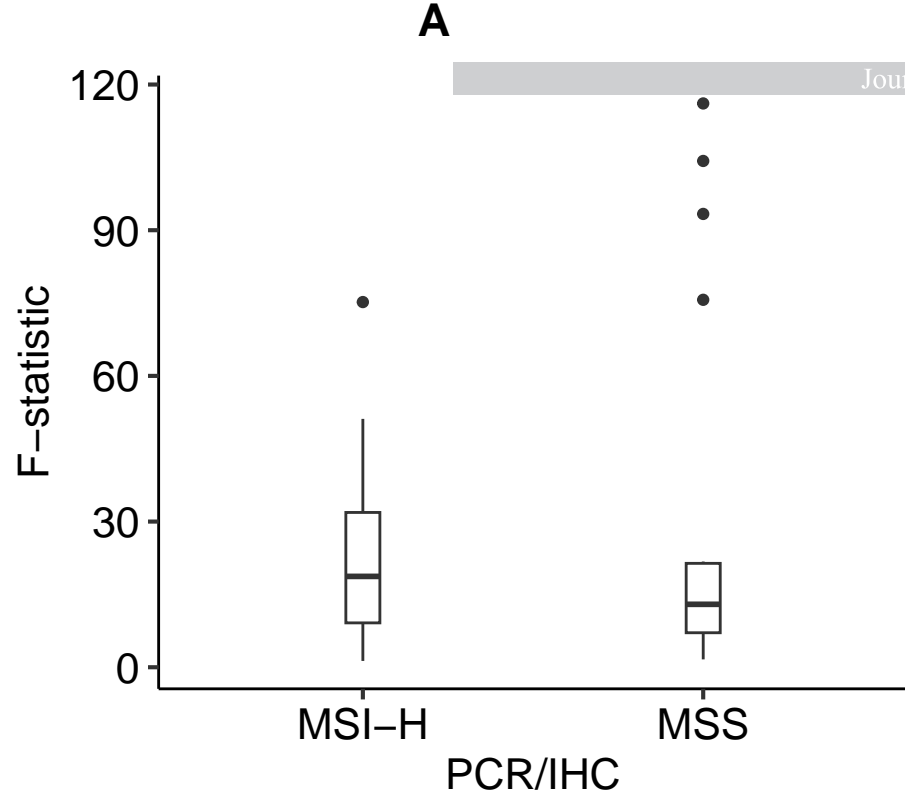
All cells

Journal Pre-proof

**B**

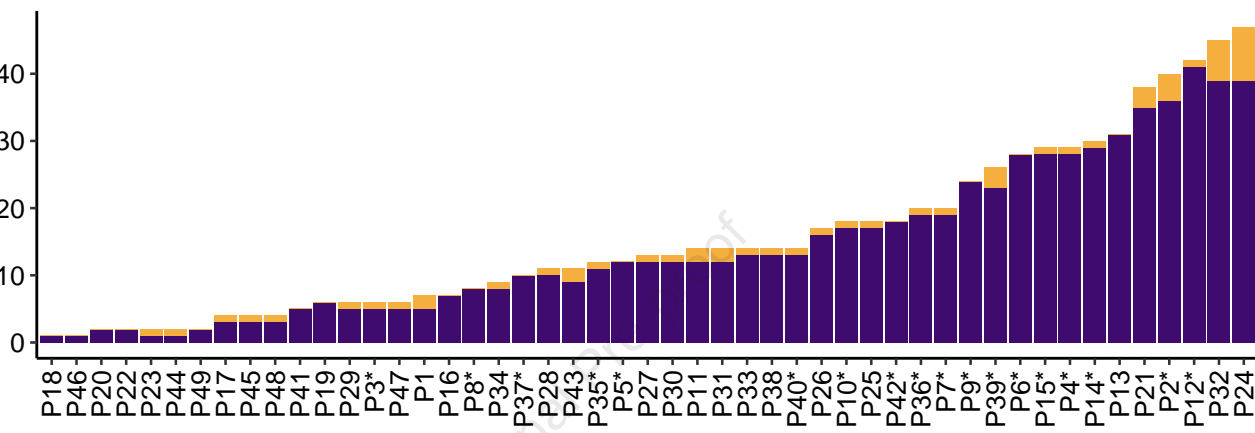
Cancer cells

**C****D**

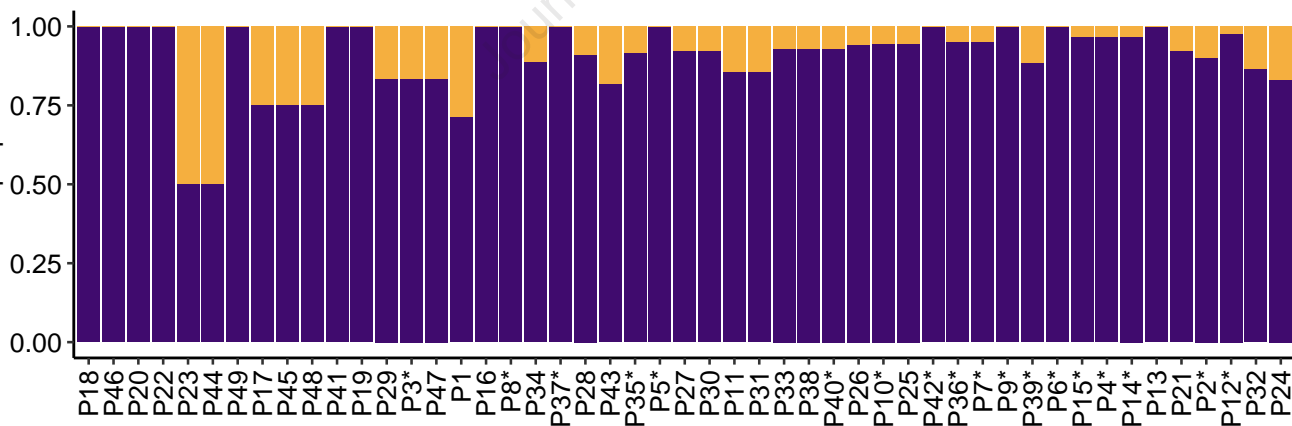


A

Subclone counts

**B**

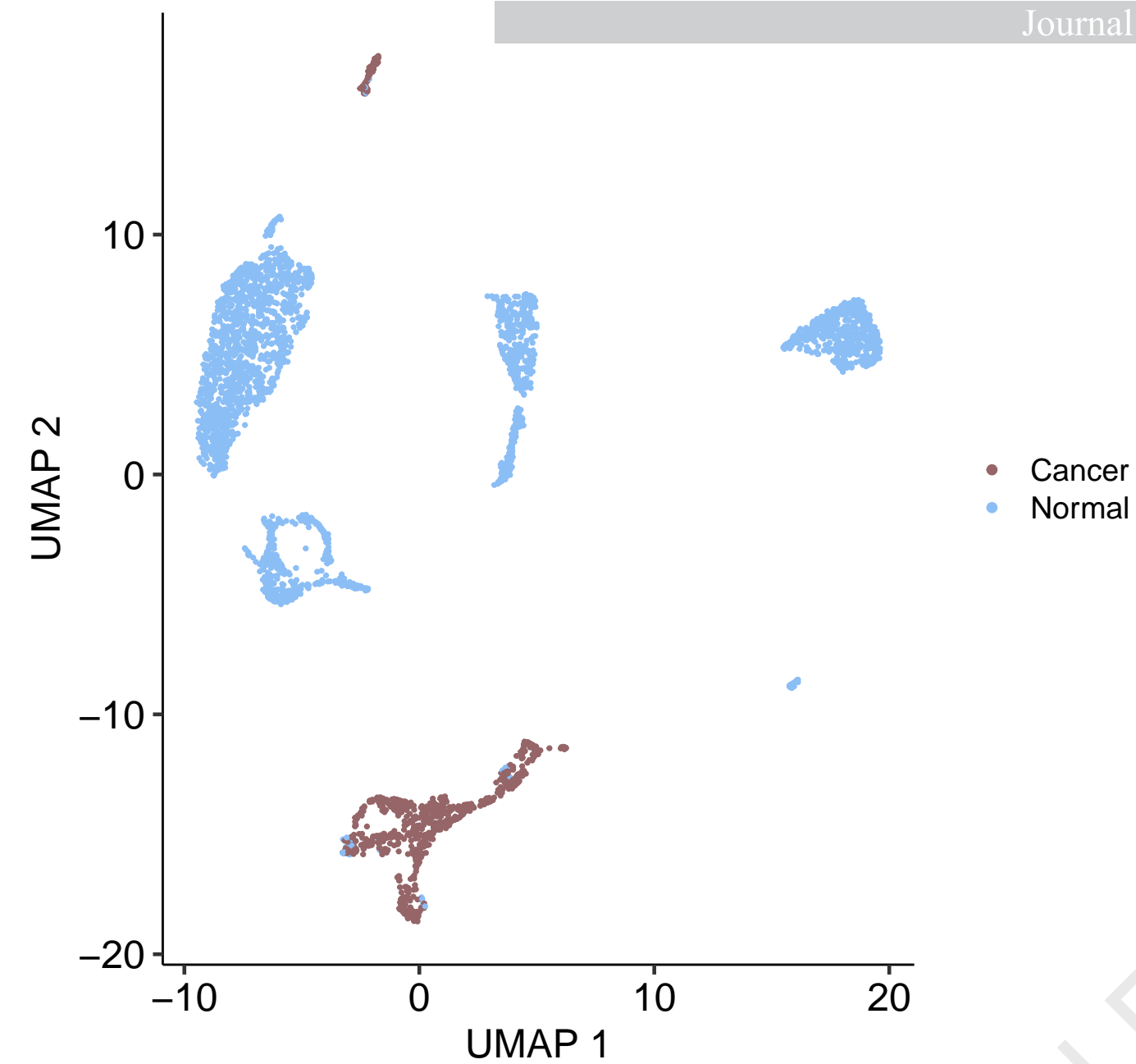
Subclone proportion



A

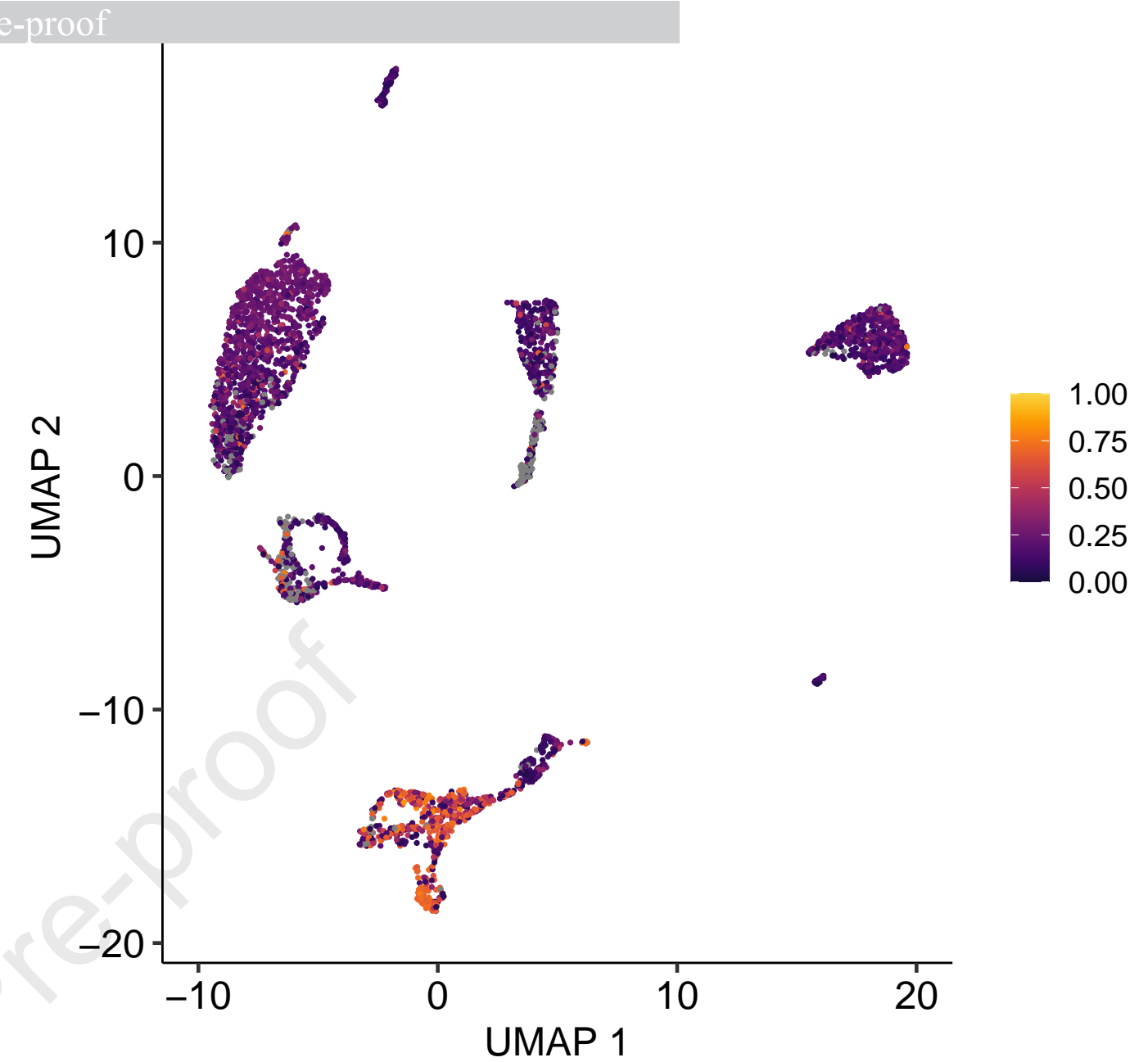
Simplified cell types

Journal Pre-proof



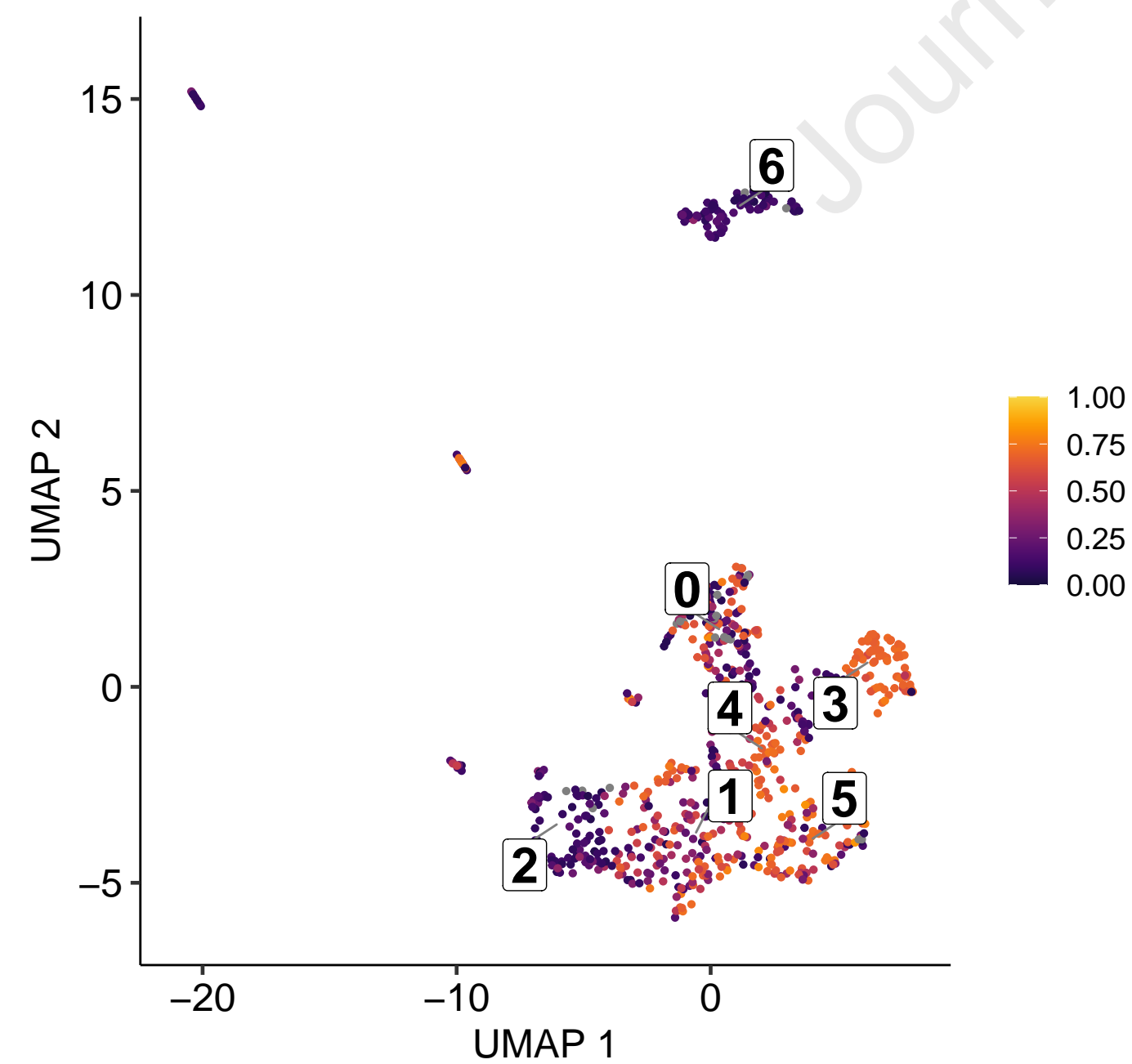
B

MSI score



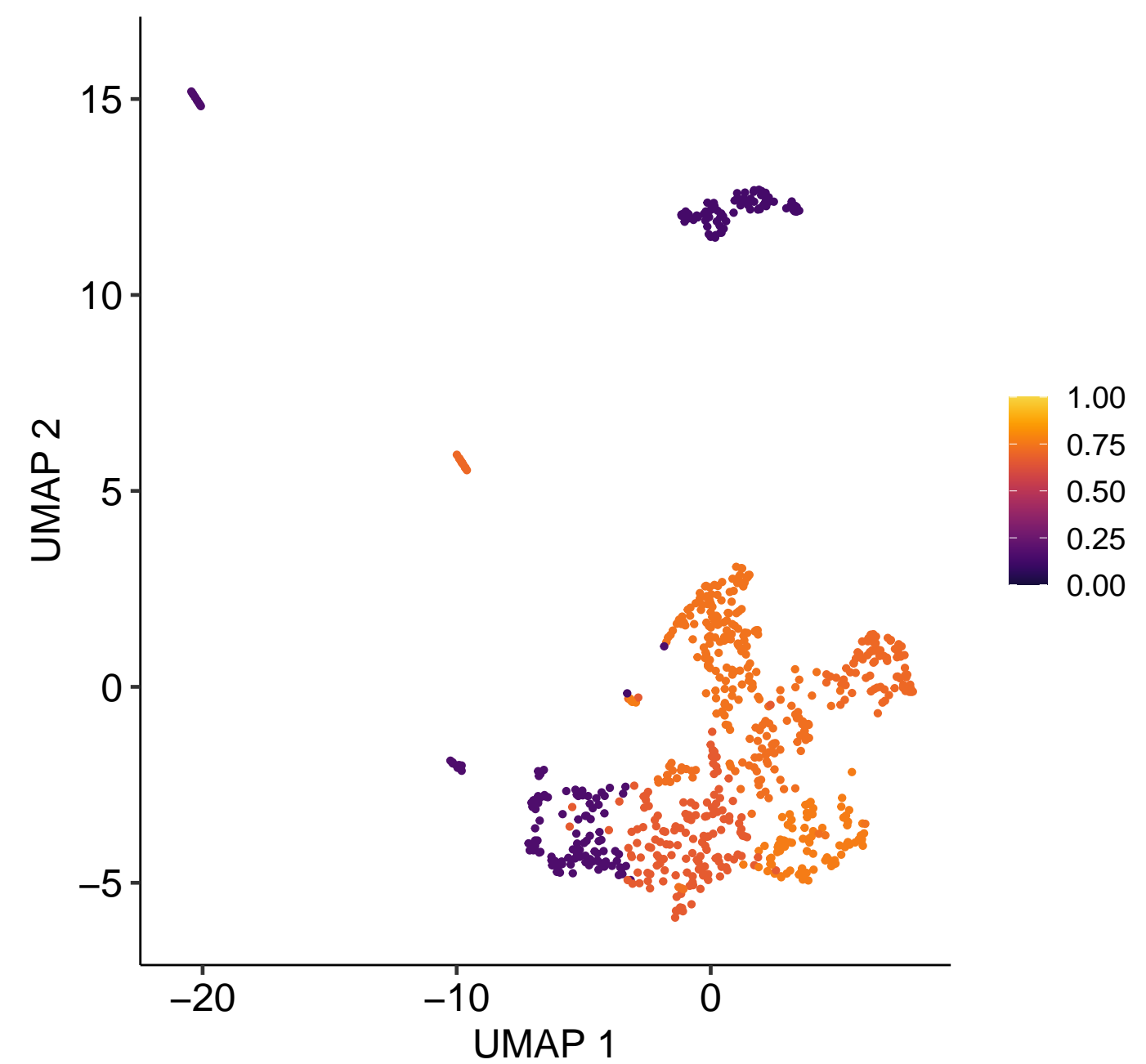
C

Recluster of cancer cells



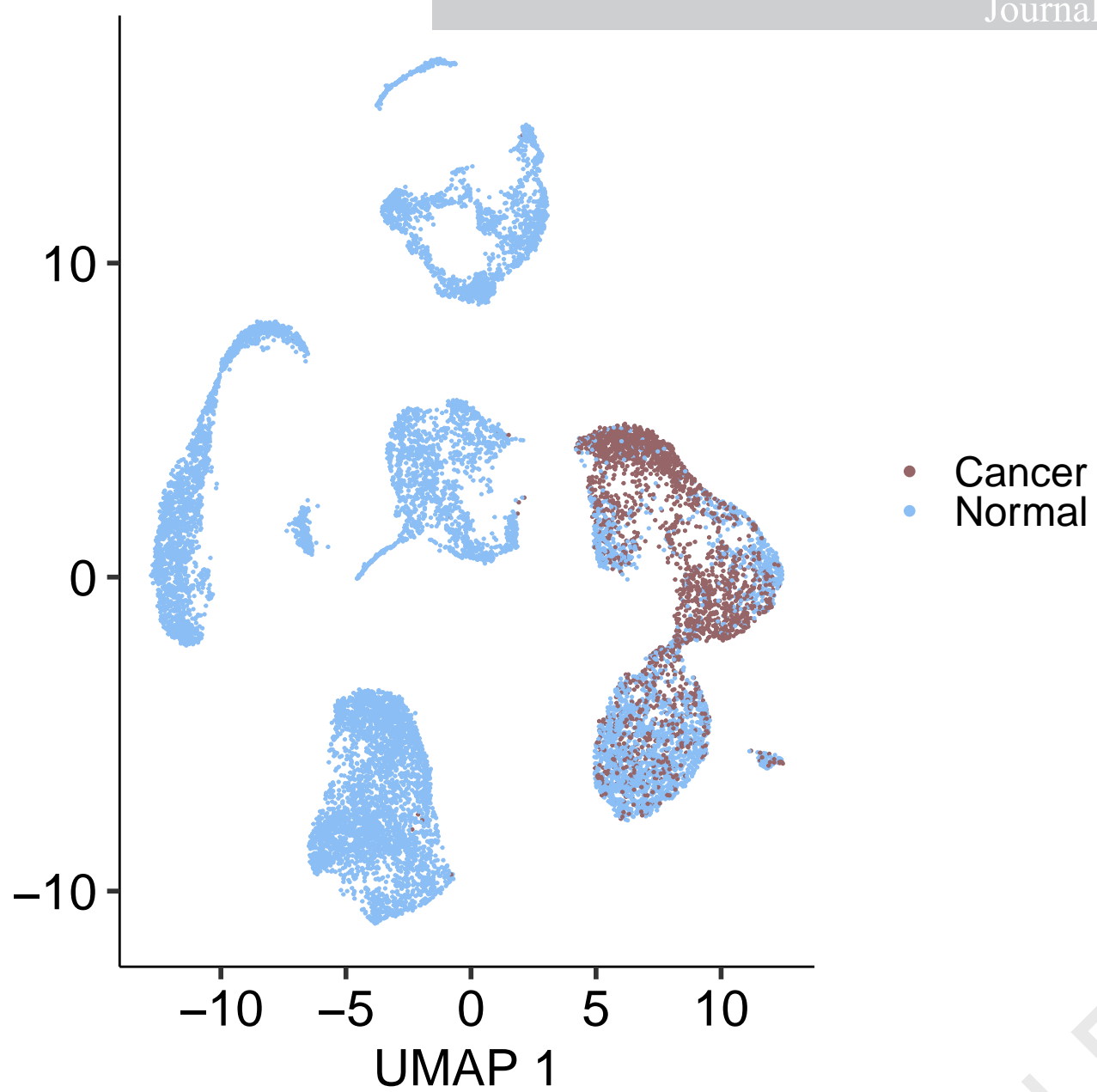
D

Cancer cells pseudobulk MSI score



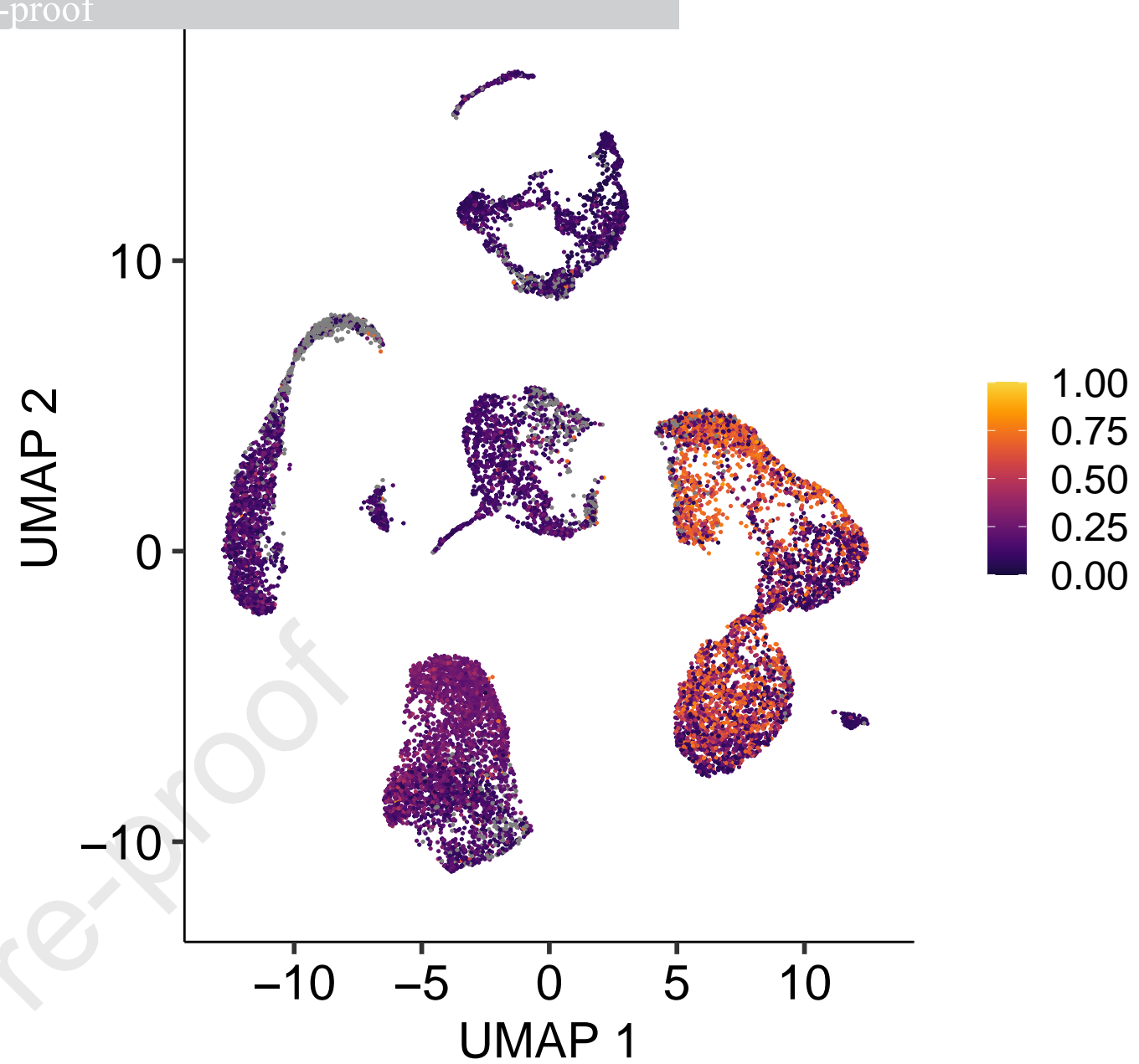
A

Simplified cell types



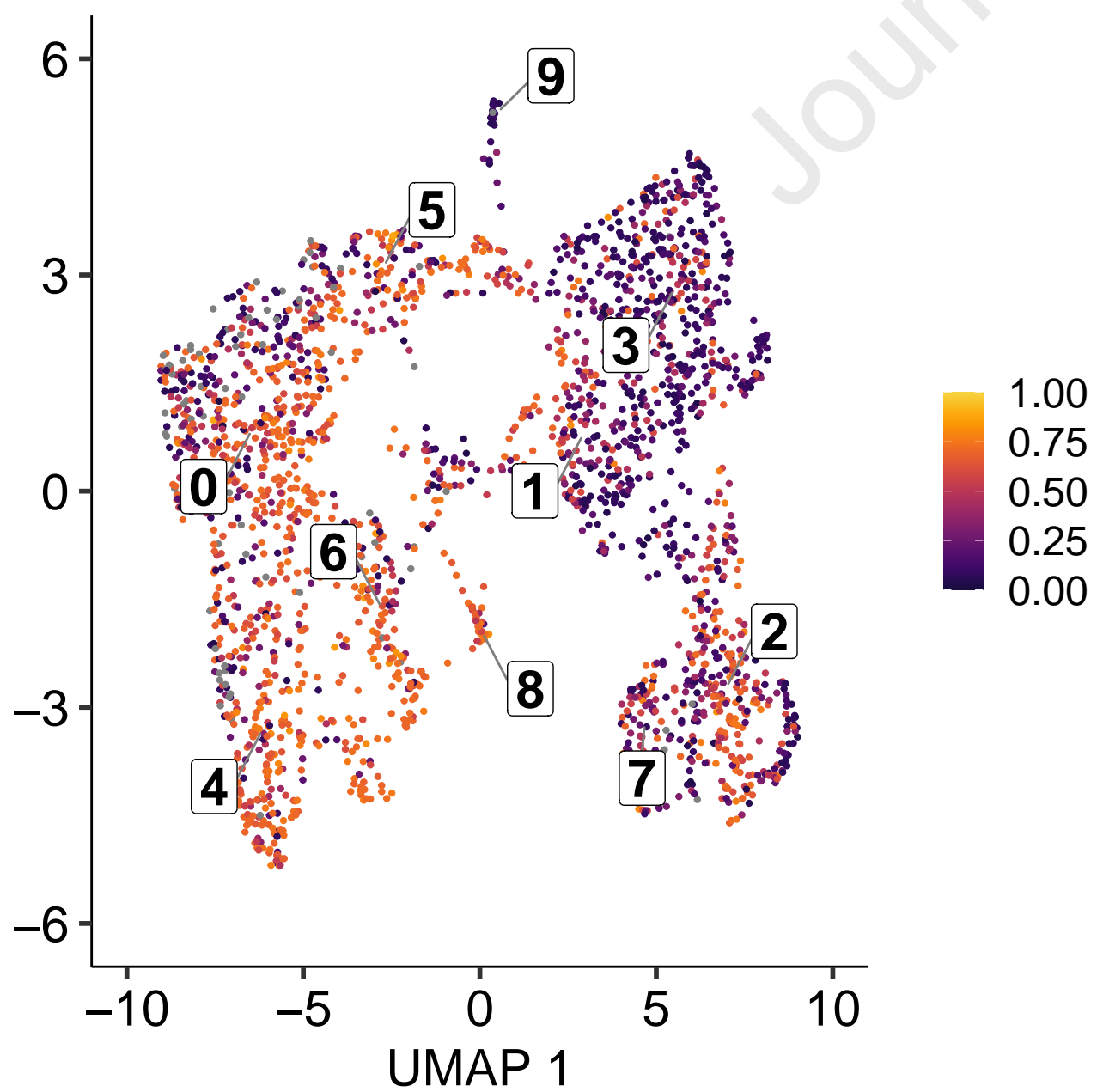
B

MSI score



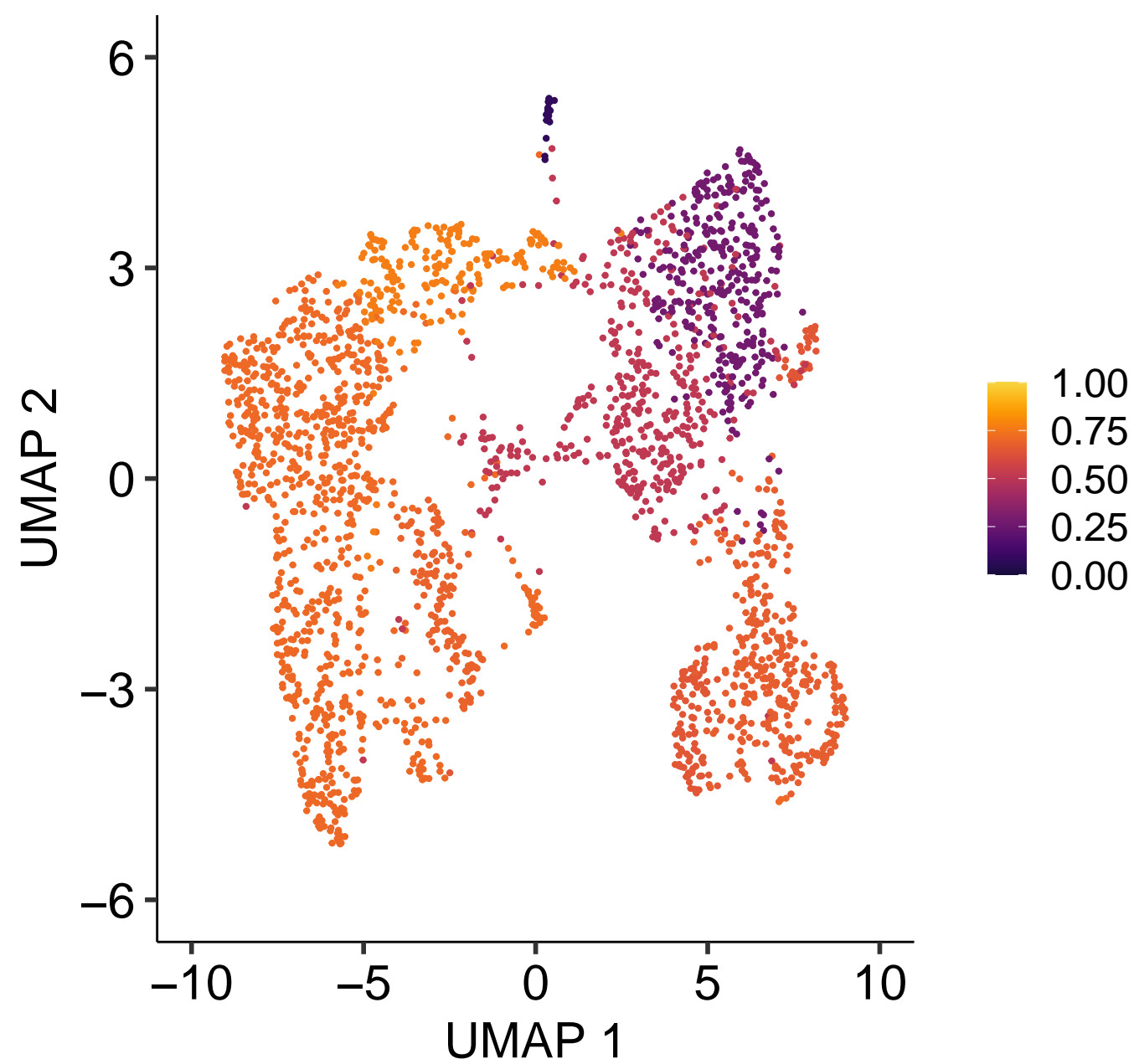
C

Recluster of cancer cells



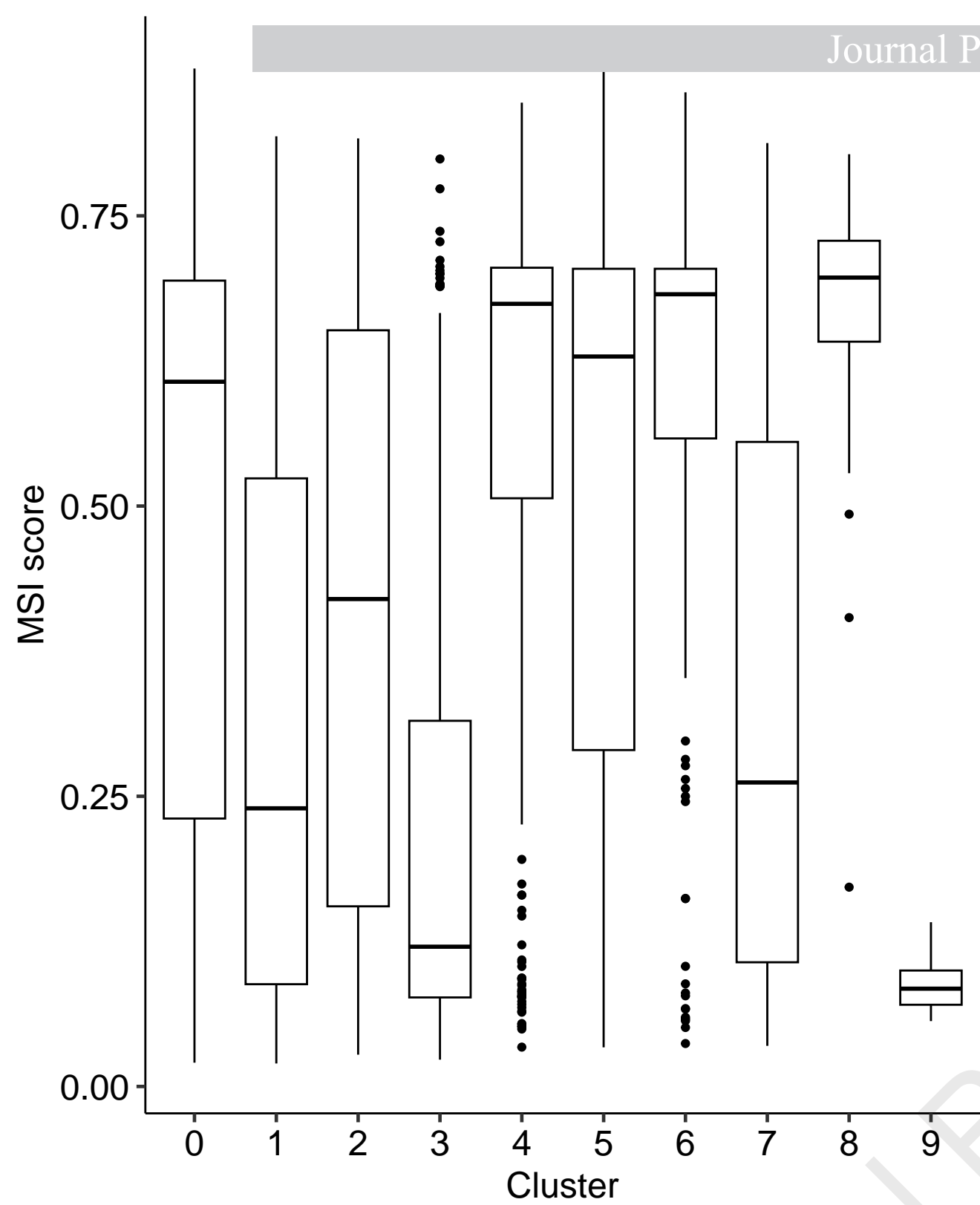
D

Cancer cells pseudobulk MSI score



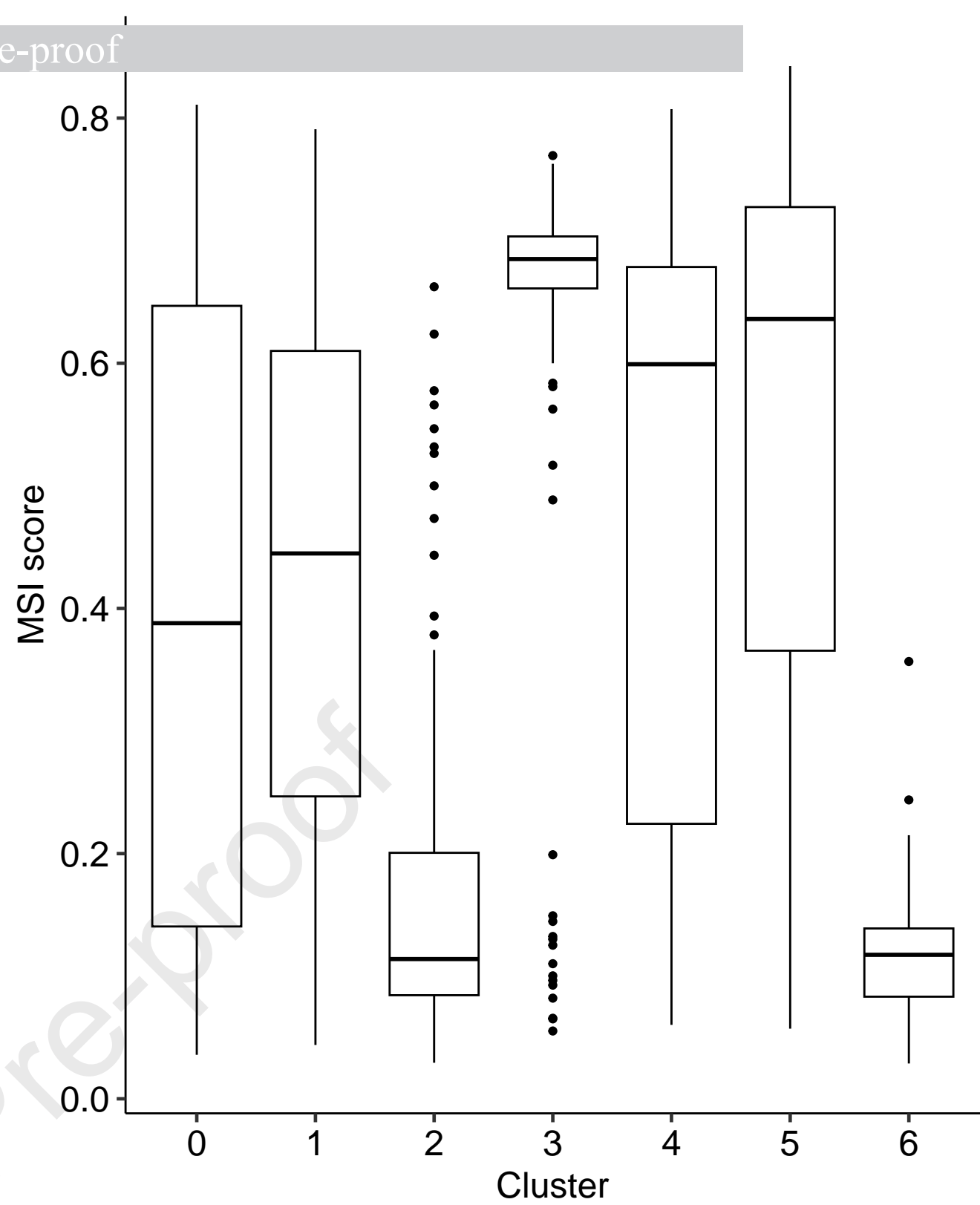
A

F=77.05



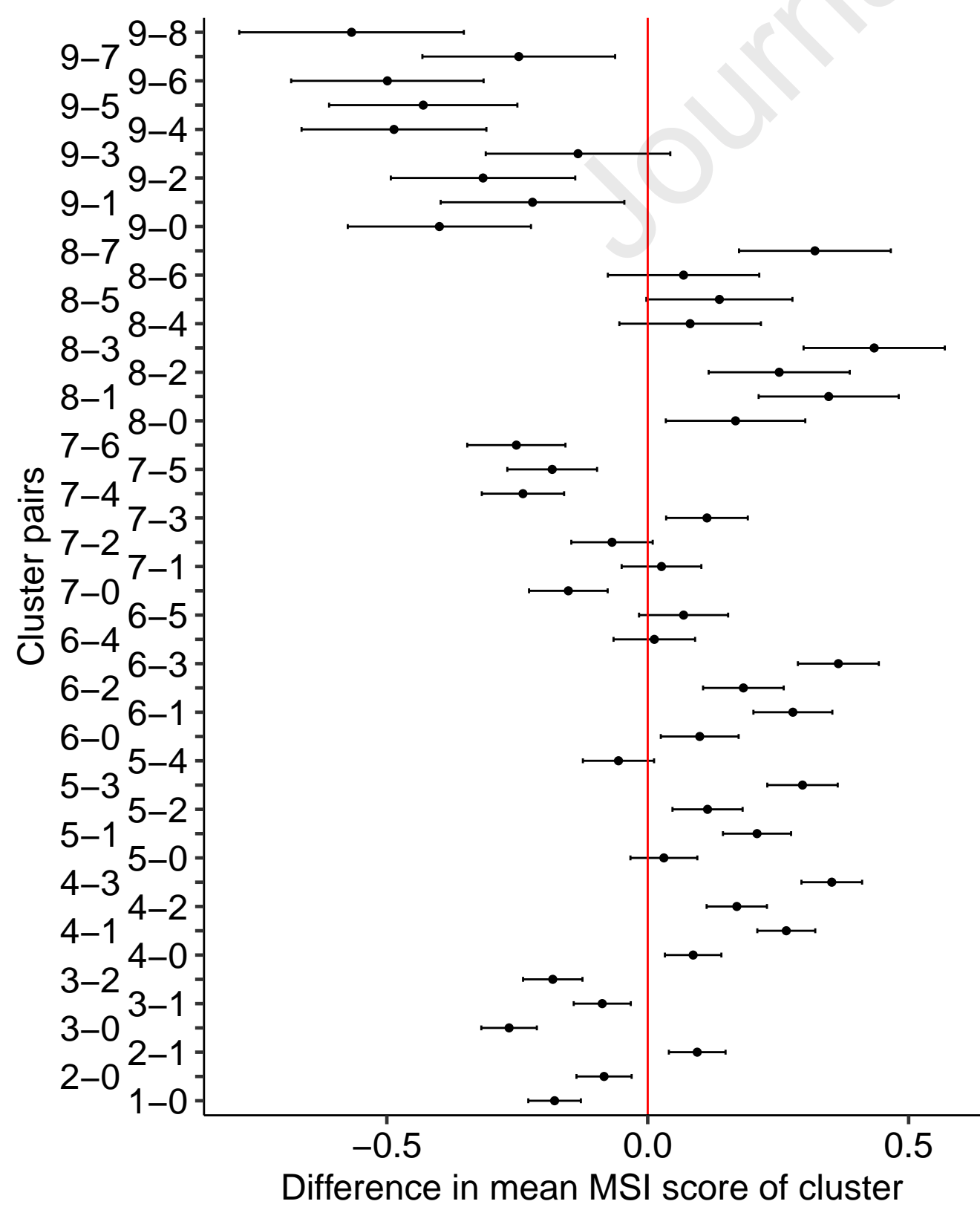
B

F=75.2



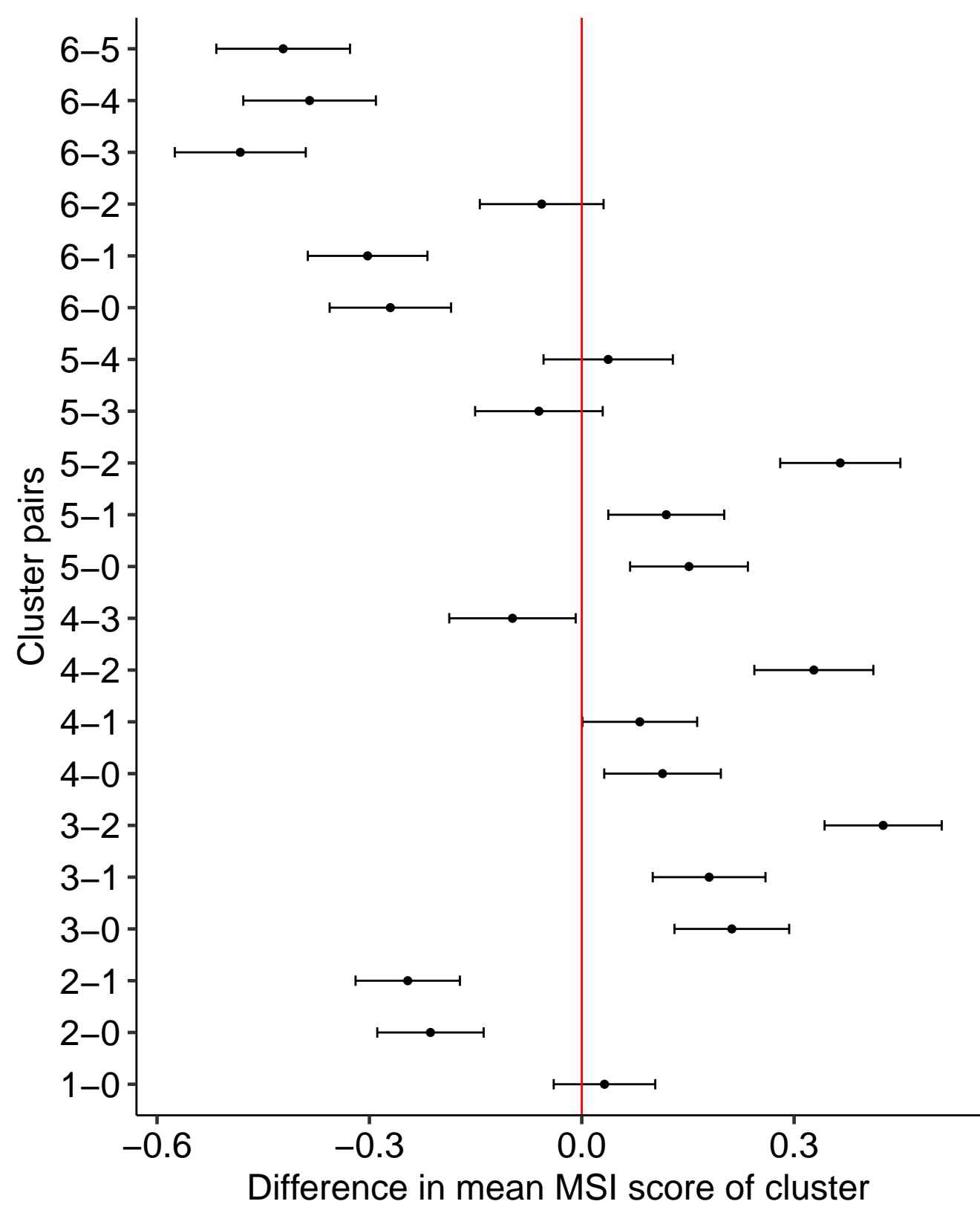
C

CRC2786 95% family-wise CI



D

P24 95% family-wise CI



- 1.) Novel computational pipeline quantifies ITH and MSI at single-cell resolution
- 2.) 15 out of 49 individuals have evidence of ITH in MSI
- 3.) Both MSI-H and MSS individuals have tumors with distinct MSI-H and MSS subclones
- 4.) Findings challenge the current binary classification of MSI status

Journal Pre-proof