Brd4::Nutm1 fusion gene initiates NUT carcinoma *in vivo*

Dejin Zheng[1,2,10], Ahmed Elnegiry[1,2,8,10], Chenxiang Luo[1,2,9,10], Mohammed Amine Bendahou[3], Liangqi Xie[3], Diana Bell[4], Yoko Takahashi[5], Ehab Hanna[5], George I. Mias[2,6], Mayra F. Tsoi[7, #] and Bin Gu[1,2, #, *]

Affiliations:

[1] Department of Obstetrics, Gynecology and Reproductive Biology, College of Human Medicine, Michigan State University East Lansing, Michigan, United States

[2] Institute for Quantitative Health Science and Engineering, Michigan State University East Lansing, Michigan, United States

[3] Infection Biology and Cancer Biology Program, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, United States.

[4] City of Hope Comprehensive Cancer Center, 20220, Pathology, Duarte, California, United States;

[5] Department of Head and Neck Surgery, The University of Texas MD Anderson Cancer Center, Houston, United States

[6] Department of Biochemistry and Molecular Biology, College of Nature Science, Michigan State University East Lansing, Michigan, United States

[7] Department of Pathobiology and Diagnostic Investigation, College of Veterinary Medicine, Michigan State University East Lansing, Michigan, United States

[8] Home Institution: Department of Cytology and Histology, Faculty of Veterinary Medicine, Aswan University, Aswan, Egypt

[9] Home Institution: Center for Reproductive Medicine and Department of Gynecology & Obstetrics, the First Affiliated Hospital, Sun Yat-Sen University, Guangzhou, PR China.

[10]Equal contribution.

#Correspondence to Bin Gu (gubin1@msu.edu) and Mayra Tsoi (tsoimayr@msu.edu)

*Lead contact

1. Abstract: Nut carcinoma (NC) is an aggressive cancer with no effective treatment. The majority (70%) of NUT carcinoma is associated with chromosome translocation events that lead to the formation of a *BRD4::NUTM1* fusion gene. However, because the *BRD4::NUTM1* gene is unequivocally cytotoxic when ectopically expressed in cell lines, questions remain on whether the fusion gene can initiate NC. Here, we report the first genetically engineered mouse model (GEMM) for NUT carcinoma recapitulating the human mutation. By stochastically inducing a chromosome translocation mirroring the human event, we demonstrated that the *Brd4::Nutm1* fusion gene could induce aggressive carcinomas in mice. The tumors present histopathological and molecular features similar to human NC, with an enrichment of undifferentiated cells. Similar to the reports of human NC incidence, *Brd4::Nutm1* can induce NC from a broad range of tissues, demonstrating that its oncogenic potential is not lineage-restricted. The consistent induction of tumors of squamous phenotypes, even from ductal epithelial and mesenchymal tissues, demonstrated a strong reprogramming activity of BRD4::NUTM1. The new mouse model provided a critical preclinical model for NC and opens new opportunities for understanding the oncogenic mechanism and developing new therapies.

2. Introduction:

NUT carcinomas (NC), formally named NUT Midline Carcinomas (NMC), are poorly-differentiated squamous cell carcinomas that often arise within midline organs[1-3]. Other primary sites, including skin[4], pancreas[5], pelvic organs[6,7], and soft tissues[8,9], have also been reported. They are aggressive tumors that typically present with metastatic disease at diagnosis[10]. The prognosis of NC is dismal, with an overall survival time of

50 6.5 months, even with intensive treatment, including surgery, radiation, and
51 chemotherapy[10]. NC is rare; a recent estimation suggests that NC accounts for
52 approximately 0.21% of all cancers, with an estimated 3500 new cases emerging in the
53 U.S. each year[11]. However, because NC lacks a distinctive diagnostic feature and is not
54 widely recognized by many clinicians around the world, the case rates are likely
55 underestimated. NC is more prevalent in young people, with a median incidence age in
56 the 20s; it causes a dramatic loss of life[12]. Thus, it is critical to understand the
57 pathogenesis of NC, which will lead to better treatment for NC patients.
58 Clinical studies have established a strong association of NC with chromosome
59 translocation events that produce fusion genes involving a testis-specific gene known as
60 Nuclear protein in testis (NUTM1)[13], with various epigenetic factors, including BRD4
61 (70% of cases), and BRD3, NSD3, ZNF532, [1,4,9,14-17]. Limited genomic sequencing and
62 karyotyping analysis suggested that these fusion genes, likely acting as an abnormal
63 epigenetic modifier, can drive NC without other mutations in a quiet genomic
64 landscape[18]. At the molecular level, it was suggested that, by recruiting the P300
65 acetyltransferase (HAT) to ectopic sites, BRD4::NUTM1, and the other fusion proteins,
66 drive malignant transformation by activating oncogenes and epithelial progenitor genes
67 such as *cMYC*, *SOX2* and *TP63*[19-23].
68 However, when ectopically expressed in non-NC cell lines, *BRD4::NUTM1* is
69 unequivocally cytotoxic[2,24,25], contradictory to the proposed oncogenic functions. To
70 reconcile these contradictory observations, we consider two possible explanations.
71 First, many fusion oncogenes have been shown to require a specific cell lineage status
72 to achieve cell transformation and are otherwise toxic to other cell types. This is coined

as a concept of 'cellular pliancy' first in pediatric cancer and then in cancer in general[25-27]. Second, it could simply be that the over-expression system used on cell lines expresses the *BRD4::NUTM1* oncogene at a too-high level. It is not uncommon to observe gene dosage-dependent toxicity of oncogenes[28,29]. Thus, the discrepancy in expression levels could be a simple explanation for the cytotoxicity of the *BRD4::NUTM1* oncogene.

To recapitulate the NC oncogenesis in an experimental model in which both critical parameters are controlled, we developed a genetically engineered mouse model (GEMM). Using cre-lox-mediated chromosome translocation, the model ensured the recapitulation of human-mimicking chromosome translocation and established syntenic control of the resultant *Brd4-Nutm1* fusion gene. Taking advantage of the intrinsic low frequency of cre-lox mediated interchromosome translocation[30-33], the model recapitulated the stochastic nature of fusion gene formation. Using various cre-drivers, we demonstrated a broad oncogenic activity of BRD4-NUTM1 across tissues, covering all reported sites of human NC. BRD4-NUTM1 drives aggressive squamous cell carcinoma with various degrees of differentiation, consistent with human NCs. Further, the model demonstrated a strong intrinsic metastasis potential of NC, recapitulating a key feature of human NC. Thus, we reported the first GEMM that recapitulates the intrinsic genetic lesion and oncogenic process of human NC. The model will create tremendous opportunities for understanding and developing treatments for this devastating disease.

3. Results:

3.1 *BRD4::NUTM1* induces NC-like tumors from epithelial progenitor cells:

96 We designed and developed a mouse model to model the NC initiation event using
97 2C-HR-CRISPR[34] (Figure 1A). Loxp sites were inserted into the appropriate introns of
98 the mouse *Brd4* and *Nutm1* genes to allow stochastic induction of a reciprocal t(2;17)
99 chromosome translocation forming the Brd4-Nutm1 fusion gene by Cre recombinase.
100 To track the development of NC *in vivo*, we inserted a Luc2TdTomato reporter cassette
101 downstream of the endogenous *Nutm1* coding sequence, separated by a T2A self-
102 cleaving sequence[35]. The Luc2TdTomato reporter expressed specifically in *Nutm1*-
103 expressing post-meiotic spermatids as expected (Supplementary Figure 1 A and B). For
104 brevity, mice carrying both loxp sites and the reporter cassette linked to $Nutm1^{loxp}$ are
105 designated NC translocators (NCT). From this point, unless otherwise explained,
106 $NCT^{+/wt}$ refers to $Brd4^{loxp/wt}$; $Nutm1^{loxp/wt}$ whereas $NCT^{+/+}$ refers to $Brd4^{loxp/loxp}$;$Nutm1$
107 $^{loxp/loxp}$.
108 Because around 40% of human NCs are poorly differentiated, squamous cell
109 carcinoma (SCC) arose in the head-and-neck region and expressing markers
110 reminiscent of basal progenitor cells[3,36-38], we first tested NC induction using the
111 *Krt14Cre* allele that expresses the Cre recombinase in ectoderm-derived basal
112 progenitor cells[39] (supplementary Figure 1C). At weaning age, $KRT14Cre^{+/-}$; $NCT^{+/wt}$
113 mice developed BLI-positive head tumors at 100% penetrance (Figure 1B and
114 Supplementary Figure 1D)(n>50 mice), with some mice also developing skin tumors.
115 These tumors grew exponentially (Figure 1D) and rapidly led to a moribund state
116 (followed by humane euthanasia) (Fig. 1E), consistent with the aggressive nature of
117 human NC, likely due to the large tumor burden in the oral cavity causing obstruction of
118 the upper airway and digestive tracts. Upon autopsy, tumors were identified in the oral

119 cavity (mandible, maxilla, and tongue), salivary glands, eyelids, ear canal, and skin
120 (Supplementary Figure 1E); all of these locations have been reported in human patients
121 and consistent with the expression pattern of Krt14Cre.
122 We validated the formation of the t(2;17) chromosome translocation by 1)duo color
123 fluorescent in situ hybridization (DC-FISH) (Figure 1F); 2) Sanger sequencing of
124 junction sequence (Supplementary Figure 1F); 3) karyotyping (Supplementary Figure
125 1G); and 4) fusion calling in whole genome sequencing (Supplementary figure 1H).
126 Furthermore, immunohistochemistry (IHC) analysis against NUTM1 detected pervasive
127 nuclear expression of the BRD4-NUTM1 protein in the tumors (Figure 1G and
128 Supplementary Figure 1I). No BRD4-NUTM1 was expressed in NC surrounding normal
129 epithelial or tumor-associated stromal cells (Figure 1G), confirming the low frequency
130 and stochastic nature of the chromosome translocation. The above-described data
131 suggest that the t(2;17) chromosome translocation that creates *Brd4::Nutm1* can induce
132 aggressive tumors from mouse epithelial basal cells. Because the mouse tumors satisfy
133 two key diagnostic criteria applied to humans – the detection of the chromosome
134 translocation by FISH and the detection of BRD4::NUTM1 proteins by IHC[15,40], we call
135 these tumors mouse NUT Carcinoma (mNC).
136 **3.2 Mouse NCs Present Histopathological and Molecular Features of Human NC.**
137 Most human NCs are categorized as poorly differentiated squamous cell carcinomas,
138 characterized by primitive round cells with variable areas of abrupt squamous
139 differentiation and keratinization[11]. This heterogeneous phenotype has caused profound
140 confusion in diagnosing NC. Also, the intra and inter-tumor heterogeneity can create a
141 plasticity space for therapy resistance. As shown in Figure 2A, mNCs consist of large

142 areas of basaloid, polygonal, or spindle-shaped cells. Many tumors exhibit areas of
143 squamous cell differentiation and keratinization that occupy variable proportions of the
144 tumor. These features are consistent with the histopathological observations in the
145 clinical setting. Furthermore, a board-certified veterinary pathologist scored 50 cases of
146 mouse NC arising in the head-and-neck and skin sites and concluded that the mouse
147 NCs present a wide range of degrees of squamous cell differentiation (Figure 2B),
148 recapitulating the human pattern. Notably, even within a single mouse, individual
149 tumors can display variable morphology (Figure 2C), demonstrating intra and inter-
150 tumor heterogeneity within the same genetic background. Thus, mNCs reproduced the
151 morphological heterogeneity of human NC.

152 A few key molecular features define human NC. First, the BRD4-NUTM1 proteins
153 form nuclear foci, possibly as biological condensates mediated by the highly disordered
154 sequences in NUTM1, in human NC[41,42]. We detected such foci in mNCs through both
155 Immunofluorescence (IF) and IHC (Figure 2D and Supplementary Figure 2A). Second,
156 human NC tissues and cell lines demonstrated nuclear foci of Histone 3 lysine 27
157 Acetylation (H3K27ace), likely caused by the recruitment of P300 Histone
158 Acetyltransferase (HAT) by NUTM1[19]. We detected such foci in mNCs through IF
159 (Figure 2E and Supplementary Figure 2B). Third, human NCs express other hallmark
160 genes, including the basal progenitor cell markers P63, Krt14, SOX2, and oncogene
161 cMYC. These genes have also been implicated as required for the progression of
162 NC[20,22,23]. The majority of oral mNC cells express P63, Krt14, and cMYC (Figure 2F,
163 Supplementary Figure 2C). However, SOX2 is expressed in a small proportion of NC
164 cells with a clustered pattern in all tested mouse NCs (Figure 2F and Supplementary

165 Figure 2C). In clinical pathology practice, SOX2 is considered a stem cell marker in NC
166 and is expressed in various proportions of NC cells[23,43] , consistent with the mouse
167 data. Whether the SOX2-positive NC cells represent a cancer stem cell population is a
168 critical question to be investigated in the future.

169 **3.3 mNCs are aggressive poorly differentiation cancer at the transcriptomics**
170 **level:**

171 To gain a global understanding of how transcriptomics reprogramming promoted
172 aggressive mNC, we performed bulk RNA sequencing analysis on six normal oral
173 mucosa (control) and six oral mNC (tumor) samples. As shown in Supplementary Figure
174 3 A-C, the RNA sequencing data is of high quality, with a clear separation of the six
175 normal oral mucosa and the six mouse NC repeats. The analysis revealed 3065
176 upregulated genes and 2654 downregulated genes in mNCs, demonstrating pervasive
177 transcriptomics reprogramming (Figure 3A). Gene set Enrichment Analysis (GSEA)
178 against the Hallmark database revealed the activation of many pathways that may
179 contribute to an aggressive phenotype, including cell proliferation and mitosis, DNA
180 repair, MYC target pathways, and metabolic reprogramming (Figure 3B). Notably,
181 downregulated genes are specifically enriched for immuno-stimulatory pathways (Figure
182 3B).

183 To understand whether the poorly differentiated morphological traits are represented
184 at the transcriptomics level, we used CIBERSORT to deconvolute the bulk RNA
185 sequencing data using published single-cell RNA sequencing (scRNA-seq)
186 signatures[44,45]. As reference data, we used the published scRNA-seq data on the oral
187 mucosa epithelial from the *Tabula Muris* project[46], which represents the cells from which

188 the NC tumors are derived. As shown in Figure 3C, compared to the normal oral
189 mucosa samples, there is a strong enrichment of the proliferating basal cell expression
190 profile in the NC samples. CIBERSEQ analysis using another independently produced
191 mouse oral mucosa scRNA-seq reference data produced similar result[47]
192 (Supplementary Figure 3D). The CIBERSORT results are consistent with the
193 histological features of these tumors; for example, case K0329020 showed more
194 differentiated histology, while K0424100NC is more undifferentiated (Figure 3D). At the
195 molecular level, the overwhelmingly enriched proliferating basal cell profile is consistent
196 with the pervasive expression of basal cell markers (Figure 2). Consistent with its highly
197 proliferative nature, mNC cells overwhelmingly expressed the cell proliferation marker
198 Ki67 (Figure 3E). In addition, the differentiated keratinocyte marker KRT13 was only
199 detected in small regions with apparent keratinization (Figure 3F). Thus, the
200 *Brd4::Nutm1* expression trapped most NC cells in a proliferative progenitor status, likely
201 contributing to its aggressive nature.
202
203 **3.4. *BRD4::NUTM1* can induce aggressive malignancy from a broad range of**
204 **tissues.**
205 In humans, although most NCs were detected within the head and neck or thoracic
206 tissues[3], there were reports of carcinoma carrying the *BRD3/4-NUTM1* fusion genes in
207 many other epithelial tissues, such as the pancreas[5], the kidney[48], and the pelvic
208 organs[7]; and even mesenchymal tissues as undifferentiated soft tissue tumors[8]. This
209 broad tissue distribution is a unique property of NC compared to many other fusion
210 gene-driven cancers and indicates that the *BRD4::NUTM1* fusion gene may transform a

211 broad range of cell types covering all three germ layers. However, due to the limited
212 number of cases, it is unclear whether the reported *NUTM1* fusion genes are the cancer
213 driver or background mutations irrelevant to oncogenesis in these cases. We tested it
214 using the GEMM.

215 There are two cases of NC involving the pancreas reported in very young children[5].
216 The congenital nature of these cases suggests that an embryonic progenitor was
217 transformed. Thus, we crossed the Pdx1Cre allele[49], which drives recombination in
218 endoderm-derived embryonic ductal progenitor cells into the NCT mouse line.
219 Pdx1Cre;NCT$^{+/-}$ mice developed pancreatic tumors with 100% penetrance(18/18)
220 (Figure 4A, B and supplementary Figure 4A). The pancreatic mNCs are fast-growing
221 SCC (Supplementary Figure 4B) presenting all histopathological and molecular markers
222 as oral mNCs (Figure 4C and Supplementary Figure 4B). Beyond the epithelial tissue,
223 a few cases of soft tissue tumors have been reported to carry BRD3/4-NUTM1 fusion
224 genes, suggesting that the fusion gene can also collaborate with mesoderm-derived
225 mesenchymal gene regulatory networks (GRNs) to drive NC[8]. To test this, we used a
226 Prrx1Cre to drive the chromosome translocation in the fetal limb mesenchyme cells[50].
227 These mice developed fast-growing tumors within limbs at 100% penetrance (9/9)
228 (Figure 4D, E and Supplementary Figure 4C). The tumors are also fast-growing SCC
229 presenting all histopathological and molecular markers of NC (Figure 4F and
230 Supplementary Figure 4D). The ability to induce SCC from pancreatic ductal progenitors
231 that normally initiate adenocarcinoma and mesenchymal progenitors that normally
232 initiate sarcomas demonstrated a strong reprogramming activity of BRD4::NUTM1[49,51].

233 The clonal and stochastic nature of the translocation induction provided a unique
234 advantage of modeling the reality of stochastic gaining of the fusion gene across tissues
235 in humans. We crossed the NCT mouse line with the NLS-Cre mouse line that
236 constitutively expresses Cre in all mouse tissues[52]. This cross will induce
237 *BRD4::NUTM1,* forming chromosome translocation stochastically across tissues. The
238 NLS-Cre; NCT$^{+/-}$ mice developed tumors detected by BLI signals at high penetrance
239 (24/25) (Supplementary Figure 5A). Autopsy guided by BLI imaging identified tumors in
240 a broad range of anatomical locations, including the head and neck region (oral
241 mucosa, paranasal sinuses, tongue, salivary glands), thoracic cavity, stomach, kidney,
242 ovary, skin, and limb soft tissues (Figure 5A), covering all reported anatomical locations
243 of human NC. Most of these tumors were diagnosed as SCC with differing degrees of
244 differentiation (Figure 5B and C and Supplementary Figure 5B and 5C), confirming a
245 consistent phenotype. Notably, although the NLC-Cre driven mNCs consistently
246 express NUTM1, cMYC and KRT14, the expression of P63 and SOX2 are variable
247 (Figure 5B and C). The inter-tumor heterogeneity of these progenitor transcription
248 factors is consistent with pathological observations in human patients and may be
249 responsible for the remarkable plasticity and resistance to therapies of NC. Critically,
250 atypical histological phenotypes were reported in isolated human NC cases, including
251 glandular and chondrogenic metaplasia[53]. Consistently, regions of chondrogenic and
252 glandular transitions were observed in some mNCs (Figure 5D). The new GEMM
253 provided a critical tool for investigating the reprogramming capability of *Brd4::Nutm1* to
254 drive NC across tissues with remarkable heterogeneity.

255 **3.5. Mouse NCs have high metastasis potential.**

256 Most human NC patients have metastatic disease at diagnosis, contributing
257 significantly to its dismal prognosis[2,54,55]. Of the more than 100 oral NC mice we
258 analyzed, there were four cases of metastasis to proximal lymph nodes (Figure 6A),
259 suggesting metastatic potentials of mouse NC. However, because the oral NCs lead to
260 a quick death of mice, they may not allow the time for metastasis to establish and
261 emerge. To further investigate, we focused on analyzing the Pdx1Cre driven pancreatic
262 NC, which survives longer. Pancreatic NC mice consistently develop broad
263 dissemination and metastasis at ten weeks of age. Large numbers of dissemination
264 nodules (carcinomatosis) develop across the mesentery membrane (Figure 6B and
265 Supplementary Figure 6A). Distant Micro-metastasis was detected in the lungs and liver
266 (Figure 6B). The metastatic tumors express NC markers, including NUTM1, cMYC, P63,
267 KRT14, and contain populations of SOX2-positive cells (Figure 6C).
268 It is unclear why NCs are extremely metastatic. Desmoplasia, characterized by the
269 recruitment of cancer-associated fibroblasts and pervasive deposition of extracellular
270 matrix proteins, is strongly associated with metastatic potential in common cancers such
271 as pancreatic ductal adenocarcinoma (PDAC)[56-58]. Desmoplastic histological features
272 have been reported in a number of human NC case reports and are routinely observed
273 in clinical diagnostic practice[59-61], yet, have never been characterized at the molecular
274 level. Through RNA sequencing analysis of mouse NC, we detected large increases in
275 the expression of fibroblast genes such as *fibroblast activation protein (FAP)*, *smooth*
276 *muscle alpha actin (Acta2)* and SM22alpha (*Tagln*) (Figure 7A). Mouse NC tissue also
277 express high levels of ECM genes including *fibronectin (Fn1)*, *collagens* (17 types) and
278 collagen cross-linking enzymes such as *lysyl oxidase (Lox)*, *Loxl2, Loxl3* and

279 *peroxidasin (Pxdn)* (Figure 7B). Confirming these results, both mouse oral and
280 pancreatic NC contain large numbers of activated fibroblast cells and abundant
281 collagen-based ECM (Figure 7C). In addition to the unique desmoplastic stromal
282 response, the epithelial-mesenchymal transition (EMT) process has been closely linked
283 to metastasis[62]. A gene set enrichment analysis (GSEA) revealed significant activation
284 of the Epithelial-mesenchymal transition program in mouse NCs (adjusted *p*-value =
285 2.16E-07 or 2.16×10^{-7}), with strong upregulation of mesenchymal markers including
286 Snai1[63], Zeb1[64] and Vimentin[65,66] (Figure 7D). Consistently, at an advanced stage,
287 regions of both mouse oral and pancreatic NC lost the membrane-localized epithelial
288 marker E-Cadherin and gained the expression of Zeb1 and Vimentin to differing
289 degrees (Figure 7E), demonstrating EMT, which might be associated with the
290 metastatic potential of NC. It is clear now that EMT is a reversible process marking a
291 continuous spectrum of phenotypes[67]. However, it is under debate whether full EMT is
292 required for metastasis and whether a reverse process called Mesenchymal Epithelial
293 Transition (MET) is required for the establishment of metastasis[62]. We analyzed EMT
294 phenotypes of the metastatic lesions in mouse NC. As shown in Figure 7F, in some
295 cases such as the lymph node metastasis of oral NC, the majority of cancer cells within
296 the metastasis tumor expressed E-Cadherin but also vimentin and Zeb1, showing a
297 mixed Epithelial-mesenchymal phenotype. On the other hand, the mesenteric nodules
298 developed from pancreatic NCs have lost membrane localized E-cadherin but did not
299 significantly upregulate Vimentin and ZEB1. Thus, there is likely a complex, non-binary
300 interplay between desmoplasia, EMT, and metastasis of NC (more examples in

301 supplementary Figure 6). The new GEMM provides a powerful tool to dissect these
302 interplays.

303 **4. Discussion:**

304 Nut carcinoma is an aggressive cancer with a dismal prognosis[2,10]. Because NC is
305 extremely rare in the human population and is often diagnosed at a late disease stage,
306 relevant *in vivo* models are vital for understanding its pathogenesis and for therapy
307 development[20,25]. The new NC GEMM addressed key issues that are challenging to
308 explore with currently available human cell line and xenograft models and created new
309 opportunities for understanding NC.

310 First, a long-standing dilemma in NUT carcinoma research is that although the
311 disease is strongly associated with chromosome translocations events that induce the
312 expression of NUTM1 fusion genes such as *BRD4::NUTM1*, the ectopic expression of
313 the fusion genes in non-NC cells are unequivocally cytotoxic[2,24]. This dilemma raises a
314 critical question as to whether the *NUTM1* fusion genes are indeed responsible for NC
315 initiation. Through analyzing hundreds of NC GEMM, our data demonstrated that when
316 expressed at the endogenous level from an endogenous chromosome translocation,
317 *BRD4::NUTM1* can efficiently induce aggressive cancers with high penetrance. The
318 mouse tumors that BRD4::NUTM1 induced presented similar histopathological and
319 molecular features of human NCs. mNCs also reproduced the intrinsic intra and inter-
320 tumor heterogeneity. This comprehensively characterized GEMM firmly demonstrates
321 that the *BRD4::NUTM1* fusion gene is sufficient to induce NC-like tumors *in vivo*.

322 Second, although human NC is rare, cases have been reported in a broad range of
323 tissues. The main affected sites were reported to be head and neck (~40% of reported

324 cases) and thoracic (~50% of reported cases)[3]. There are two possible explanations for
325 the predilection for these sites: 1) The cell types in these tissues are more receptive to
326 transformation by *BRD4::NUTM1*. 2) Fewer cases were diagnosed at other sites due to
327 limited clinical awareness. The second explanation is not trivial, considering the history
328 of NUT carcinoma. It was first thought to be a pediatric thymic cancer[1]. With the
329 increased awareness and better availability of diagnostic methods such as the NUTM1
330 IHC and next-generation sequencing, NC is now considered a preferentially young adult
331 cancer that can occur in any age group and many tissue sites[3]. Our data showed that
332 when *Brd4::Nutm1* fusion was induced in tissue progenitors of all three germ layers
333 using cell type-specific cre drivers, NC can be induced at 100% penetrance in the
334 respective tissue. Furthermore, when *BRD4::NUTM1* fusion was induced stochastically
335 using a constitutively active Cre driver, mNCs can arise in many tissues, such as the
336 paranasal sinuses, ovary and kidney, previously reported in humans, and previously not
337 reported tissues, such as the gastric tissues. Our data raised a critical possibility that
338 the low case rate outside the head and neck/thoracic region might be a result of
339 diagnosis bias, either due to limited clinical awareness or in the case of gastric cancer,
340 the low general prevalence and research intensity in North America[68]. Broader and
341 global surveys of NUT Carcinoma cases in hard-to-diagnose cancers with monolithic,
342 poorly differentiated morphology in all tissues may establish a more complete picture of
343 NC in humans. At the molecular level, the broad cell type receptivity and the universal
344 SCC phenotype implied that BRD4::NUTM1 coerces a shared molecular program in cell
345 types across all germ layers to initiate NC. Yet, BRD4::NUTM1 has a strong
346 reprogramming ability to drive an aggressive poorly differentiated SCC phenotype of NC

347 regardless of lineage context. The new GEMM provided a unique model to further
348 investigate the key molecular mechanisms underlying NC oncogenesis.
349 Fourth, human NCs are highly metastatic, yet, currently, there is no established
350 model to study this key aggressive phenotype. The GEMM uncovered an extremely
351 strong metastatic potential of NC. GEMMs are not always optimal for studying
352 metastasis. In many GEMMs of common cancers, mice reach their endpoint due to the
353 growth of primary tumors before significant metastasis[69]. In GEMMs that consistently
354 develop metastasis, such as the mutant K-Ras driven pancreatic ductal carcinoma
355 (PDAC) model, metastasis takes a long time to develop (15-40 weeks) and is
356 challenging to study[49,70]. The pancreatic NC GEMM develops widespread metastasis in
357 a short time window (5-10 weeks), providing ample materials to investigate the
358 mechanisms of NC metastasis in vivo. Critically, the primary mouse NCs do present
359 phenotypes, including desmoplasia and EMT that are associated with metastatic
360 potentials. Further investigation using the new GEMM will lead to a fruitful
361 understanding of the metastasis mechanism and therapeutic strategies of NC.
362 As immune therapy has transformed the therapeutic landscape of aggressive cancers
363 such as melanoma and lymphoma, it is critical to investigate the immune system's
364 interaction with targeted or immune therapies for NC. The GEMM reported here is the
365 first immune-competent experimental model for NC and can provide a valuable platform
366 for studying the immune therapies for NC.
367 After our preprint was online, another NC mouse model was submitted and published
368 [71]. This model uses the *cre*-dependent one-way genetic switch (FLEx) conditional
369 inversion system and a Sox2-CreERT2 tamoxifen driver to induce the *Brd4::Nutm1*

370 fusion gene from the mouse Brd4 allele. In parallel, these two GEMM each have unique
371 advantages and disadvantages in modeling NC. First, the FlEx model is driven by
372 Tamoxifen inducible Cre, thus have better temporal control. However, the high
373 recombination efficiency will mean that the fusion gene will be induced in a large
374 number of cells simultaneously, mimicking the sequencing mutagenesis and field
375 carcinogenesis in common cancers[72], different from the clonal stochastic reality of
376 fusion gene driven cancer. Our translocation model recapitulates such a reality.
377 Second, the FlEx model induces mNCs only in the esophagus, providing a clear and
378 easily accessible target tissue for study. However, there has been no report of primary
379 NC in the esophagus. Whether this inconsistency is due to species difference, miss-
380 assigning of human primary NC, or an artifact due to pervasive induction of
381 *Brd4::Nutm1* in all esophagus basal cells by the Sox2-CreERT, remains to be
382 determined. In contrast, our translocation model closely recapitulated the human tissue
383 spectrum, including the rare cases outside the head-and-neck and thoracic regions,
384 providing a richer resource understanding of the NC oncogenesis from different tissues.
385 Furthermore, the broad tissue distribution of the translocation model will allow
386 intersectional analysis to find the core mechanisms that drive NC and model more
387 tissue contexts, which will be critical for future drug delivery and immune therapy
388 research.
389 In summary, we have reported the first GEMM for NC that recapitulates the
390 endogenous chromosome translocation. The new GEMM can enhance our
391 understanding of this disease and open up many opportunities for developing targeted

392 therapies. The GEMM will serve as a valuable preclinical tool for the global community
393 to study the mechanism of NC and develop new treatments.
394
395 **4. Materials and Methods:**
396 **4.1 Ethical statement:**
397 All animal work was carried out under PROTO202000143 and PROTO 202300127,
398 approved by the Michigan State University (MSU) Campus Animal Resources (CAR)
399 and Institutional Animal Care and Use Committee (IACUC) in AAALAC credited
400 facilities.
401
402 **4.2 Mouse line generation:**
403 All mouse generation work follows our published 2-cell (2C)- Homologous
404 Recombination (HR)- CRISPR (2C-HR-CRISPR) method and protocol and is described
405 briefly as follows[34,73,74]. First, 2-cell stage embryos of the CD1 strain were microinjected
406 with Cas9-mSA mRNA (75ng/ul), sgRNA targeting around the stop codon of the Nutm1
407 gene (GGUGGCCCUCUGCUUCCUAC, chemically modified sgRNA from Synthego Inc.
408 50ng/ul) and a biotinylated PCR donor consisted of homology arms (5'arm 1000bp and
409 3'arm 600bp) spanning the coding sequence of T2A-Luc2TdTomato cassette (5ng/ul).
410 The Luc2TdTomato cassette was cloned from the pCDNA3.1(+)/Luc2-TdT plasmid
411 (Addgene 32904, kind gift from Christopher Contag)[35]. Injected embryos were implanted
412 into pseudopregnant females the same day to generate founder pups. Positive founders
413 were screened by long-range PCR spanning homology arms using the following primer
414 sets. 5' arm gtF: GAGACTGTCATAGACAGCATCCAAGAT and 5'arm gtR:

415 tcatggctttgtgcagctgc; 3’arm gtF: ggactacaccatcgtggaacagta and 3’arm gtR:
416 CATATTTAACAGTCCCACGGAGAG. A positive founder mouse was out-crossed to
417 CD1 mice to generate N1 mice. The N1 mice were genotyped by PCR. Additionally,
418 genomic regions spanning the targeting cassette and 3’ and 5’ homology arms were
419 Sanger-sequenced to validate correct targeting. Insertion copy number was evaluated
420 by digital droplet PCR ddqPCR. Heterozygous N1 mice have only one insertion copy,
421 demonstrating single-copy insertion. An N1 that passed all quality controls was further
422 outcrossed with CD1 for two generations to remove any potential off-target mutations
423 introduced by CRISPR-Cas9 editing to generate N3. Heterozygous N3s were bred
424 together to establish a homozygous colony of the Nutm1-T2A-Luc2TdTomato
425 mousseline.
426 Then two additional mouse lines were generated in parallel. First, a loxp site was
427 inserted into the intron 10 of the Brd4 gene (Brd4loxp) by electroporating CD1 zygotes
428 with Cas9 proteins, a chemically modified sgRNA (GAAUAGAGUGCUGCUACACU,
429 Synthego Inc.), and a single strand oligo DNA donor (ssODN), with 30 base homology
430 arms on each side spanning the loxp site
431 (TCTCCCTGCAGAGTGGCAAGATGCCCAGTGataacttcgtatagcatacattatacgaagttatTAG
432 CAGCACTCTATTCTCTCCATCTGGGTA). Electroporated zygotes were implanted into
433 pseudopregnant females the same day to generate founder pups. Positive founder pups
434 were screened by PCR using primers: BRD4 loxp gt F:
435 TGCCTGGGCTCTTCTGTCCAC, and BRD4 loxp gt R:
436 CAACCCAAGGTTGATGGGACCC. A positive founder mouse was out-crossed to CD1
437 mice to generate N1 mice. The N1 mice were genotyped by PCR. Additionally,

438 genomic regions spanning the targeting cassette and 3’ and 5’ homology arms were
439 Sanger-sequenced to validate correct targeting. An N1 that passed all quality controls
440 was further outcrossed with CD1 for two generations to remove any potential off-target
441 mutations introduced by CRISPR-Cas9 editing to generate N3. Second, a loxp site was
442 inserted into the intron 1 of the Nutm1-T2A-Luc2TdTomato gene (Nutm1-loxp) by
443 microinjecting Nutm1-T2A-Luc2TdTomato homozygous 2-cell embryos. (75ng/ul),
444 sgRNA targeting around the intron (ACCGGAUCCCAACCGCUCUA, chemically
445 modified sgRNA from Synthego Inc. 50ng/ul) and a biotinylated PCR donor consisting of
446 homology arms (5’arm 500bp and 3’arm 500bp) spanning a loxp site and artificial
447 genotyping primer sequences (5ng/ul). Injected embryos were implanted into
448 pseudopregnant females the same day to generate founder pups. Positive founders
449 were screened by long-range PCR spanning homology arms using the following primer
450 sets. 5’ arm gtF: GCATGAAATCAGACCAAGTGGGC and 5’arm gtR:
451 AAGCTGACCCTGAAGTTCATCTG; 3’arm gtF: ggactacaccatcgtggaacagta and 3’arm
452 gtR: CATATTTAACAGTCCCACGGAGAG. A positive founder mouse was out-crossed
453 to CD1 mice to generate N1 mice. The N1 mice were genotyped by PCR. Additionally,
454 genomic regions spanning the targeting cassette and 3’ and 5’ homology arms were
455 Sanger-sequenced to validate correct targeting. An N1 that passed all quality controls
456 was further outcrossed with CD1 for two generations to remove any potential off-target
457 mutations introduced by CRISPR-Cas9 editing to generate N3. Starting from N3s, the
458 Brd4loxp and Nutm1loxp strains were bred to double homozygous to generate the NC
459 translocator (NCT) mouse line.

460 For inducing the BRD4::NUTM1 fusion, NCT+/+ mice were crossed with heterozygous
461 KRT14Cre mice (Jackson Lab 018964) for oral cancer, Pdx1Cre mice (Jackson Lab
462 014647) for pancreatic NC, Prrx1Cre mice(Jackson Lab 005584) for limb soft tissue NC
463 and NLS-Cre (MGI:3586452) for universal screen. The Cre mouse lines from Jackson
464 Lab was on the B6 background and had black fur. To facilitate sensitive bioluminescent
465 imaging, the strain was outcrossed for two generations with CD1 mice to generate
466 albino mice. Progenies from the crosses were genotyped; the cre-positive ones were
467 used as experimental groups, while the cre-negative ones were used as controls.
468 Bioluminescence imaging:
469 NC mice were subjected to regular bioluminescence imaging to monitor tumor signals.
470 During the imaging procedure, mice received an intraperitoneal injection of 150 mg/kg D-
471 Luciferin Potassium (Goldbio). After 20 minutes, images of both dorsal and ventral
472 positions were captured using IVIS Spectrum In Vivo Imaging System (Perkin Elmer)
473 under anesthesia. To quantify the tumor signal intensity, regions of interest (ROIs) were
474 delineated around the head area. The total flux (photons/sec) within the ROIs was
475 measured using the Living Image® Software. The average value of the total flux from
476 both dorsal and ventral images was used for generating tumor growth curves.
477
478 **4.3. Necropsy and tissue collection**
479 Routine necropsies were performed by ventral midline approach. Tumors, draining
480 lymph nodes, heads, lungs, and liver were harvested and fixed in 10% neutral buffered
481 formalin for 24 h or flash-frozen in liquid nitrogen for further analyses. Tissue samples
482 were either fixed in 4% paraformaldehyde at 4° C for 24 h or 10% neutral buffered

483 formalin at room temperature for 24 h. Heads were decalcified in 10% EDTA for 12-14
484 days at 4° C by replacing the decalcification solution every two days. Tissues were
485 either routinely processed, and embedded in paraffin wax or used for cryo-sections.
486 4.4. Histopathology
487 Serial, 5-µm thick sections were cut and stained with hematoxylin-eosin
488 (H&E). Neoplasms were characterized based on anatomic location, necrosis,
489 inflammation (polymorphonuclear and mononuclear), and histologic features (focal
490 dysplasia, undifferentiated carcinoma, poorly differentiated carcinoma with regional
491 squamous differentiation, and well-differentiated squamous cell carcinoma as
492 described[75]. Histopathologic examination was performed by a board-certified veterinary
493 pathologist (MFT).
494 **4.5. Immunohistochemistry**
495 Antigen retrieval was done using sodium citrate pH 6 for 20 min at 97 ^{o}C.
496 Immunohistochemistry was performed using antibodies targeting NUTM1 (1:100,
497 Huabio HA721690), TdTomato (1:500, Origene, ABOO40-500), cMYC (1:500,
498 Proteintech, 10828-1-AP), SOX2 (1:1000, R&D Systems, AF2018), TP63 (1:1000,
499 Sigma Aldrich, AMAB91224), KRT14 (1:500, Novus Biologicals, NBP2-67585), and
500 involucrin (1:1000, BioLegend, 924401) using Agilent Dako Autostainer Link 48 and
501 EnVision FLEX kit (Dako). Immunoreactivity for each antibody was assessed by light
502 microscopy.
503 **4.6. Cryosection:**

504 Tissues for cryosections were dehydrated using 30% sucrose at 4° C overnight.
505 Samples were embedded in OCT and snap-frozen on a steel bench block immersed in
506 liquid nitrogen. Samples were sectioned at 4-µm thickness using a cryostat (Leica).

507 **4.7. Immunofluorescence:**

508 Slides were dried at room temperature for 30 minutes and permeabilized using 0.1%
509 Tween 20 for 30 minutes. Samples were blocked using PBST containing 5% normal
510 donkey serum for 1 hour at room temperature. The primary antibodies (same as
511 described in the IHC section) diluted in blocking solution were applied on tissue sections
512 and incubated overnight at 4 ^{o}C. Slides were rinsed three times for 5 minutes each with
513 PBS. Secondary antibodies diluted in blocking solution were applied on tissue sections
514 and incubated in a dark humid chamber for 1 hour at room temperature. Slides were
515 rinsed three times for 5 minutes each with PBS. Nuclei were stained with Hoechst
516 33258 and then mounted with Vectashield® mounting medium. Slides were imaged with
517 a Leica Thunder microscope imaging system.

518 **4.8. Duo color FISH:**

519 The Oligopaint FISH probe libraries were constructed as described previously (Xie et
520 al., Nature Methods, 2020). A ssDNA oligo pool was ordered and synthesized as
521 custom arrays from GenScript. Each oligo consists of a 30-37 nucleotide (nt) homology
522 to the 30kb region, flanking the mBrd4-mNut1m fusion point from the mm10 genome
523 assembly. Each Oligo was run through the OligoArray2.0 with the parameters -n 22 -D
524 1000 -l 32 -L 32 -g 52 -t 75 -T 85 -s 60 -x 60 -p 35 -P 80-m "GGGGG;CCCCC;
525 TTTTT;AAAAA" according to algorithms developed from the Wu laboratory
526 (https://oligopaints.hms.harvard.edu/). The mBrd4 and mNut1m oligo subpool consists

527 of a unique set of primer pairs for PCR amplification, a 20 nt T7 promoter sequence for
528 in vitro transcription, and a 20 nt region for reverse transcription. The mBrd4 oligo pool
529 was amplified by GCGGGACGTAAGGGCAACCG and GCGTTGCGGTGCGATCTTCT.
530 The mNut1m was amplified by TGATAACCCACGCACGGCTG and
531 GACCCGGGCCACTAACACGA. Following PCR amplification, each Oligopaint probes
532 were generated by in vitro transcription and reverse transcription, in which ssDNA oligos
533 conjugated with ATTO565 and ATTO647 fluorophores were introduced during the
534 reverse transcription step. The Oligopaint covered genomic regions (mm10) used in this
535 study are listed below: mBrd4, chr17:32,210,998-32,244,827. mNut1m,
536 chr2:112,227,883-112,257,004. The oligo sequences for conjugating fluorophores
537 during reverse transcription are: /5ATTO565N/AATTCGGCAGACCCGAATGC
538 (mNut1m) and /5ATTO647N/CCTATCCCCTGTGTGCCTTG (mBrd4).
539
540 DNA FISH is performed on cryosections prepared as described above. Tissue slides
541 were washed 2X with 1XPBS and permeabilized in 0.5% TritonX100 in 1XPBS for
542 60min in a Coplin Jar. After 2X wash in 1XPBS, slides were treated with 0.1M HCl for
543 5min, followed by 3X washes with 2XSSC and 30 min incubation in 2X SSC + 0.1%
544 Tween20 (2XSSCT) + 50% (v/v) formamide (EMD Millipore, cat#S4117). For each
545 sample, we prepare a 25ul hybridization mixture containing 2XSSCT+ 50% formamide
546 +10% Dextran sulfate (EMD Millipore, cat#S4030) supplemented with 0.5µl 10mg/mL
547 RNaseA (Thermo Fisher Scientific, cat# 12091021) +0.5µl 10mg/mL salmon sperm
548 DNA (Thermo Fisher Scientific, cat# 15632011) and 20pmol probes with distinct
549 fluorophores. The probe mixture was thoroughly mixed by vortexing and briefly

550 microcentrifuged. The hybridization mix was transferred directly onto the slide covered
551 with a clean coverslip. The coverslip was sealed onto the slide by adding a layer of
552 rubber cement (Fixo gum, Marabu) around the edges. Each slide was denatured at
553 80°C for 5 min, then transferred to a humidified hybridization chamber and incubated at
554 42°C for 16 hours in a heated incubator. After hybridization, samples were washed 2X
555 for 15 minutes in prewarmed 2XSSCT at 60 °C and then were further incubated at
556 2XSSCT for 10min at RT, at 0.2XSSC for 10min at RT, at 1XPBS for 2X5min with DNA
557 counterstaining with DAPI. Then coverslips were mounted on slides with Prolong
558 Diamond Antifade Mountant (Thermo Fisher Scientific Cat#P36961) for imaging
559 acquisition.
560
561 DNA FISH images were acquired using a structured illumination microscopy on a Zeiss
562 AxioObserver microscope (Zeiss Elyra7). Images were taken with a Plan-Apochromat
563 63x/1.40 oil DIC objective in a lens immersion medium having a refractive index of
564 1.515 and two pco.edge 4.2 CL HS sCMOS cameras. We used 405nm (Excitation
565 wavelength) and 460nm (Emission wavelength) for the DAPI channel, 561nm
566 (Excitation wavelength) and 579nm (Emission wavelength) for the ATTO565 channel,
567 and 633nm (Excitation wavelength) and 654nm (Emission wavelength) for the ATTO647
568 channel. We acquired 13 phase images for each focal plane. Image post-processing
569 was performed with the ZEN 3.0 SR (black) software.
570 **4.9. RNA sequencing**
571 Total RNA was prepared from tumors and normal tissues using the RNeasy Mini Kit
572 (Qiagen Inc.). The quality control, cDNA library preparation and bulk RNA sequencing

573 were performed by Novogene Inc following standard protocol using the Novaseq 6000
574 sequencer. Each sample were sequenced for 20M reads.
575
576 FastQC v0.11.7[76] was used to evaluate sequence quality, GC content, overrepresented
577 sequences and adapter contamination of the raw paired-end FASTQ reads for each
578 sample. Trimmomatic v0.39[77] was used to trim raw reads using following settings: PE
579 ILLUMINACLIP: TruSeq3-PE.fa:2:30:10:8:true HEADCROP:15 CROP:110. All samples
580 passed quality control based on the results of FastQC. Quality trimmed reads were
581 pseudoaligned to GENCODE mouse transcript reference (GRCm39, release M32) and
582 transcription levels were quantified using kallisto[78] (version: 0.46.1, parameters: -b 100 -
583 l 110). Gene expression levels were generated using tximport v1.28.0[79] and filtered for
584 lowly expressed genes (0.5 count per million in at least six samples). Differential
585 expression analysis were carried out using the Bioconductor v3.17.1[80] package DESeq2
586 v1.40.2[81]. rlog-normalized counts were obtained from the differential expression analysis
587 and used for principal component analysis, sample correlation calculation and k-means
588 clustering[81]. The Wald test was used by DESeq2 to identify genes that are differentially
589 expressed. Benjamini-Hochberg was used to adjust the p values for multiple testing[82].
590 The log2 fold change (LFC) was shrunken by using the “apeglm” method to provide more
591 accurate estimates[83]. Differentially expressed genes (DEGs) were defined based on LFC >
592 1 or < -1 and adjusted *p*-value < 0.05. ClusterProfiler v4.8.2[84] was used to perform
593 functional enrichment analysis by using different databases: Gene Ontology[85,86], KEGG
594 PATHWAY[87], Reactome[88] and the Hallmark gene sets of Molecular Signatures Database
595 (MSigDb)[89]. Lists of downregulated DEGs and upregulated DEGs were separately

596 examined for Over Representation Analysis (ORA)[90]. The shrunken LFCs were used to
597 rank all DEGs decreasingly to generate a gene list as an input for Gene Set Enrichment
598 Analysis (GSEA)[91]. Redundant GO terms were removed by GOSemSim with a cut-off of
599 0.7[92]. Pathways or gene sets with adjusted p value < 0.05 were considered significantly
600 enriched. RNA-seq data will be available through GEO after publication.
601
602 **4.10. Whole genome sequencing:**
603 Genomic DNAs was extracted from normal and tumor samples using The GeneJet DNA
604 kit (Thermo Fisher), according to the manufacturer's instructions. The quality control,
605 cDNA library preparation and RNA sequencing were performed by Novogene Inc
606 following standard protocol. A total of 1.0 ug DNA per sample was used for library
607 preparations. Each sample were sequenced for 10-16X coverage. WGS data will be
608 available through GEO after publication.
609 Raw reads for each sample were trimmed using Trimmomatic v0.39[77] (PE
610 ILLUMINACLIP:2:30:10:8:true HEADCROP:5 CROP:120). Then the trimmed reads were
611 aligned to the GENCODE mouse primary genome assembly (GRCm39, release M32)
612 using the BWA-MEM algorithm in the Burrows-Wheeler Aligner (BWA) v0.7.17[93] with
613 following options: -M to flag shorter split hits as secondary; -R to add read group
614 information. The SortSam tool of Picard v2.25.0[94] was used to sort the output files from
615 BWA by queryname and export output as binary Sam (BAM) files. Subsequently,
616 duplicate reads were marked using markDuplicates tool in Picard and the resulting files
617 were sorted using SortSam by coordinates. The AlignmentSummaryMetrics and
618 CollectWgsMetrics tools of Picard were used to collect alignment and genome coverage

619 statistics. The Mutect2 tool of GATK v4.1.4.1[95] used to identify and call variants in each
620 sample individually. Variants were firstly filtered using FilterMutectCalls of GATK to
621 remove probable technical or germline artifacts. Additional filters can be used to decrease
622 the false-positive rate via SnpSift v4.1[96]: mutant allele frequency (≥5%), coverage at
623 particular positions in tumor and normal samples (≥5×) and supporting reads for the
624 mutation in the tumor samples (at least two). Filtered variants were further compared to
625 exclude known polymorphisms by using bcftools v1.9.64[97]. snpEff v4.1[96] was used to
626 anotate the effects of variants. Large structural variations were identified using Delly
627 v0.7.8[98].

628 **4.11. Cell-type deconvolution analysis**

629 The signature matrixes were separately generated from two published single cell RNA-
630 seq datasets for implementation in the CIBERSORTx deconvolution algorithm[99]. Briefly,
631 the preprocessed count matrix of mouse tongue from 10x genomics sequencing was
632 extracted from the Tabula Muris. Quality control, normalization, and clustering were
633 performed as described in the original study[46]. The transcripts per kilobase million (TPM)
634 gene expression profiles of 7538 labeled cells, pertaining to 7 distinct cell types
635 (proliferating basal cells, basal cells, suprabasal differentiating keratinocytes, suprabasal
636 differentiated keratinocytes, differentiated keratinocytes, filiform differentiated
637 keratinocytes and Langerhans cells), were used to build a signature matrix file
638 (q value = 0.1). The number of barcode genes per cell type was held between 300 and
639 500, resulting in a matrix consisting of 2256 genes. With the same strategy, the
640 preprocessed, mouse oral epithelial basal layer cells-derived single-cell RNA-seq dataset
641 from Jones KB, et al.[47] was also used to generate a signature matrix of with 8 cell

642 subtypes and 1970 genes included. The two signature matrixes were then applied to
643 deconvolute the bulk RNA-seq TPM matrix of mNC tumor and normal control to predict
644 the fraction of cell subtypes, with the absolute mode, S-mode batch correction, and 100
645 permutations.
646
647 **Karyotyping:**
648 Karyotyping analysis was performed on two early passage cell lines derived from
649 mouse oral NC by Karyologic Inc. following standard protocol.
650 **Software and statistics:**
651 The tumor growth and survival data were analyzed and visualized using Prism 9 software
652 (GraphPad, USA). Animal sacrifice was considered the endpoint for both analyses. The
653 median survival times were calculated utilizing the Kaplan-Meier analysis, and
654 significance values were determined by conducting the log-rank test. A p-value of less
655 than 0.05 defined statistical significance.
656
657 Author contribution:
658 D.Z: formal analysis, investigation, methodology, validation, visualization, writing review,
659 and editing;
660 A.E: formal analysis, investigation, methodology, validation, visualization, writing review
661 and editing;
662 C.L: formal analysis, investigation, methodology, validation, visualization, writing review
663 and editing;
664 M.B: methodology and visualization for FISH

L.X: methodology and visualization for FISH

D.B: Pathology analysis, human pathology insights, writing, review and editing;

Y.T: Human clinical insights, writing, review and editing;

E.H: Human clinical insights, writing, review and editing;

G.M: Bioinformatics analysis, writing, review and editing;

M.T: histopathological and IHC analysis, writing, review and editing;

B.G: conceptualization, design and generation of NC GEMM, formal analysis, investigation, methodology, validation, visualization, writing original draft, writing review and editing; providing funding support;

Funding: This research is supported by an R37 grant (R37CA269076) from the National Institute of Science (NIH)/National Cancer Institute (NCI) and startup funding both awarded to B.G.

Conflict of Interest:

The Authors declear no conflict of interest.

Acknowledgment:

We thank Dr. Andrew Sikora, Dr. Xiao Zhao and Dr. Jennifer Wang from the MD Anderson Cancer Center for constructive discussions. We thank MSU Transgenic and Genome Editing Facility (TGEF) for their help in generating the NC GEMM and the Histology laboratory at Veterinary Diagnostic Laboratory (Taylor Vaughn) for their help

687 in histopathology and IHC analysis. We would also like to thank MSU Precision Health
688 Program Tissue Analysis Core for assistance in slide scanning.

689
690
691 References:

692 1 Kubonishi, I. *et al.* Novel t(15;19)(q15;p13) chromosome abnormality in a thymic
693 carcinoma. *Cancer Res* **51**, 3327-3328 (1991).
694 2 French, C. A. Pathogenesis of NUT midline carcinoma. *Annu Rev Pathol* **7**, 247-265,
695 doi:10.1146/annurev-pathol-011811-132438 (2012).
696 3 Chau, N. G. *et al.* An Anatomical Site and Genetic-Based Prognostic Model for Patients
697 With Nuclear Protein in Testis (NUT) Midline Carcinoma: Analysis of 124 Patients. *JNCI*
698 *Cancer Spectr* **4**, pkz094, doi:10.1093/jncics/pkz094 (2020).
699 4 Nishimura, Y., Ryo, E., Yamazaki, N., Yatabe, Y. & Mori, T. Cutaneous Primary NUT
700 Carcinoma With BRD3-NUTM1 Fusion. *Am J Surg Pathol* **45**, 1582-1584,
701 doi:10.1097/PAS.0000000000001801 (2021).
702 5 Shehata, B. M. *et al.* NUT midline carcinoma in a newborn with multiorgan disseminated
703 tumor and a 2-year-old with a pancreatic/hepatic primary. *Pediatr Dev Pathol* **13**, 481-
704 485, doi:10.2350/09-10-0727-CR.1 (2010).
705 6 Ball, A. *et al.* A rare case of NUT midline carcinoma. *Gynecol Oncol Case Rep* **3**, 1-3,
706 doi:10.1016/j.gynor.2012.09.004 (2012).
707 7 Jung, M., Kim, S. I., Kim, J. W., Jeon, Y. K. & Lee, C. NUT Carcinoma in the Pelvic Cavity
708 With Unusual Pathologic Features. *Int J Gynecol Pathol* **41**, 292-297,
709 doi:10.1097/PGP.0000000000000801 (2022).
710 8 Dickson, B. C. *et al.* NUTM1 Gene Fusions Characterize a Subset of Undifferentiated Soft
711 Tissue and Visceral Tumors. *Am J Surg Pathol* **42**, 636-645,
712 doi:10.1097/PAS.0000000000001021 (2018).
713 9 Chien, Y. W. *et al.* Primary malignant epithelioid and rhabdoid tumor of bone harboring
714 ZNF532-NUTM1 fusion: the expanding NUT cancer family. *Genes Chromosomes Cancer*
715 **58**, 809-814, doi:10.1002/gcc.22785 (2019).
716 10 Bauer, D. E. *et al.* Clinicopathologic features and long-term outcomes of NUT midline
717 carcinoma. *Clin Cancer Res* **18**, 5773-5779, doi:10.1158/1078-0432.CCR-12-1153 (2012).
718 11 French, C. A. NUT Carcinoma: Clinicopathologic features, pathogenesis, and treatment.
719 *Pathol Int* **68**, 583-595, doi:10.1111/pin.12727 (2018).
720 12 Lauer, U. M., Hinterleitner, M., Horger, M., Ohnesorge, P. V. & Zender, L. NUT
721 Carcinoma-An Underdiagnosed Malignancy. *Front Oncol* **12**, 914031,
722 doi:10.3389/fonc.2022.914031 (2022).
723 13 Shiota, H. *et al.* Nut Directs p300-Dependent, Genome-Wide H4 Hyperacetylation in
724 Male Germ Cells. *Cell Rep* **24**, 3477-3487 e3476, doi:10.1016/j.celrep.2018.08.069
725 (2018).
726 14 Kees, U. R., Mulcahy, M. T. & Willoughby, M. L. Intrathoracic carcinoma in an 11-year-
727 old girl showing a translocation t(15;19). *Am J Pediatr Hematol Oncol* **13**, 459-464,
728 doi:10.1097/00043426-199124000-00011 (1991).

729 15 French, C. A. *et al.* Midline carcinoma of children and young adults with NUT
730 rearrangement. *J Clin Oncol* **22**, 4135-4139, doi:10.1200/JCO.2004.02.107 (2004).
731 16 French, C. A. *et al.* BRD4-NUT fusion oncogene: a novel mechanism in aggressive
732 carcinoma. *Cancer Res* **63**, 304-307 (2003).
733 17 French, C. A. *et al.* NSD3-NUT fusion oncoprotein in NUT midline carcinoma: implications
734 for a novel oncogenic mechanism. *Cancer Discov* **4**, 928-941, doi:10.1158/2159-
735 8290.CD-14-0014 (2014).
736 18 He, M. *et al.* Tumor Mutation Burden and Checkpoint Immunotherapy Markers in NUT
737 Midline Carcinoma. *Appl Immunohistochem Mol Morphol* **28**, 495-500,
738 doi:10.1097/PAI.0000000000000781 (2020).
739 19 Alekseyenko, A. A. *et al.* The oncogenic BRD4-NUT chromatin regulator drives aberrant
740 transcription within large topological domains. *Genes Dev* **29**, 1507-1523,
741 doi:10.1101/gad.267583.115 (2015).
742 20 Eagen, K. P. & French, C. A. Supercharging BRD4 with NUT in carcinoma. *Oncogene* **40**,
743 1396-1408, doi:10.1038/s41388-020-01625-0 (2021).
744 21 French, C. A. *et al.* BRD-NUT oncoproteins: a family of closely related nuclear proteins
745 that block epithelial differentiation and maintain the growth of carcinoma cells.
746 *Oncogene* **27**, 2237-2242, doi:10.1038/sj.onc.1210852 (2008).
747 22 Grayson, A. R. *et al.* MYC, a downstream target of BRD-NUT, is necessary and sufficient
748 for the blockade of differentiation in NUT midline carcinoma. *Oncogene* **33**, 1736-1742,
749 doi:10.1038/onc.2013.126 (2014).
750 23 Wang, R. *et al.* Activation of SOX2 expression by BRD4-NUT oncogenic fusion drives
751 neoplastic transformation in NUT midline carcinoma. *Cancer Res* **74**, 3332-3343,
752 doi:10.1158/0008-5472.CAN-13-2658 (2014).
753 24 Haruki, N. *et al.* Cloned fusion product from a rare t(15;19)(q13.2;p13.1) inhibit S phase
754 in vitro. *J Med Genet* **42**, 558-564, doi:10.1136/jmg.2004.029686 (2005).
755 25 Gu, B. & Hakun, M. C. Challenges and Opportunities in NUT Carcinoma Research. *Genes*
756 **12**, 235 (2021).
757 26 Chen, X., Pappo, A. & Dyer, M. A. Pediatric solid tumor genomics and developmental
758 pliancy. *Oncogene* **34**, 5207-5215, doi:10.1038/onc.2014.474 (2015).
759 27 Puisieux, A., Pommier, R. M., Morel, A. P. & Lavial, F. Cellular Pliancy and the Multistep
760 Process of Tumorigenesis. *Cancer Cell* **33**, 164-172, doi:10.1016/j.ccell.2018.01.007
761 (2018).
762 28 Minas, T. Z. *et al.* Combined experience of six independent laboratories attempting to
763 create an Ewing sarcoma mouse model. *Oncotarget* **8**, 34141-34163,
764 doi:10.18632/oncotarget.9388 (2017).
765 29 Amin, A. D., Rajan, S. S., Groysman, M. J., Pongtornpipat, P. & Schatz, J. H. Oncogene
766 Overdose: Too Much of a Bad Thing for Oncogene-Addicted Cancer Cells. *Biomark*
767 *Cancer* **7**, 25-32, doi:10.4137/BIC.S29326 (2015).
768 30 Forster, A. *et al.* Engineering de novo reciprocal chromosomal translocations associated
769 with Mll to replicate primary events of human cancer. *Cancer Cell* **3**, 449-458,
770 doi:10.1016/s1535-6108(03)00106-5 (2003).

31 Van Deursen, J., Fornerod, M., Van Rees, B. & Grosveld, G. Cre-mediated site-specific translocation between nonhomologous mouse chromosomes. *Proc Natl Acad Sci U S A* **92**, 7376-7380, doi:10.1073/pnas.92.16.7376 (1995).

32 Drynan, L. F. *et al.* Mll fusions generated by Cre-loxP-mediated de novo translocations can induce lineage reassignment in tumorigenesis. *EMBO J* **24**, 3136-3146, doi:10.1038/sj.emboj.7600760 (2005).

33 Buchholz, F., Refaeli, Y., Trumpp, A. & Bishop, J. M. Inducible chromosomal translocation of AML1 and ETO genes through Cre/loxP-mediated recombination in the mouse. *EMBO Rep* **1**, 133-139, doi:10.1093/embo-reports/kvd027 (2000).

34 Gu, B., Posfai, E. & Rossant, J. Efficient generation of targeted large insertions by microinjection into two-cell-stage mouse embryos. *Nat Biotechnol* **36**, 632-637, doi:10.1038/nbt.4166 (2018).

35 Patel, M. R. *et al.* Longitudinal, noninvasive imaging of T-cell effector function and proliferation in living subjects. *Cancer Res* **70**, 10141-10149, doi:10.1158/0008-5472.CAN-10-1843 (2010).

36 Arnold, K. *et al.* Sox2(+) adult stem and progenitor cells are important for tissue regeneration and survival of mice. *Cell Stem Cell* **9**, 317-329, doi:10.1016/j.stem.2011.09.001 (2011).

37 Senoo, M., Pinto, F., Crum, C. P. & McKeon, F. p63 Is essential for the proliferative potential of stem cells in stratified epithelia. *Cell* **129**, 523-536, doi:10.1016/j.cell.2007.02.045 (2007).

38 Murphy, M. J., Wilson, A. & Trumpp, A. More than just proliferation: Myc function in stem cells. *Trends Cell Biol* **15**, 128-137, doi:10.1016/j.tcb.2005.01.008 (2005).

39 Dassule, H. R., Lewis, P., Bei, M., Maas, R. & McMahon, A. P. Sonic hedgehog regulates growth and morphogenesis of the tooth. *Development* **127**, 4775-4785, doi:10.1242/dev.127.22.4775 (2000).

40 Haack, H. *et al.* Diagnosis of NUT midline carcinoma using a NUT-specific monoclonal antibody. *Am J Surg Pathol* **33**, 984-991, doi:10.1097/PAS.0b013e318198d666 (2009).

41 Reynoird, N. *et al.* Oncogenesis by sequestration of CBP/p300 in transcriptionally inactive hyperacetylated chromatin domains. *EMBO J* **29**, 2943-2952, doi:10.1038/emboj.2010.176 (2010).

42 Tripathi, S. *et al.* Defining the condensate landscape of fusion oncoproteins. *Nature Communications* **14**, 6008, doi:10.1038/s41467-023-41655-2 (2023).

43 Sun, H., McGuire, M. F., Zhang, S. & Brown, R. E. NUT Midline Carcinoma: Morphoproteomic Characterization with Genomic and Therapeutic Correlates. *Ann Clin Lab Sci* **45**, 692-701 (2015).

44 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457, doi:10.1038/nmeth.3337 (2015).

45 Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37**, 773-782, doi:10.1038/s41587-019-0114-2 (2019).

46 Tabula Muris, C. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372, doi:10.1038/s41586-018-0590-4 (2018).

814 47 Jones, K. B. *et al.* Quantitative Clonal Analysis and Single-Cell Transcriptomics Reveal
815 Division Kinetics, Hierarchy, and Fate of Oral Epithelial Progenitor Cells. *Cell Stem Cell*
816 **24**, 183-192 e188, doi:10.1016/j.stem.2018.10.015 (2019).
817 48 Yang, F., Shen, D. & Shi, J. Primary renal NUT carcinoma identified by next-generation
818 sequencing: a case report and literature review. *Int J Clin Exp Pathol* **14**, 662-669 (2021).
819 49 Hingorani, S. R. *et al.* Preinvasive and invasive ductal pancreatic cancer and its early
820 detection in the mouse. *Cancer Cell* **4**, 437-450, doi:10.1016/s1535-6108(03)00309-x
821 (2003).
822 50 Logan, M. *et al.* Expression of Cre Recombinase in the developing mouse limb bud
823 driven by a Prxl enhancer. *Genesis* **33**, 77-80, doi:10.1002/gene.10092 (2002).
824 51 Xiao, W., Mohseny, A. B., Hogendoorn, P. C. & Cleton-Jansen, A. M. Mesenchymal stem
825 cell transformation and sarcoma genesis. *Clin Sarcoma Res* **3**, 10, doi:10.1186/2045-
826 3329-3-10 (2013).
827 52 Belteki, G. *et al.* Conditional and inducible transgene expression in mice through the
828 combinatorial use of Cre-mediated recombination and tetracycline induction. *Nucleic*
829 *Acids Res* **33**, e51, doi:10.1093/nar/gni051 (2005).
830 53 den Bakker, M. A. *et al.* NUT midline carcinoma of the parotid gland with mesenchymal
831 differentiation. *Am J Surg Pathol* **33**, 1253-1258, doi:10.1097/PAS.0b013e3181abe120
832 (2009).
833 54 Bell, D. *et al.* High-grade sinonasal carcinomas and surveillance of differential expression
834 in immune related transcriptome. *Ann Diagn Pathol* **49**, 151622,
835 doi:10.1016/j.anndiagpath.2020.151622 (2020).
836 55 Thawani, R. *et al.* The contemporary management of cancers of the sinonasal tract in
837 adults. *CA Cancer J Clin* **73**, 72-112, doi:10.3322/caac.21752 (2023).
838 56 Whatcott, C. J. *et al.* Desmoplasia in Primary Tumors and Metastatic Lesions of
839 Pancreatic Cancer. *Clin Cancer Res* **21**, 3561-3568, doi:10.1158/1078-0432.CCR-14-1051
840 (2015).
841 57 Ren, B. *et al.* Tumor microenvironment participates in metastasis of pancreatic cancer.
842 *Mol Cancer* **17**, 108, doi:10.1186/s12943-018-0858-1 (2018).
843 58 Mascharak, S. *et al.* Desmoplastic stromal signatures predict patient outcomes in
844 pancreatic ductal adenocarcinoma. *Cell Reports Medicine*,
845 doi:10.1016/j.xcrm.2023.101248 (2023).
846 59 Stelow, E. B. A review of NUT midline carcinoma. *Head Neck Pathol* **5**, 31-35,
847 doi:10.1007/s12105-010-0235-x (2011).
848 60 Agaimy, A. *et al.* NUT Carcinoma of the Salivary Glands: Clinicopathologic and Molecular
849 Analysis of 3 Cases and a Survey of NUT Expression in Salivary Gland Carcinomas. *Am J*
850 *Surg Pathol* **42**, 877-884, doi:10.1097/PAS.0000000000001046 (2018).
851 61 Gokmen-Polar, Y., Cano, O. D., Kesler, K. A., Loehrer, P. J. & Badve, S. NUT midline
852 carcinomas in the thymic region. *Mod Pathol* **27**, 1649-1656,
853 doi:10.1038/modpathol.2014.63 (2014).
854 62 Bakir, B., Chiarella, A. M., Pitarresi, J. R. & Rustgi, A. K. EMT, MET, Plasticity, and Tumor
855 Metastasis. *Trends Cell Biol* **30**, 764-776, doi:10.1016/j.tcb.2020.07.003 (2020).
856 63 Nieto, M. A. The snail superfamily of zinc-finger transcription factors. *Nat Rev Mol Cell*
857 *Biol* **3**, 155-166, doi:10.1038/nrm757 (2002).

858 64 Liu, Y., El-Naggar, S., Darling, D. S., Higashi, Y. & Dean, D. C. Zeb1 links epithelial-
859 mesenchymal transition and cellular senescence. *Development* **135**, 579-588,
860 doi:10.1242/dev.007047 (2008).
861 65 Ridge, K. M., Eriksson, J. E., Pekny, M. & Goldman, R. D. Roles of vimentin in health and
862 disease. *Genes Dev* **36**, 391-407, doi:10.1101/gad.349358.122 (2022).
863 66 Berr, A. L. *et al.* Vimentin is required for tumor progression and metastasis in a mouse
864 model of non-small cell lung cancer. *Oncogene* **42**, 2074-2087, doi:10.1038/s41388-023-
865 02703-9 (2023).
866 67 Brabletz, S., Schuhwerk, H., Brabletz, T. & Stemmler, M. P. Dynamic EMT: a multi-tool
867 for tumor progression. *EMBO J* **40**, e108647, doi:10.15252/embj.2021108647 (2021).
868 68 Xie, Y., Shi, L., He, X. & Luo, Y. Gastrointestinal cancers in China, the USA, and Europe.
869 *Gastroenterol Rep (Oxf)* **9**, 91-104, doi:10.1093/gastro/goab010 (2021).
870 69 Giacobbe, A. & Abate-Shen, C. Modeling metastasis in mice: a closer look. *Trends Cancer*
871 **7**, 916-929, doi:10.1016/j.trecan.2021.06.010 (2021).
872 70 Chen, Y. *et al.* Dual reporter genetic mouse models of pancreatic cancer identify an
873 epithelial-to-mesenchymal transition-independent metastasis program. *EMBO Mol Med*
874 **10**, doi:10.15252/emmm.201809085 (2018).
875 71 Durall, R. T. *et al.* The BRD4-NUT fusion alone drives malignant transformation of NUT
876 carcinoma. *Cancer Res*, doi:10.1158/0008-5472.CAN-23-2545 (2023).
877 72 Sinjab, A., Han, G., Wang, L. & Kadara, H. Field Carcinogenesis in Cancer Evolution: What
878 the Cell Is Going On? *Cancer Res* **80**, 4888-4891, doi:10.1158/0008-5472.CAN-20-1956
879 (2020).
880 73 Gu, B., Posfai, E., Gertsenstein, M. & Rossant, J. Efficient Generation of Large-Fragment
881 Knock-In Mouse Models Using 2-Cell (2C)-Homologous Recombination (HR)-CRISPR. *Curr*
882 *Protoc Mouse Biol* **10**, e67, doi:10.1002/cpmo.67 (2020).
883 74 Gu, B., Gertsenstein, M. & Posfai, E. Generation of Large Fragment Knock-In Mouse
884 Models by Microinjecting into 2-Cell Stage Embryos. *Methods Mol Biol* **2066**, 89-100,
885 doi:10.1007/978-1-4939-9837-1_7 (2020).
886 75 Stelow, E. B. *et al.* NUT rearrangement in undifferentiated carcinomas of the upper
887 aerodigestive tract. *Am J Surg Pathol* **32**, 828-834, doi:10.1097/PAS.0b013e31815a3900
888 (2008).
889 76 Bioinformatics, B. *FastQC: A quality control tool for high throughput sequence data*,
890 <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).
891 77 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
892 sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170
893 (2014).
894 78 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
895 quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).
896 79 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-
897 level estimates improve gene-level inferences. *F1000Res* **4**, 1521,
898 doi:10.12688/f1000research.7563.2 (2015).
899 80 Gentleman, R. C. *et al.* Bioconductor: open software development for computational
900 biology and bioinformatics. *Genome Biol* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).

81 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
82 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B* **57**, 289-300, doi:DOI 10.1111/j.2517-6161.1995.tb02031.x (1995).
83 Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084-2092, doi:10.1093/bioinformatics/bty895 (2019).
84 Yu, G. C., Wang, L. G., Han, Y. Y. & He, Q. Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omics* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).
85 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
86 Gene Ontology, C. *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, doi:10.1093/genetics/iyad031 (2023).
87 Kanehisa, M. A database for post-genome analysis. *Trends Genet* **13**, 375-376, doi:10.1016/s0168-9525(97)01223-7 (1997).
88 Vastrik, I. *et al.* Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* **8**, R39, doi:10.1186/gb-2007-8-3-r39 (2007).
89 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).
90 Boyle, E. I. *et al.* GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710-3715, doi:10.1093/bioinformatics/bth456 (2004).
91 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
92 Yu, G. C. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976-978, doi:10.1093/bioinformatics/btq064 (2010).
93 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
94 Institute, B. *"Picard Toolkit."*, <GitHub Repository. https://broadinstitute.github.io/picard/> (2019).
95 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
96 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92, doi:10.4161/fly.19695 (2012).
97 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

98 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339, doi:10.1093/bioinformatics/bts378 (2012).
99 Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773-782, doi:10.1038/s41587-019-0114-2 (2019).

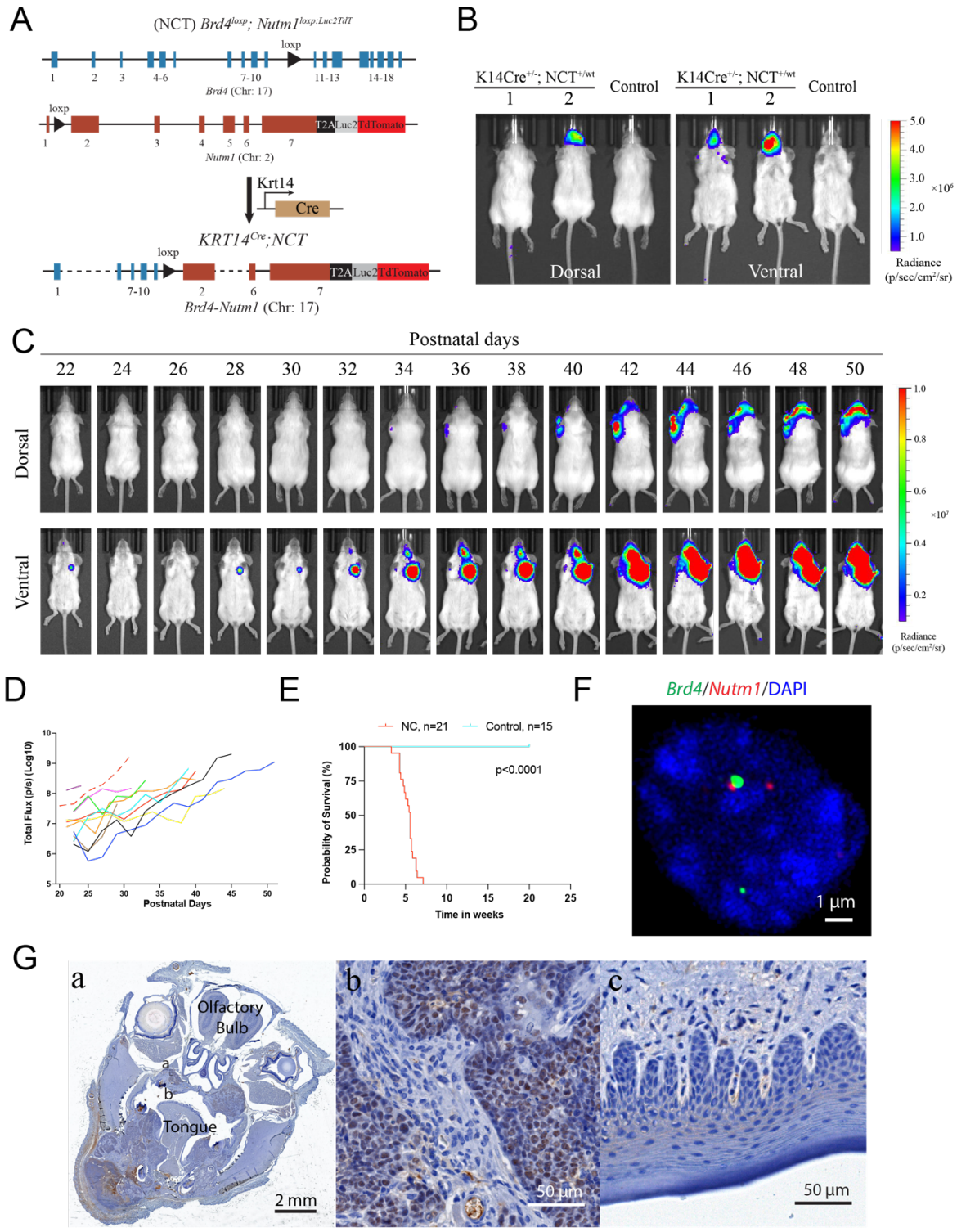
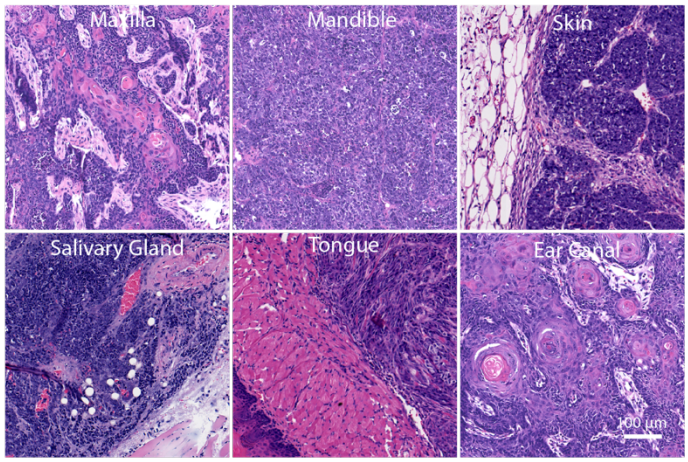


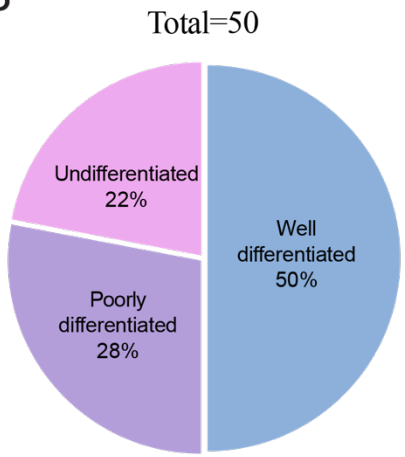
Figure 1: Generation of the NC GEMM: A. Genetic design of the Nut carcinoma translocator (NCT) and the generation of the Krt14Cre induced mNC. B. Representative BLI images of mice with head NC. C. BLI images of the growth kinetics of mNC. D.

Growth curves of mNC in 8 mice quantified by BLI signals; each curve represents one mouse. E. Kaplan Meier survival curve of Krt14Cre induced mNC mice. P value calculated by log-rank test. F. DC-FISH of the *Brd4* (Green) and *Nutm1* (Red) on mNC. G. A representative NUTM1 IHC image of NC bearing whole mouse head (a) and zoom-ins on the tumor (b) and normal oral mucosa (c) regions.

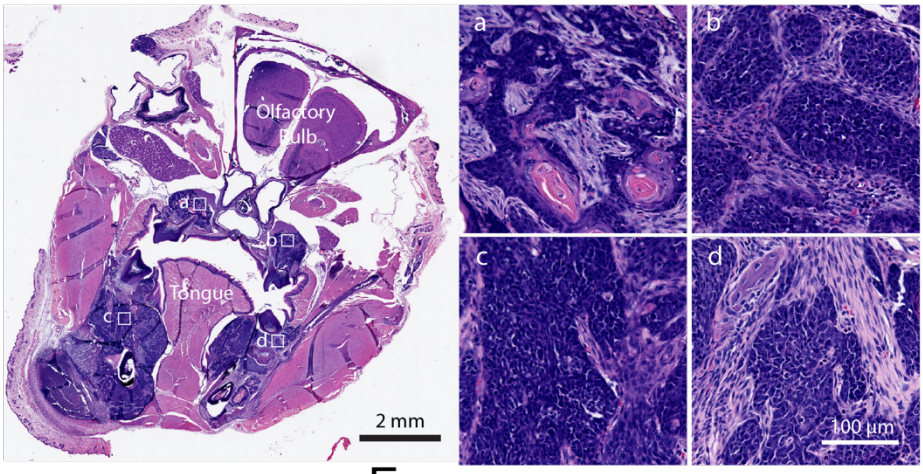
A


Maxilla
Mandible
Skin
Salivary Gland
Tongue
Ear Canal
100 µm

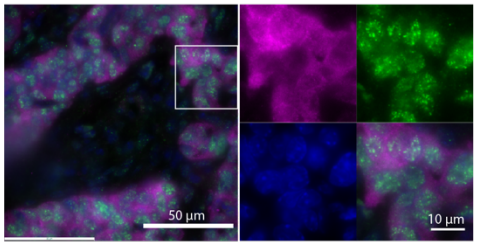
B


Total=50
Undifferentiated 22%
Well differentiated 50%
Poorly differentiated 28%

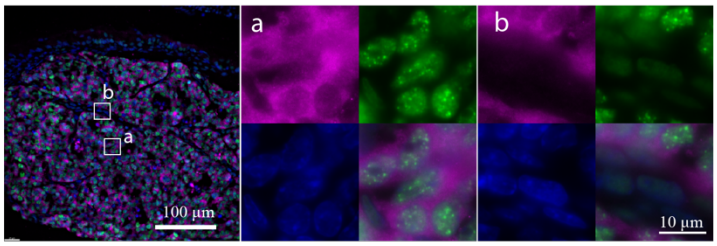
C


Olfactory Bulb
Tongue
a
b
c
d
2 mm
100 µm

D


TdTomato/NUTM1/Hoechst
50 µm
10 µm

E


TdTomato/H3K27ac/Hoechst
a
b
100 µm
10 µm

F

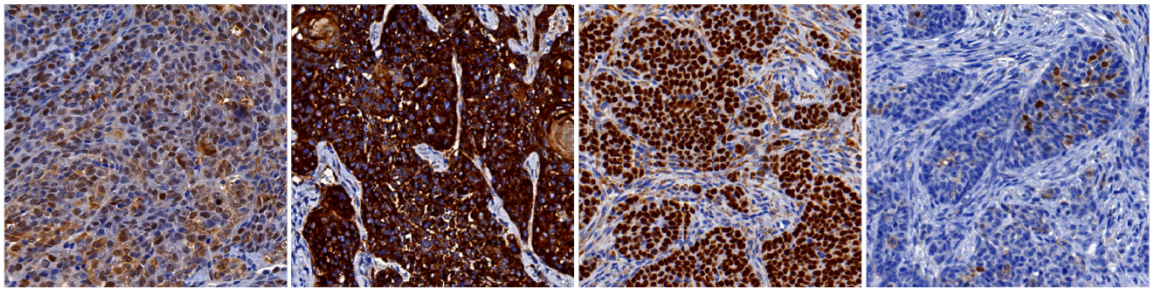

cMYC
KRT14
P63
SOX2

Figure 2: Histological and molecular characteristics of mNCs: A. Representative H&E staining images of Krt14Cre mNC tumors at different anatomical locations. B. The distribution of pathological features of mNCs. C. Tumors formed in the same mouse presenting variable histological features. D. Representative image of IF staining of NUTM1 in mNC showing nuclear foci. TdTomato staining demarcates tumor cells E. Representative images of IF analysis of H3K27ace in mNCs, depicting nuclear foci in tumor cells (b) versus a diffused nuclear stain in peri-tumor stromal cells (c). F. Representative IHC images of cMYC, KRT14, P63, and SOX2 in oral mNCs.

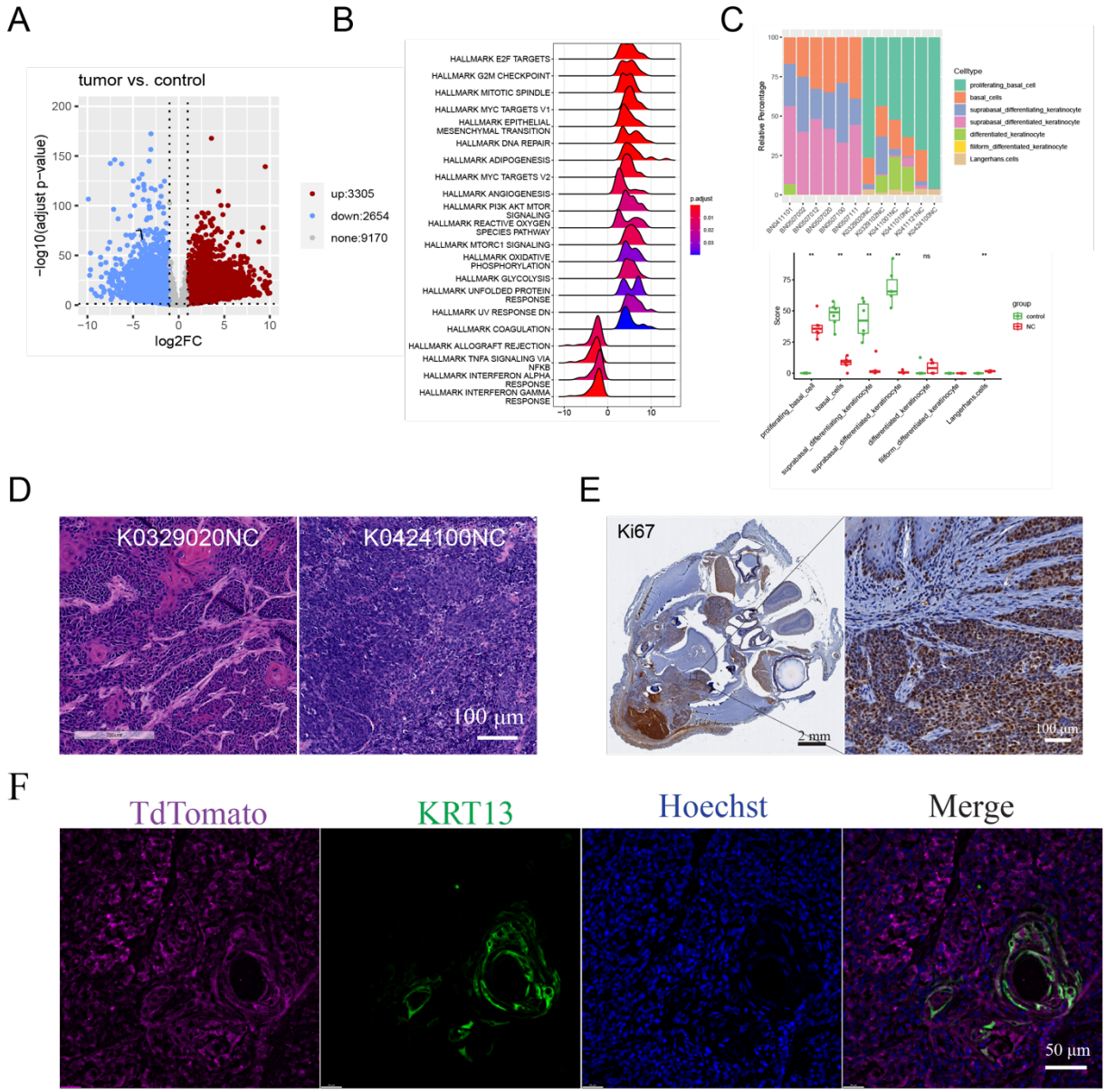


Figure 3: mNCs largely consist of proliferative progenitor cells. A. Volcano plot of differentially expressed genes in mNCs compared to normal oral mucosa. B. GSEA enrichment of HALLMARK pathways in mNCs. C. CIBERSORT deconvolution of the transcriptome of mNCs according to the *Tabula Muris* dataset. D. Representative histology images of mNCs that enriches the proliferative basal cell profile at different

level. E. A representative Ki67 IHC image of mNCs. F. A representative IF image of NC cells (tdTomato) and the ones expressing keratinization marker KRT13.

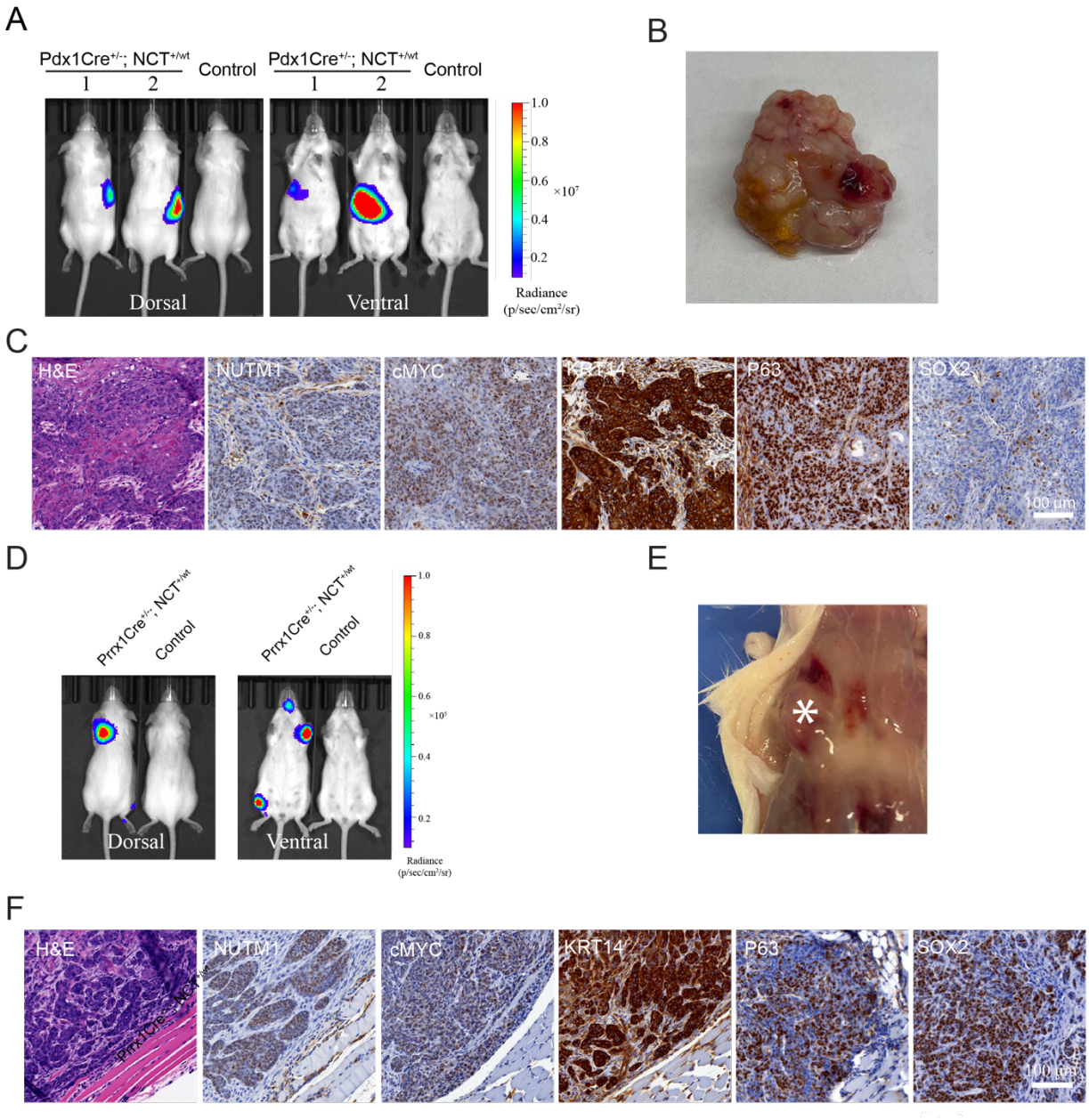


Figure 4: *Brd4::Nutm1* initiates pancreatic and soft tissue tumors. A. Representative BLI images of mice with pancreatic mNCs. B. A representative macroscopic image of pancreatic mNCs. C. Representative H&E, and IHC images of NUTM1, cMYC, KRT14, P63, and SOX2 in pancreatic mNCs. D. Representative BLI images of mice with soft tissue mNCs. E. A representative macroscopic image of soft tissue mNCs. F.

Representative H&E, and IHC images of NUTM1, cMYC, KRT14, P63, and SOX2 in pancreatic mNCs.

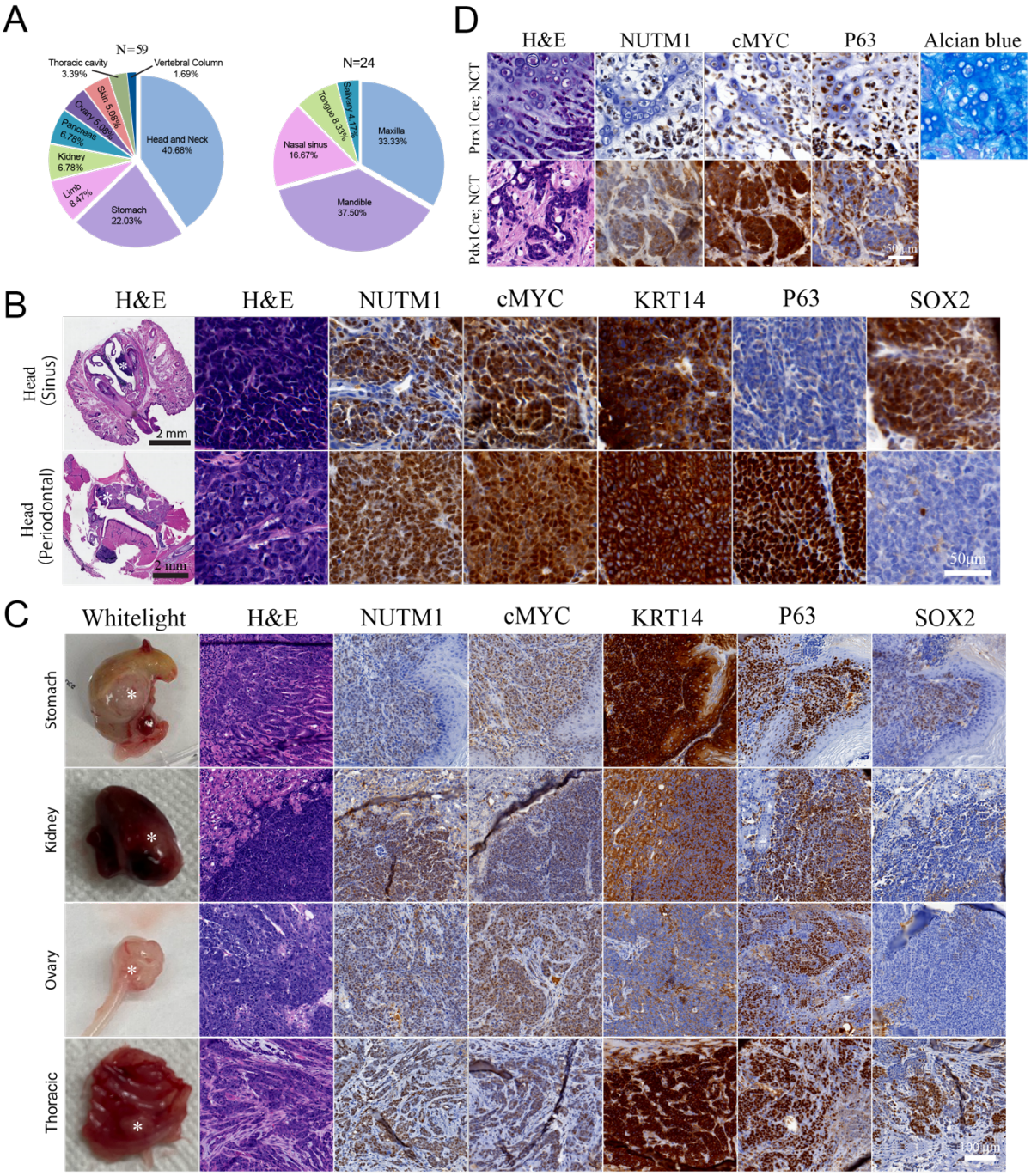


Figure 5: *Brd4::Nutm1* initiates NC across tissues. A. Tissue distribution of the NLS-Cre driven mNCs; left: distribution throughout body. Right. Distribution of the head and neck tumors. B. Representative whole-head H&E, high magnification H&E, and IHC images of nasal and oral NLS-Cre driven mNCs. C. Representative macroscopic, H&E, and IHC

images of NLS-Cre driven mNCs outside the head and neck region. D. Representative H&E, IHC and Alcian Blue cartilage staining images of chondrogenic and ductal metaplasia in mNCs.

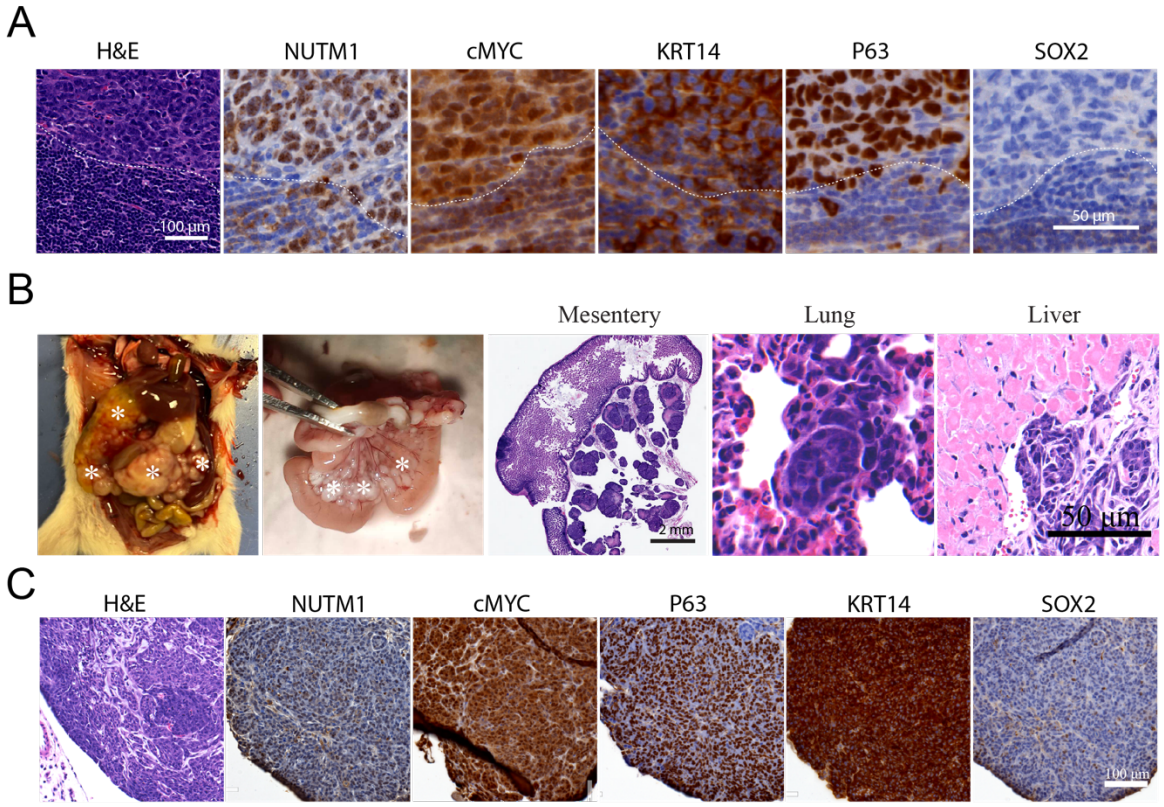


Figure 6: mNCs are highly metastatic. A. Representative H&E and IHC images of a lymph node metastasis from oral mNC. B. Representative macroscopic and H&E images of peritoneal/mesentry, lung and liver metastasis of pancreatic mNCs. C. Representative H&E and IHC images of peritoneal metastasis of pancreatic mNCs.

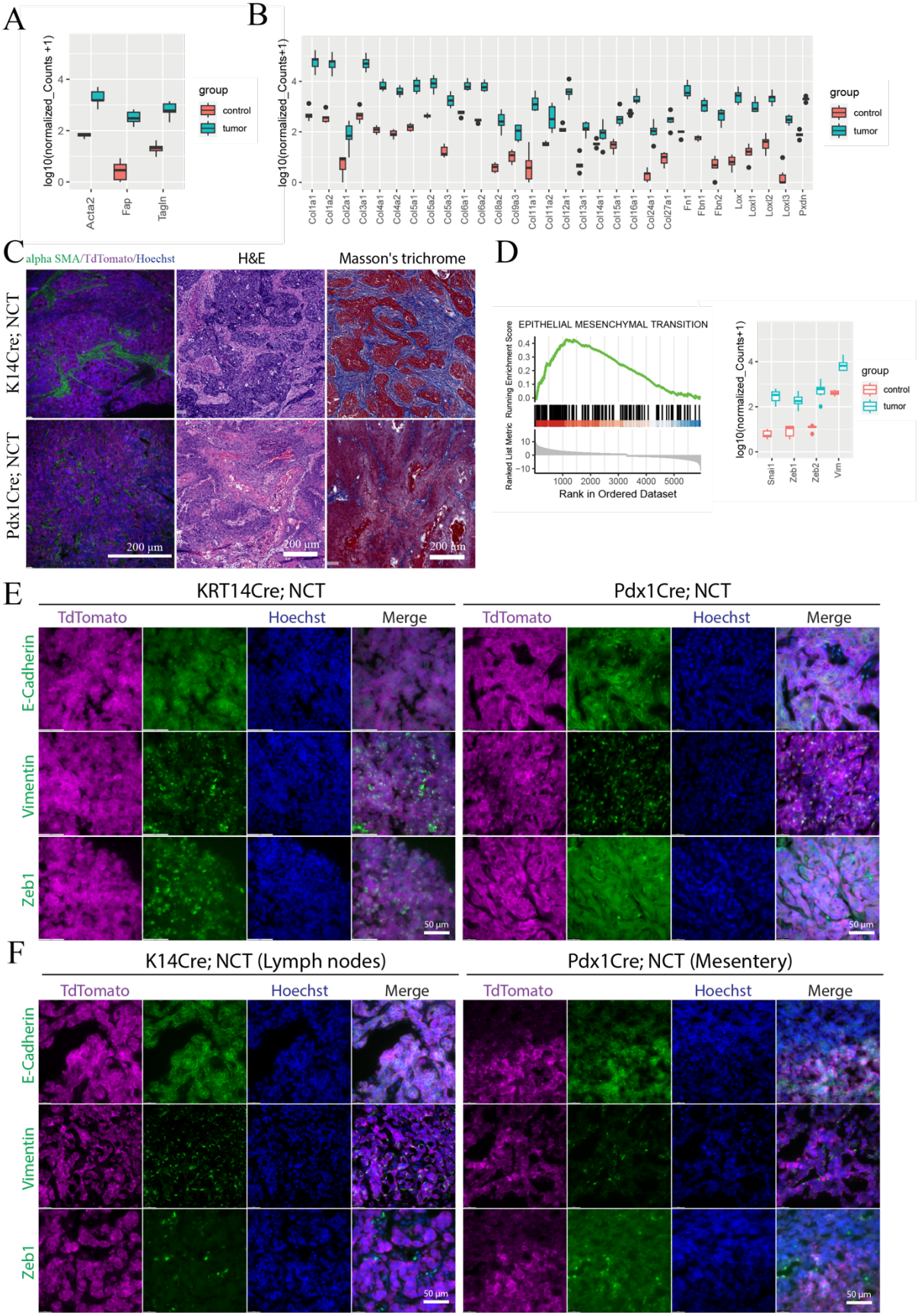

A
log10(normalized_Counts +1)
Acta2
Fap
Tagln
group
control
tumor
B
log10(normalized_Counts+1)
Col1a1
Col1a2
Col2a1
Col3a1
Col4a1
Col4a2
Col5a1
Col5a2
Col5a3
Col6a1
Col6a2
Col8a2
Col9a3
Col11a1
Col11a2
Col12a1
Col13a1
Col14a1
Col15a1
Col16a1
Col24a1
Col27a1
Fn1
Fbn1
Fbn2
Lox
Loxl1
Loxl2
Loxl3
Pxdn
group
control
tumor
C
alpha SMA/TdTomato/Hoechst
H&E
Masson's trichrome
K14Cre; NCT
Pdx1Cre; NCT
200 µm
D
EPITHELIAL MESENCHYMAL TRANSITION
Running Enrichment Score
Ranked List Metric
Rank in Ordered Dataset
log10(normalized_Counts+1)
Snai1
Zeb1
Zeb2
Vim
group
control
tumor
E
KRT14Cre; NCT
Pdx1Cre; NCT
TdTomato
Hoechst
Merge
E-Cadherin
Vimentin
Zeb1
50 µm
F
K14Cre; NCT (Lymph nodes)
Pdx1Cre; NCT (Mesentery)
TdTomato
Hoechst
Merge
E-Cadherin
Vimentin
Zeb1
50 µm

Figure 7: mNCs present desmoplastic and EMT features that could contribute to its metastatic potential. A. Upregulation of markers of activated fibroblasts in mNCs. B. Upregulation of ECM and their modifier genes in mNCs. C. Representative images of desmoplastic features of oral (upper) and pancreatic (bottom) mNCs. D. GSEA enrichment of EMT genes (left) and upregulation of key EMT markers in oral mNCs. E. Expression of the epithelial marker E-Cadherin and the Mesenchymal marker Vimentin and Zeb1 in advanced oral and pancreatic mNCs, showing regions of EMT. TdTomato staining demarcates NC cells. F. Expression of the epithelial marker E-Cadherin and the Mesenchymal marker Vimentin and Zeb1 in metastatic mNCs, presenting heterogenous EMT status. TdTomato staining demarcates NC cells.

Supplementary Figure legends:

Supplementary Figure 1. The validation of the NC GEMM. A. Representative BLI images of NCT mice. B. A representative IF image of NUTM1 and TdTomato in the testis of NCT mice. C. A representative IF image of TdTomato in a whole head section of Krt14Cre; R26-LSL-TdTomato mice, depicting the tissue distribution of Cre activities. D. BLI images of 19 representative Krt14Cre-dirven NC mice. E. The tissue distribution of Krt14Cre-dirven mNCs. F. A representative sange sequencing track of the *Brd4::Nutm1* fusion junction. G. A representative karyotype of Krt14Cre-dirven mNC cells. Arrows indicates the t(2;17) translocation. H. Circos chart of chromosome translocation detected by whole genome sequencing of two mNCs. The thick blue line indicates t(2;17) forming *Brd4::Nutm1* is the only recurrent chromosome translocation. I. Two additional whole head NUTM1 IHC in mNC bearing mice.

Supplementary Figure 2. Molecular features if mNCs. A. A representative IHC image showing NUTM1 nuclear foci in mNCs. B. Representative IF images showing H3K27ace nuclear foci in mNCs. C. Representative whole head section IHC images of P63, KRT14, cMYC and SOX2 in two mNC bearing mice.

Supplementary Figure 3. RNA sequencing analysis of mNCs. A. RNA integrity number of samples demonstrated appropriate quality. B. Clustering analysis of RNA sequencing data. C. Principal Component Analysis (PCA) of RNA sequencing data D. CIBERSORT deconvolution of the bulk RNA sequencing data using a second set of single-cell RNA sequencing signatures of oral mucosa epithelium.

Supplementary Figure 4. BRD4::NUTM1 induced pancreatic and soft tissue mNCs. A. BLI images of 18 Pdx1Cre;NCT mice. B. Representative H&E and IHC images of pancreatic NC from Pdx1Cre;NCT mice. C. BLI images of 9 Prrx1Cre;NCT mice. D. B. Representative H&E and IHC images of soft tissue NC from Prrx1Cre;NCT mice.

Supplementary Figure 5. BRD4::NUTM1 induced NC from a broad range of tissues. A. BLI images of 25 NLS-Cre; NCT mice. B. Histopathological diagnosis of 62 mNCs derived from NLS-Cre; NCT mice. C. Additional representative H&E images of mNCs derived from NLS-Cre; NCT mice. D. Additional representative H&E images of metaplastic mNCs derived from NLS-Cre; NCT mice.

Supplementary Figure 6. Metastatic lesions from pancreatic mNC. A. Representative macroscopic images of metastatic mNCs. B. Representative IF of EMT markers of abdominal metastasis from pancreatic mNCs. C. Representative IF of EMT markers of diaphram metastasis from pancreatic mNCs. D. Representative IF of EMT markers of thymic metastasis from pancreatic mNCs.