**Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor EVOLUTION**

Michelle Chan-Seng-Yue1,2\*, Jaeseung C. Kim1,3,\*, Gavin W. Wilson1, Karen Ng1, Eugenia Flores Figueroa1, Grainne M. O’Kane2,4, Ashton A. Connor5, Robert E. Denroche2, Robert C. Grant4, Jessica McLeod1, Julie M. Wilson2, Gun Ho Jang2, Amy Zhang2, Sheng-Ben Liang6, Ayelet Borgida7, Dianne Chadwick6, Sangeetha Kalimuthu8, Ilinca Lungu9, John M. S. Bartlett9, Paul M. Krzyzanowski3, Vandana Sandhu1, Hervé Tiriac10,11, Fieke E. M. Froeling10,12, Joanna M. Karasinska13, James T. Topham13, Daniel J. Renouf13,14, David F. Schaeffer13,15, Steven J. M. Jones16,17, Marco A. Marra16,17, Janessa Laskin14, Runjan Chetty8, Lincoln D. Stein18,19, George Zogopoulos20,21, Benjamin Haibe-Kains1,2,22,23, Peter J. Campbell24,25, David A. Tuveson10, Jennifer J. Knox2,4, Sandra E. Fischer8,26, Steven Gallinger2,5,7,27**‡**, Faiyaz Notta1,2,22**‡**

1Princess Margaret Cancer Centre, Toronto, Ontario, Canada,

2PanCuRx Translational Research Initiative, Ontario Institute for Cancer Research, Toronto, Ontario, Canada,

3Genomics Technology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada,

4Wallace McCain Centre for Pancreatic Cancer, Department of Medical Oncology, Princess Margaret Cancer Centre, University Health Network (UHN), University of Toronto, Toronto, Ontario, Canada,

5Hepatobiliary/Pancreatic Surgical Oncology Program, University Health Network, Toronto, Ontario, Canada,

6UHN Program in BioSpecimen Sciences, University Health Network, Toronto, ON, Canada,

7Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada,

8Department of Pathology, Uni aryversity Health Network and University of Toronto, Toronto, Ontario, Canada,

 9Diagnostic Development, Ontario Institute for Cancer Research, Toronto, Ontario, Canada,

10Cold Spring Harbor Laboratory, Cold Spring Harbor, New York; Lustgarten Foundation Pancreatic Cancer Research Laboratory, Cold Spring Harbor, NY, USA,

11current address: University of California San Diego, Department of Surgery, NCI-designated Comprehensive Cancer Center, La Jolla, CA, USA,

12Department of Surgery and Cancer, Imperial College London, London, United Kingdom

13Pancreas Centre BC, Vancouver, British Columbia, Canada,

14Department of Medical Oncology, BC Cancer, Vancouver, British Columbia, Canada,

15Department of Pathology and Laboratory Medicine, Vancouver Generate Hospital, Vancouver, British Columbia, Canada,

16Canada’s Michael Smith Genome Sciences Centre, Vancouver, British Columbia, Canada,

17Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia,

18Bioinformatics, Ontario Institute for Cancer Research, Toronto, Ontario, Canada,

19Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

20Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada,

21The Goodman Cancer Research Centre of McGill University, Montreal, Quebec, Canada,

22Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada,

23Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

24Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK,

25Department of Haematology, University of Cambridge, Cambridge, UK,

26Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada,

27Department of Surgery, University of Toronto, Toronto, Ontario, Canada

\* These authors contributed equally to this work

**‡** To whom correspondence should be addressed:

Faiyaz Notta, PhD (faiyaz.notta@gmail.com);

Steven Gallinger, MD, MSc, FRCSC (steven.gallinger@uhn.ca)

**Pancreatic adenocarcinoma presents as a spectrum of a highly aggressive disease in patients. The basis of this disease heterogeneity has been difficult to resolve due to poor tumor cellularity and extensive genomic instability. To address this, a dataset of whole genomes and transcriptomes was generated from purified epithelium of primary and metastatic tumors. Transcriptome analysis demonstrated that molecular subtypes are a product of a gene-expression continuum driven by a mixture of intratumoral subpopulations, which was confirmed by single cell analysis. Integrated whole genome analysis uncovered that molecular subtypes are linked to specific copy number aberrations in genes such as mutant *KRAS* and *GATA6*. By mapping tumor genetic histories, tetraploidization emerged as a key mutational process behind these events. Taken together, these data support that the constellation of genomic aberrations in the tumor give rise to the molecular subtype, and that disease heterogeneity is due to ongoing genomic instability during progression.**

Pancreatic adenocarcinoma (herein pancreatic cancer) is a highly lethal cancer. Most patients present with Stage IV disease and will succumb to it within a year of diagnosis. Although perceived to be uniformly aggressive, there is extensive heterogeneity amongst patients with respect to treatment response. Excluding rare cases1,2, most patients respond for a limited duration ranging from months to over a year, whereas others stabilize briefly before progressing. Still, others do not respond at all and show tumor growth on first line chemotherapy. Among patients with surgically resectable tumors (Stage I or II), some progress to Stage IV within months, while others recur after more than a year, and a minority are cured3. Genomic studies, however continue to support that this disease is homogeneous with recurrent mutations in four genes: *KRAS*, *CDKN2A*, *TP53*, and *SMAD4*4-7. It remains unclear how the clinical heterogeneity arises if most tumors develop through the same mutational path.

Limited mutational heterogeneity among primary and metastatic tumors8,9 suggests that non-genetic mechanisms, such as cell-of-origin or stromal cell interactions, may be the predominant source of this variation10. However, most genomic studies addressing this question have been performed using tumors with low cellularity and have focused on single-nucleotide variants (SNV), which can be more readily detected than DNA copy number alterations. Pancreatic tumors exhibit a high frequency of chromothripsis11,12; however, whether mutational processes related to genomic instability influence clinical disease heterogeneity in pancreatic cancer is unknown.

Two experimental hurdles hinder the ability to address this issue. First, most genomic studies on pancreatic cancer are based on early stage disease (Stage I/II), which are a minority of the patient population. This is mostly due to difficulty in obtaining research material from metastatic patients. In December 2015, we launched the COMPASS trial (NCT02750657) which recruits only advanced patients for genomic analysis. Because our patient cohort included Stage IV disease, which accounts for more than half of the pancreatic cancer patient population, we were able to capture a wider spectrum of disease. Second, low tumor cellularity remains a major concern in molecular subtyping of this disease. Recent work by Raphael *et al.*13 confirmed that molecular classification of pancreatic cancer, with the exception of Moffitt *et al.*14, are influenced by a high degree of stromal cell infiltration. Therefore, tumor expression features may be obscured in the absence of cell purification. To improve tumor cellularity for genomic analyses, local infrastructure was built to purify tumor epithelium via laser capture microdissection (LCM) from a large number of cases. We performed whole-genome sequencing (WGS) of 330 LCM-purified tumors from 314 patients and whole-transcriptome sequencing (RNA-seq) of 248 tumors. This study represents the most comprehensive analysis of the molecular subtypes of pancreatic cancer to date. We used this data to investigate whether there is a genetic basis to clinical disease heterogeneity in this disease.

**Results**

**Disease subtypes and their association with clinical stage**

Notably, current subtyping schemes14-16 can identify prognostic subgroups in resectable tumors but not in advanced disease (Extended Data Fig. 1). This motivated us to perform a *de novo* reclassification of disease subtypes using our cohort (Supplementary Table 1). Non-negative matrix factorization (NMF) was used to extract tumor-specific expression signatures (Sigs. 1, 2, 6 and 10; Supplementary note), and these expression patterns were confirmed in the single-cell dataset (Supplementary note). Genes from these 4 signatures were then used for consensus clustering17. The cohort segregated into five subtypes (Fig. 1a, n=248 tumors) based on a gene-expression continuum of the tumor NMF signatures, and was independent of tumor cellularity (Fig. 1a, Supplementary note). We labelled them Basal-like-A, Basal-like-B, Hybrid, Classical-A and Classical-B. By aligning them to previous classification models (Fig. 1b, Supplementary note), we learned that our classification splits each of the previously defined ‘Basal-like’ and ‘Classical’ subtypes into two disease subtypes (Fig. 1b). The fifth subtype was inconsistently classified by all previous schemes, due to the presence of multiple expression signatures (Fig. 1a, Extended Data Fig. 2, Supplementary note). Accordingly, we termed this subset as ‘Hybrids’.

 The above subtypes were then evaluated for clinical disease stage. Classical-A/B tumors were more frequent in early stage (Stage I/II – 62%, n=98/159) compared to Stage IV (46%; n=34/74; p=0.03, Fisher’s exact test; Fig. 1c, d). Accordingly, Basal-like-A subtype, which is rare in resectable disease (5%, n=8/159) and absent in locally advanced tumors (n=14), accounts for nearly a quarter of Stage IV disease (24%, n=18/74; p=0.00003, Fisher’s exact test). Interestingly, previous ‘Basal-like’ tumors in resectable disease predominantly consisted of Basal-like-B (9%; n=14/159) and Hybrid tumors (24%; n=38/159; Fig. 1c). This suggests that the Basal-like phenotype between resectable and advanced disease is different. Distinguishing Basal-like-A, Basal-like-B, and Hybrids provided two important insights: 1) in resectable disease, Basal-like-B and Hybrid tumors identify two prognostic subgroups which were considered to be uniformly aggressive under previous classification schemes (Quasi-mesenchymal15, Basal-like14, Squamous16; Extended Data Fig. 3a,b); and 2) in advanced disease, Basal-like-A, not Basal-like-B, are highly chemoresistant, and trend towards worse survival (Extended Data Fig. 3c,d). The separation of Basal-like-A from Basal-like-B and Hybrid tumors allowed us to observe differences in chemotherapy response (Extended Data Fig. 3e,f). Large independent cohorts will be required to validate these findings in the future. These data show that tumor subtypes are not uniform at different clinical stages and also highlight the importance of including Stage IV tumors in molecular analysis.

**Basal-like and Classical programs co-exist intratumorally**

To investigate how the expression signatures from bulk RNA-seq distribute intratumorally, we performed single-cell RNA-seq (scRNA-seq) on 15 patient tumors (13 resectable, 2 metastatic) (Extended Data Fig. 4a). Tumor epithelium was enriched using negative cell selection (see Methods) and a total of 31,195 cells were studied. Leftover immune cells, fibroblasts, and endothelial cells were identified using established lineage marker genes (Extended Data Fig. 4b,c). Epithelial cells were identified using lineage antigens using a previous strategy18 (Extended Data Fig. 4c). Marker genes from the single-cell transcriptome atlas of the normal human pancreas19 were used to discriminate normal from malignant epithelial cells (Extended Data Fig. 4d,e). In support of our strategy, rare mutant *KRAS* reads and inferred copy number aberrations were only observed in cells defined as tumor (Extended Data Fig. 4f,g,h and Supplementary note).

 Tumor cells were clustered and scored for Basal-like (Sigs. 2 and 10) and Classical signatures (Sigs. 1 and 6). In the tumor shown in Fig. 2a and b, we observed 7 single cell clusters: clusters 0, 1, and 3 were enriched for Basal-like signatures, clusters 2, 4, and 6 were enriched for Classical signatures, and cluster 5 did not show high scores for any of the signatures (Fig. 2b). In 13 out of 15 tumors, both Basal-like and Classical clusters were found in the same tumor; however, these cell clusters commonly demarcated distinct tumor cell populations (Extended Data Fig. 4i). Pairwise comparisons of Basal-related and Classical-related expression signatures showed that they were negatively correlated at the single cell level (Fig. 2c; n=15 samples). From the bulk RNA-seq analysis, we found a relationship between Basal-like signatures and epithelial mesenchymal transition (EMT) program (Supplementary note and below). Within single cells, EMT program was positively correlated with Basal-like signatures and negatively correlated with Classical signatures (Fig. 2d). In summary, we observe that 1) the Basal-like and the Classical expression signatures can exist intratumorally, 2) these signatures commonly segregate to different tumor cell subsets, and 3) there is a correlation between EMT and the Basal-like expression program.

***SMAD4* and *GATA6* alterations enriched in Classical tumors**

Next, we analyzed the WGS data for genetic features of the tumor subtypes. Tumors with homologous recombination defects (HRD) and DNA mismatch repair deficiency (MMR) were excluded due to their unique mutational signatures (Extended Data Fig. 5a). By examining point mutations, no gene other than *TP53* or gene network was found to be enriched in Basal-like-A/B subtypes (Extended Data Fig. 5b-f, Supplementary note). Gene set enrichment analysis (GSEA) demonstrated that EMT and TGF-β signaling are enriched in Basal-like-A/B tumors (Extended Data Fig. 6a), thus we analyzed the status of the *SMAD4* gene – a key player in TGF- signaling. All major forms of genetic mutations (i.e. SNV, structural variant (SV), loss of heterozygosity (LOH), homozygous deletions, and amplifications) were considered. An intact allele of *SMAD4* was observed in 78% (n=18/23) of Basal-like-A tumors. By contrast, Classical-A tumors showed the lowest frequency of intact *SMAD4* (48%; n=45/94; Basal-like-A vs Classical-A, p=0.01, Fisher’s exact test; Extended Data Fig. 6b, Supplementary Table 2). Also, in accordance with GSEA findings, Basal-like-A/B tumors were enriched for complete loss of *CDKN2A* (E2F targets) (87%; n=40/46) and *TP53* mutations (TP53 signaling) (78%; n=36/46) (Extended Data Fig. 6b).

We then searched for molecular aberrations enriched in Classical-A/B tumors. A hallmark of Classical-A/B tumors is the upregulation of transcription factors (TFs) related to pancreatic lineage differentiation (e.g. *HNF1A*, *HNF4G*, *GATA4*, *GATA6*, *ONECUT2,* and *NKX2-2*; Fig. 3a). However, of these transcription factors, only *GATA6* was found to be recurrently amplified in the genome (16%, n=42/279; Fig. 3b). Moreover, the frequency of copy number gains in *GATA6* was significantly higher in Classical-A/B than Basal-like-A/B tumors (Fig. 3c,d; p=0.0015, Kruskal-Wallis test; Extended Data Fig. 6c,d). There was also a correlation between *GATA6* DNA copy number and gene expression in Classical-A, (=0.47; p=2.2 × 10-6; Spearman’s correlation), Classical-B (=0.69; p=1.6 × 10-5) and Hybrid (=0.42; p=0.002) tumors, but not in Basal-like-A (=0.16; p=0.45) or Basal-like-B tumors (=0.36; p=0.09) (Extended Data Fig. 6e). Notably, many Classical-A/B tumors expressed high levels of *GATA6* but did not have genomic amplifications of the gene. Due to its crucial role in the normal exocrine pancreas20, it is possible that many Classical-A/B tumors simply maintain *GATA6* expression during transformation. We then performed RNA in situ hybridization (RNAish) of benign adjacent tissue. *GATA6* was found to be highly expressed in both duct and acini (Extended Data Fig. 6f). This suggests that the Classical phenotype is the default molecular path in pancreatic cancer pathogenesis. Although conjectural, this idea resonates with the observation that Classical tumors are the numerically most abundant subtype regardless of clinical stage. In summary, we observe that increased frequency of complete *SMAD4* loss and *GATA6* amplification were associated with Classical-A/B tumors; whereas, higher frequency of *TP53* mutations and complete loss of *CDKN2A* were associated with Basal-like-A/B tumors.

***KRAS* imbalance, disease stage and Basal subtype are linked**

Amplification of mutant *KRAS*, which creates an allelic imbalance favoring the mutant allele, has been linked to the Basal-like phenotype in mice21. In our dataset, while Basal-like-B tumors showed increased frequency of *KRAS* imbalance in primary tumors, all subtypes showed similar frequencies in mutant *KRAS* imbalance in metastatic tumors (56~81%; Extended Data Fig. 7a). We investigated whether the high complexity of human tumor genomes was masking the relationship between mutant *KRAS* imbalance and tumor subtypes.

To address this, we first compared the genomic complexity of primary and metastatic diseases. Overall, metastatic disease showed a significantly higher frequency of genome doubling than primary tumors (71% (n=40/56) vs. 47% (n=70/148); p=0.003, Fisher’s exact test) (Extended Data Fig. 7b, left). Importantly, the same molecular subtype was more unstable in metastatic setting compared to primary disease (Extended Data Fig. 7b, c). Genome duplicated (also referred to as tetraploid) tumors commonly lose a quarter of their genomes (ploidy ~ 3.3n, not 4n; ref. 22), resulting in global copy number imbalances. To accurately define copy number imbalances in mutant *KRAS* and account for the high frequency of tetraploidy in metastatic tumors, we assigned patients into three distinct allelic states of mutant *KRAS* (Extended Data Fig. 7d.i and d.ii.): 1) balanced (*KRAS*ba) – equal copy numbers of mutant and wildtype alleles; 2) minor imbalance (*KRAS*mi) – a small imbalance favoring the mutant allele; and, 3) major imbalance (*KRAS*Ma) – a considerable imbalance favoring the mutant allele over the wildtype allele. Using this method, we found *KRAS*Ma was rare in primary tumors (primary – 4% [n=6/141]; mets - 29% [n=17/58]; p=2.8 × 10-6, Fisher’s exact test). Moreover, metastatic Basal-like-A/B tumors were enriched for *KRAS*Ma compared to metastatic Classical-A/B tumors (44% (n=11/25) vs. 14% (n=3/21); p=0.029, Fisher’s exact test; Extended Data Fig. 8a, top row). Classical tumors that harbored *KRAS*Ma also carried other alterations linked to the Classical lineage (Supplementary note). In primary tumors, Basal-like-B tumors were significantly enriched for any imbalance in mutant *KRAS* (*KRAS*mi or *KRAS*Ma) compared to stage-matched Classical-A/B tumors (66.7% (n=8/12) vs. 32.6% (n=30/92); p=0.0282, Fisher’s exact test; Extended Data Fig. 8a, bottom row). As there were only seven primary Basal-like-A tumors, we could not make any statistically powered observations involving them. Overall, the two key takeaway points are: 1) the link between mutant *KRAS* imbalance and the Basal-like phenotype appears to be stage-dependent ⎯ metastatic Basal-like tumors are enriched for *KRAS*Ma whereas primary Basal-like tumors mostly exhibit *KRAS*mi, and; 2) the presence of *KRAS*Ma in some Classical tumor suggest that the entire constellation of genomic aberrations contributes to the molecular phenotype of the tumor. More simply, there are likely multiple genetic paths beyond *KRAS*Ma that produce these phenotypes.

We then assessed the clinical utility of segregating tumors with varying degrees of mutant *KRAS* imbalance. Chemotherapy responses was assessed using RECIST (Response evaluation criteria in solid tumors). We found that Stage IV tumors with *KRAS*Ma were more chemoresistant compared to *KRAS*mi or *KRAS*ba (Fig. 3e,f; p=0.0002, Wilcoxon rank-sum test). For advanced patients, a worse overall outcome was observed for patients with *KRAS*Ma (Fig. 3g, left panel; p=0.0094, log-rank test), and is supported by previous work23. Interestingly, no difference in overall survival was observed between advanced patients with *KRAS*mi and those with *KRAS*ba, suggesting that, at least in part, some of *KRAS*mi may be passenger to genome duplication (Extended Data Fig. 8b; p=0.13, log-rank test). In Stage I/II disease, which are mostly diploid, there was separation in overall survival between each state of mutant *KRAS* imbalance (Fig. 3g, right panel; p=0.019, log-rank test). Importantly, tumors with *KRAS*Ma were not more genomically unstable compared to *KRAS*mi or *KRAS*ba in primary or metastatic disease (Extended Data Fig. 8c). In line with this, neither ploidy nor the number of structural variants (SVs), two independent measures of genomic instability, predicted survival or chemoresistance (Extended Data Fig. 9a-c). This supports that *KRAS*Ma, and not indiscriminate genomic instability, is driving the prognostication. Thus, segregating distinct allelic states of mutant *KRAS* allows identification of patient subgroups with different chemotherapy responses and outcomes.

**Genome duplication promotes major imbalances in mutant *KRAS***

To investigate the mutational processes that drive imbalances in mutant *KRAS*, we mapped the evolutionary trajectory of this specific copy number aberration. To demonstrate the principles of the approach, we have used a single case of liver metastasis (Fig. 4a; methods, Supplementary note). Pcsi\_0729 was a tetraploid tumor with *KRAS*Ma, which occurred through a chromosome arm level event (Fig. 4b; Segment A in top panel). Germline SNP ratios in this region indicated that the wildtype allele was lost, resulting in 4 mutant and 0 wildtype copies (Fig. 4b, bottom panel). Using the copy numbers of mutations on Segment A, we identified 3 distinct clusters of mutations: [**a**] those on all 4 copies (including *KRAS* mutation), [**b**] those on 2 copies, and [**c**] those on 1 copy (Fig. 4c). Mutations labelled as [**c**] are the most recent mutations that occurred after the genome doubled (Fig. 4d, Step 4). Mutations labelled as [**b**], which reside at a copy number state of 2, occurred on a single copy of DNA before the genome doubled (Fig. 4d, Steps 3-4). Mutations on all 4 copies of DNA labelled [**a**] occurred on one of the two original parental DNA strands that underwent copy-neutral LOH (CN-LOH) followed by genome doubling (Fig. 4d, Steps 1-4). Of these three sets of mutations, [**a**] are the earliest and have been clonally accumulated in this region from birth to sample collection. Because there is no *KRAS* wildtype allele in this tumor, we speculate that it was lost after the *KRAS* mutation (part of set [**a**]) as this mutation is an initiating event and likely precedes onset of genomic instability. We then inferred the molecular time of each evolutionary phase using these three sets of mutations [**a**, **b**, and **c**] (Fig. 4e). By dividing the evolutionary trajectory into three distinct phases, we estimate that Pcsi\_0729 spent 69% of its molecular time with balanced mutant *KRAS* (1 mut, 1 wt) as a diploid; 17% of molecular time was spent in *KRAS*mi (2 mut, 0 wt) also as diploid; in the final phase of its evolution, the tumor genome doubled and acquired a *KRAS*Ma (4 mut, 0 wt).

 To investigate whether genome doubling was driving *KRAS*Ma, the above analysis was applied to 48 additional tumors where the mutant *KRAS* imbalance occurred as part of a large copy number alteration. Eleven of these tumors harbored *KRAS*Ma, which were a direct consequence of genome doubling. Moreover, in the entire cohort, nearly all tumors with *KRAS*Ma were tetraploid (Fig. 4f, top row; Fig. 4g). As in Pcsi\_0729, genome doubling likely exacerbates a preexisting minor imbalance to generate a major imbalance. Taken together, this finding implicates genome doubling as a key molecular event underlying *KRAS*Ma in pancreatic cancer evolution.

**Rare subtype switch due to outgrowth of minor clones**

Considering genomic instability in these tumors24, it was possible that subtypes may not remain static with disease progression. In our cohort, 6 patient samples had biopsies from different timepoints (12 tumors). In 4 patients, there was no change in tumor subtype with progression. Accordingly, their genomes were nearly identical at the progression biopsy. In two patients, a pronounced change in the tumor phenotype was observed with progression. Compass\_0003 initially presented with a resectable tumor (Stage I/II - primary) that was classified as ‘Hybrid’ by our scheme and ‘Classical’ by the Moffitt *et al.*14 scheme (Fig. 5a,b; Extended Data Fig. 10a). This patient rapidly progressed to Stage IV (liver met) after resection and did not receive adjuvant therapy. Biopsy from the liver metastasis was highly necrotic and did not yield enough material for RNA-seq. To infer molecular subtype of the metastasis, RNAish for *GATA6*25 was used(Extended Data Fig. 6f). *GATA6* was low to absent in the metastasis suggesting that the tumor phenotype had changed to ‘Basal-like’ (Extended Data Fig. 10b). We could not classify this liver metastasis as Basal-like-A or -B without RNA-seq. WGS was performed from tumors from both timepoints. No significant changes occurred to the number of SNVs/indels or number of SVs (Extended Data Fig. 10c), however, the primary tumor was diploid whereas the metastatic tumor was tetraploid (Fig. 5c). The majority of chromosome 12p was at CN-LOH (copy number = 2) in the primary tumor even though the copy number of mutant *KRAS* was only 1. SVs in this region supported that one of the mutant copies of *KRAS* was lost due to a minor chromothripsis event (Extended Data Fig. 10d). The simplest explanation is that the chromothripsis event must have followed chr 12 CN-LOH. In the liver metastases, the copy number of mutant *KRAS* allele was 4, suggesting this allele had quadrupled with metastatic progression. A single tandem duplication flanking the mutant *KRAS* locus was observed in the metastatic tumor (Extended Data Fig. 10d, black arrow) and likely occurred before genome doubling because all the copy number states on chr12 were in multiples of 2 (i.e. 2/4/6). This suggests that tandem duplication occurred in a diploid clone raising the mutant *KRAS* copy number from 1 to 2. This diploid clone then sustained a genome doubling event bringing the mutant *KRAS* copy number to 4 (Fig. 5d). This shift to a *KRAS*Ma was accompanied by the switch of molecular phenotype to Basal-like. Whether the genome doubling occurred in a subclone after seeding the liver or in the primary tumor was unknown. We xenografted the primary tumor and observed a tetraploid clone that harbored the same tandem duplication from the liver metastasis (Extended Data Fig. 10e,f). This indicates that the origin of this metastatic disease was a Basal-like *KRAS*Ma tetraploid subclone in the primary tumor that was below our detection limit by WGS.

 In a second advanced patient, Compass\_0064 (Fig. 5e,f), liver metastases were biopsied before and after chemotherapy from the same radiologic region. RNA-seq from these two timepoints demonstrated that the diagnostic tumor was ‘Basal-like-B’ whereas the post-therapy tumor was ‘Classical-A’. A shift towards a less aggressive transcriptional phenotype after therapy is not known to occur in other cancers with Basal-like features such as breast cancer26; however, metastatic Classical pancreatic cancer is still a very aggressive disease. WGS from both timepoints showed no major changes in overall mutation status of the tumor (SNVs, indels and SVs; Extended Data Fig. 10c). However, as with Compass\_0003, there was a dramatic shift in tumor ploidy; the ploidy of the diagnostic tumor was consistent with 2 genome doublings (4.3) whereas the ploidy of the post-therapy tumor suggested only a single genome doubling (3.0) (Fig. 5g). Importantly, the lesion before therapy showed a *KRAS*Ma (5 Mut, 2 Wt), but the second tumor was balanced (*KRAS*ba – 2 wildtype, 2 mutant) (Extended Data Fig. 10g). The loss of the *KRAS*Ma status was unexpected, however the *KRAS*Ma was part of an unstable tumor clone1. Again, this shift from *KRAS*Ma to *KRAS*ba was accompanied by a change in subtype (Fig. 5h), likely driven by outgrowth of a minor clone. In summary, the molecular subtype of the tumor can change after surgery or therapy in rare cases. Together, these data support that ongoing genomic instability in the tumor shapes the molecular subtype of the tumor.

**Discussion**

Detailed transcriptomic and genomic analyses have shown that the two accepted subtypes of pancreatic cancer ⎯ Basal-like and Classical ⎯ are still heterogeneous. There are at least two distinct molecular subclusters within each of these subtypes. Basal-like tumors, which are more aggressive, exist as two states that we refer to as Basal-like-A and Basal-like-B. The degree of the squamous expression program (Sig. 2) is the main distinguishing factor that separates these two Basal-like phenotypes. Clinically, Basal-like-A tumors are enriched in metastatic disease whereas Basal-like-B are enriched in resectable disease, supporting that the squamous expression program is selected for with disease progresses. At single cell resolution, most tumors harbor both Basal-like and Classical tumor cells. The varying proportion of these cells creates a transcriptional continuum at the bulk RNA-seq level and Hybrid tumors are an outcome of this issue. Importantly, variables such as cohort size and composition (resectable/metastatic), tumor purity, and the clustering algorithm used to analyze the datasets will continue to impact classification schemes for pancreatic cancer.

 Beyond mouse models21, there have been limited data to indicate that human pancreatic cancer subtypes have a genetic basis. Here we linked copy number events in mutant *KRAS* to the Basal-like phenotype. Both the severity of the imbalance and clinical stage contribute to this relationship. Some degree of mutant *KRAS* imbalance is present in the vast majority of metastatic tumors (~70%), and we speculate that the high frequency of genome doubling in metastases (71%) associated with this phenomenon. There are two ways by which genome doubling fosters *KRAS* imbalance; 1) the generally higher degree of instability in polyploid tumors increases the probability of producing an allelic imbalance in the genome, and 2) minor imbalances can be exacerbated to a major imbalance as the genome doubles with metastatic progression (model in Fig. 6). It is interesting that the Basal-like-B phenotype (low squamous signature) is linked to minor *KRAS* imbalances in early stage disease, which is mostly diploid, whereas the Basal-like-A phenotype (high squamous signature) is linked to major *KRAS* imbalances in late stage disease, which are mostly tetraploid. The relationship between genome doubling, major imbalances in mutant *KRAS* and increased expression of squamous differentiation genes is unknown. Nevertheless, they do suggest increased mutant *KRAS* dosage, which may lead to increased RAS signaling, promotes metastases. RAS ‘addicted’ cells are more epithelial in nature27, which is partly contrary to our findings that major imbalances in mutant *KRAS* show increased EMT. It would be valuable to revisit the concept of RAS addiction in newer culture models28 or in the context of metabolic changes in the tumor29. Beyond mutant *KRAS* dosage, it is important to recognize that each pancreatic cancer carries a large number of genomic events. Given the random nature of genomic instability, one would expect that genomic alterations associated with different tumor programs are occasionally altered together. For example, some *KRAS*Ma tumors also harbor *FOXA2* or *GATA6* amplification, which are related to the Classical phenotype. It is more probable that this constellation of aberrations, rather than a single event, contributes to the transcriptional phenotype of the tumor. This is likely why not all *KRAS*Ma tumors are Basal-like. Almost certainly, non-genetic factors, such as tumor microenvironment30,31 and epigenome32, also play roles in the transcriptional phenotypes but have not been examined here.

 Lastly, our data inform on how disease heterogeneity arise from the same preneoplastic mutational path (*KRAS, TP53, CDKN2A, SMAD4*). As genomic instability is a hallmark of invasive cancer, these copy number events occur at or after preneoplastic changes in the tumor. This helps to explain how different invasive phenotypes emerge in pancreatic cancer. Based on our work, our hypothesis is that the default path of pathogenesis leads to the Classical phenotype. This mainly stems from the observation that pancreatic lineage transcription factors such as *GATA6* are already highly expressed in normal exocrine tissue. A corollary of this hypothesis is that the Basal-like phenotype is acquired in a subset of tumors, and is a hallmark of progression. In conclusion, our data support that genomic events in tumor evolution establishes the molecular subtypes of pancreatic cancer. The pursuit of subtype-based therapy for this lethal disease would benefit from a proper mechanistic understanding of how this collective set of genomic events contribute to the overall tumor phenotype.

**Acknowledgements**

We would like to thank all members of the Notta laboratory and PanCuRx program at the Ontario Institute for Cancer Research (OICR) for their critical review of the manuscript, and Genomics Technology Program at the OICR for technical support. We would also like to thank Francisco Real and Ben Stanger for intellectual input and critical discussion regarding the manuscript. We would also like to acknowledge the participation of patients and their families, and the POG and PanGen teams. This study was conducted with the support of the Ontario Institute for Cancer Research (PanCuRx Translational Research Initiative) through funding provided by the Government of Ontario, the Wallace McCain Centre for Pancreatic Cancer supported by the Princess Margaret Cancer Foundation, the Terry Fox Research Institute, the Canadian Cancer Society Research Institute, and the Pancreatic Cancer Canada Foundation. The study was also supported by charitable donations from the Canadian Friends of the Hebrew University (Alex U. Soyka). SG is the recipient of an Investigator Award from OICR. G.Z. is a clinical research scholar of the Fonds de recherche du Québec – Santé. F.N. is supported by the Gattuso-Slaight Personalized Cancer Medicine Fund, funding from the OICR, the Canadian Institutes of Health Research (#388785) and the Cancer Research Society (#23383). G.Z. is a Clinician–Scientist of the Fonds de la Recherche en Sante du Quebec, P.J.C. is a Wellcome Trust Senior Clinical Fellow, and L.D.S. and S.G. are recipients of Senior or Clinician–Scientist Awards from the OICR.

**Author Contributions**

Data analysis and interpretation: M.C.-S.-Y., J.C.K., S.G., F.N.; sample processing: K.N., E.F.F., J.M., J.M.W., S.-B.L., D.C.; genomics: M.C.-S.-Y., J.C.K., G.W.W., K.N., R.E.D., G.H.J., A.Z., P.M.K., V.S., B.H.-K., L.D.S.; LCM: S.-B.L., D.C., I.L., J.M.S.B.; single-cell RNA-seq: M.C.-S.-Y., K.N.; pathology: S.K., R.C., S.E.F.; sample acquisition and clinical annotation: G.M.O., A.A.C., R.C.G., J.M.W., A.B., G.Z., J.J.K., S.G.; data validation: H.T., F.E.M.F., J.M.K., J.T.T., D.J.R., D.F.S., S.J.M.J, M.A.M., J.L., D.A.T.; manuscript writing: M.C.-S.-Y., J.C.K., F.N.; manuscript editing: E.F.F., H.T., F.E.M.F., P.J.C., D.A.T.; and project lead and supervision: S.G., F.N.

**Competing Interests**

Authors declare no competing interest.

**References**

1. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518,** 495–501 (2015).

2. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *The New England journal of medicine* **372,** 2509–2520 (2015).

3. Yu, J., Blackford, A. L., Dal Molin, M., Wolfgang, C. L. & Goggins, M. Time to progression of pancreatic ductal adenocarcinoma from low-to-high tumour stages. *Gut* **64,** 1783–1789 (2015).

4. Hruban, R. H., Goggins, M., Parsons, J. & Kern, S. E. Progression model for pancreatic cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **6,** 2969–2972 (2000).

5. Jones, S. *et al.* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science* **321,** 1801–1806 (2008).

6. Waddell, N. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491,** 399–405 (2012).

7. Witkiewicz, A. K. *et al.* Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat Commun* **6,** 6744 (2015).

8. Makohon-Moore, A. P. *et al.* Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nature genetics* **49,** 358–366 (2017).

9. Connor, A. A. *et al.* Integration of Genomic and Transcriptional Features in Pancreatic Cancer Reveals Increased Cell Cycle Progression in Metastases. *Cancer cell* (2019). doi:10.1016/j.ccell.2018.12.010

10. McDonald, O. G. *et al.* Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. *Nature genetics* **49,** 367–376 (2017).

11. Notta, F. *et al.* A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538,** 378–382 (2016).

12. Cortés-Ciriano, I. *et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *bioRxiv* **8,** 333617 (2018).

13. Raphael, B. J. *et al.* Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer cell* **32,** 185–203.e13 (2017).

14. Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics* **47,** 1168–1178 (2015).

15. Collisson, E. A. *et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature medicine* **17,** 500–503 (2011).

16. Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531,** 47–52 (2016).

17. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics (Oxford, England)* **26,** 1572–1573 (2010).

18. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171,** 1611–1624.e24 (2017).

19. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3,** 385–394.e3 (2016).

20. Martinelli, P. *et al.* Gata6is required for complete acinar differentiation and maintenance of the exocrine pancreas in adult mice. *Gut* **62,** 1481–1488 (2013).

21. Mueller, S. *et al.* Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes. *Nature* **554,** 62 (2018).

22. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* **45,** 1134–1140 (2013).

23. Krasinskas, A. M., Moser, A. J., Saka, B., Adsay, N. V. & Chiosea, S. I. KRAS mutant allele-specific imbalance is associated with worse prognosis in pancreatic cancer and progression to undifferentiated carcinoma of the pancreas. *Mod Pathol* **26,** 1346–1354 (2013).

24. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–1113 (2010).

25. Aung, K. L. *et al.* Genomics-Driven Precision Medicine for Advanced Pancreatic Cancer: Early Results from the COMPASS Trial. *Clinical cancer research : an official journal of the American Association for Cancer Research* (2017). doi:10.1158/1078-0432.CCR-17-2994

26. Cejalvo, J. M. *et al.* Intrinsic Subtypes and Gene Expression Profiles in Primary and Metastatic Breast Cancer. *Cancer research* **77,** 2213–2221 (2017).

27. Singh, A. *et al.* A Gene Expression Signature Associated with ‘K-Ras Addiction’ Reveals Regulators of EMT and Tumor Cell Survival. *Cancer cell* **15,** 489–500 (2009).

28. Boj, S. F. *et al.* Organoid models of human and mouse ductal pancreatic cancer. *Cell* **160,** 324–338 (2015).

29. Kerr, E. M., Gaude, E., Turrell, F. K., Frezza, C. & Martins, C. P. Mutant Kras copy number defines metabolic reprogramming and therapeutic susceptibilities. *Nature* **531,** 110–113 (2016).

30. Ligorio, M. *et al.* Stromal Microenvironment Shapes the Intratumoral Architecture of Pancreatic Cancer. *Cell* (2019). doi:10.1016/j.cell.2019.05.012

31. Aiello, N. M. *et al.* EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. *Developmental cell* **45,** 681–+ (2018).

32. Roe, J.-S. *et al.* Enhancer Reprogramming Promotes Pancreatic Cancer Metastasis. *Cell* **170,** 875–888.e20 (2017).

**Figure Legends**

**Figure 1: Molecular classification of the disease cohort. a,** Heatmap of the 5 consensus clusters based on tumor transcription signatures deconvoluted by non-negative matrix factorization (NMF). Sigs. 2 and 10 are Basal-related signatures while Sigs. 1 and 6 are Classical-related signatures. Samples were clustered using ConsensusClusterPlus17 (PAM, Pearson correlation distance) for 10,000 iterations. A consensus clustering confidence is shown below the dendrogram (gray scale). Heatmap contains 248 tumors from 242 patients. Expression is shown in transcripts per million (TPM). **b,** Comparing disease clusters in (**a**) to previous classification schemes by Collisson *et al.*15, Moffitt *et al.*14, Bailey *et al.*16 and Puleo *et al.*33. **c,** Pie charts showing proportions of the 5-disease subtype in each clinical stage. **d,** Stacked bar plot of tumor clinical stage for each disease subtype.

**Figure 2: Tracking Basal and Classical-related signatures with scRNA-seq. a,** UMAP plots of 1987 epithelial tumor cells clustered from sample 100070.Known marker genes were used to identify non-epithelial cells (see Extended Data Fig. 2) which were then removed from downstream analysis. 7 single-cell clusters belonging to the tumor were identified in this sample. This workflow was applied to an additional 14 cases (not shown). **b,** Heatmap of marker genes identified in each of the 7 tumor clusters (gene list in Supplementary Table 3). Scores for the tumor (Sig. 1, 2, 6 and 10) and EMT signature are shown for each cell. **c,** *left,* Scatter plot of a representative tumor sample (same as **a**) showing signature 2 score versus signature 6 score with each dot representing a single tumor cell (two-sided Spearman correlation). *right,* Heatmap showing the Spearman correlation coefficients for pairwise comparisons between the 4 tumor signatures (Sigs 1, 2, 6 and 10) across 15 samples. Grey represents correlation without statistical significance (Holm-Bonferroni adjusted p-value > 0.05). (Cl, Classical signature (Sig. 1 or 6); Ba, Basal signature (Sig. 2 or 10)) **d,** *left,* Spearman correlation plot of a representative tumor sample (same as **a**) showing signature 2 score versus EMT score with each dot representing a single tumor cell (Spearman correlation). *right,* Heatmap of Spearman correlation coefficients of EMT score versus each of the 4 tumor signatures (Sigs 1, 2, 6 and 10) across 15 single-cell samples. Grey represents correlation without statistical significance (adjusted p-value > 0.05, same as c).

**Figure 3: DNA copy number analysis of the molecular subtypes. a,** Gene expression levels of common pancreatic lineage transcription factors (TFs). Box whisker plots show median +/- first and third quartiles. p-values are from the Kruskal-Wallis test (n=228 tumors). **b,** Frequency of DNA amplification events in pancreatic lineage TFs from (**a**). Amplification was defined as 2.5x ploidy of the tumor. (n=279 tumors, WGS cohort). **c,** Representative plot of chromosome 18 from a tumor with a *GATA6* amplification. **d,** Frequency of *GATA6* amplification among disease subtypes. Scatter plot depicts *GATA6* copy number normalized to tumor ploidy. Four distinct copy number states are provided (‘Loss’, ‘Neutral’, ‘Gain’, ‘High copy number event’). Copy number was inferred from WGS. P-values are from the Kruskal-Wallis test (n=228 tumors). BA – Basal-like-A; BB – Basal-like-B; Hy – Hybrid; CA – Classical-A; CB – Classical-B. **e,** Representative computerized tomography (CT) images for a patient with major mutant *KRAS* imbalance (top panel) and another where the mutant allele was balanced (bottom panel) pre- (baseline) and post-therapy. **f,** Waterfall plot of best tumor response assessed by RECIST 1.1 criteria in 61 Advanced patients with major (n=13), minor (n=27) or balanced (n=21) alleles for mutant *KRAS*. (p=0.0002, Two-sided Wilcoxon rank-sum test, Major imbalances versus rest). **g,** Survival analysis for patients with advanced disease (Left panel: Stage III/IV; p=0.0094, two-sided log-rank rest), and resectable disease (Right panel: Stage I/II; p=0.019, two-sided log-rank test) based on mutant imbalances in *KRAS*. Survival analysis for advanced disease for all three categories of mutant *KRAS* is shown in Extended Data Fig. 8b).

**Figure 4: Evolution of mutant *KRAS* amplifications. a,** Contour plot of germline SNP ratio and copy number of Pcsi\_0729. Pattern indicates a tetraploid11. **b,** Representative plot of chr12 showing amplification of mutant *KRAS* (at ~ 25 Mb). Amplification of the mutant allele occurs as a large segment of the p arm (Segment A). Copy number (top panel) and germline SNPs ratios (bottom panel) are shown. **c,** Density plot of the mutated copies of somatic mutations on Segment A in Pcsi\_0729. Three sets of mutations are observed [**a,b,c**]. **d,** Representation of how mutation sets [**a,b,c**] were acquired in evolution (Details in Supplementary Fig. 9). **e,** Evolutionary trajectory of mutant *KRAS* imbalance in Pcsi\_0729. Known molecular timepoints (black circles) are calculated based on mutated copies of somatic mutations. When we are unable to determine the exact molecular time due to loss of information (i.e. loss of WT allele), their predicted locations (white circles) are depicted in relation to known events. **f,** *Top row:* Proportion of mutant *KRAS* imbalance (balanced, minor, and major) in diploid and tetraploid tumors segregated by cohort and clinical stage. p-values from fisher’s exact test (Major versus rest). n values for each subgroup shown on the plots. *Bottom row:* Number of structural variants in diploid and tetraploid tumors. Box whisker plots show median and first/third quartiles with min/max points. p-values from the two-sided Wilcoxon rank-sum test. n: 204 (Cohort); 148 (primary); 56 (mets). **g,** Scatter plot of mutant *KRAS* imbalance in diploid (blue dot) and tetraploid (yellow dot) tumors. Each dot represents one tumor (n=204; only *KRAS* mutants used).

**Figure 5: Switch in patient subtype linked to copy number changes in mutant *KRAS*. a,** Compass\_0003, initially a resectable patient (left panel), progressed to Stage IV within 8 weeks of surgery (right panel – liver mets). Fresh tumor tissue from the resection was subjected to RNA-seq, WGS and xenografted (PDX - patient derived xenograft). At progression, patient was enrolled into COMPASS and a liver biopsy was obtained. Due to extreme necrosis, there was only enough tumor material for WGS and RNAish for *GATA6* (Extended Data Fig. 10b).  **b,** Compass\_0003 summary of subtype change from diagnosis to progression (Extended Data Fig. 10a). **c,** Tumor ploidy plots for Compass\_0003. Patterns supports the primary tumor was diploid whereas the liver metastasis was tetraploid. **d,** Trajectory of mutant *KRAS* imbalance and the change in molecular subtype in Compass\_0003. **e,** Compass\_0064, an advanced disease patient treated with Gem/Nab-paclitaxel and responded briefly (18% shrinkage at 16wk by RECIST). Tumor biopsies were obtained before therapy and at progression (22 weeks). Biopsies were radiologically mapped to the same region. WGS and RNA-seq were performed from both timepoints. **f,** Summary of the change in subtype from diagnosis to progression in Compass\_0064 (Extended Data Fig. 10a). **g,** Celluloid plots of tumor ploidy of diagnosis (top panel) and progression (bottom panel) (Compass\_0064). Patterns support that the diagnostic tumor underwent 2 genome duplications (ploidy=4.3), whereas the clone that emerged at progression underwent one genome duplication event (ploidy=3.0). **h,** Evolutionary trajectory of mutant *KRAS* and change in molecular subtype in Compass\_0064.

**Figure 6: Genomic evolution of the molecular subtypes of pancreatic cancer.** A possible model for evolution of pancreatic cancer molecular subtypes. The molecular subtype is a consequence of a gene expression continuum due to a mixture of expression programs from heterogenous cell populations (both Basal-like and Classical cell populations). Molecular subtypes are linked to imbalances in mutant *KRAS*; however, this relationship is dependent on the clinical stage of the tumor as metastatic disease is more copy number unstable compared to primary tumors. In primary tumors, the Basal-like phenotype is linked to minor imbalances whereas in the metastatic tumors, the Basal-like phenotype is linked to major imbalances. For simplicity, alternate genetic alterations linked to the Basal-like phenotype (ex. *KDM6A*, *YAP1*, *TP53*) are not shown.

**Methods**

**Cohort description**

The study cohort included 206 resectable (stage I and II) and 111 advanced (stage III and IV; total 317 patients) pancreatic ductal adenocarcinoma (PDA) patients. All patients signed a written informed consent that allowed the molecular characterization of their tumor samples and the follow up on their clinical information under the International Cancer Genome Consortium (ICGC) protocol. Patient samples were mostly accrued at Princess Margaret Cancer Centre (PMCC) at the University Health Network (Toronto, Canada). Resectable tumors were obtained from the UHN Biospecimens Program and advanced tumors were obtained from the COMPASS trial (NCT02750657). As part of previous studies(9,11,25,34), some resectable tumors were also obtained from Sunnybrook Health Sciences Centre (Toronto), Kingston General Hospital (Kingston), McGill University (Montreal), Mayo Clinic (Rochester), Massachusetts General Hospital (Boston), Sheba Medical Centre (Tel Aviv). Approval for the study was obtained through the University Health Network Research Ethics Board (15-9596, 13-6377 and 32517). A summary of the patient cohort data is provided in Supplementary Table 1. Patient outcomes were analyzed using Kaplan-Meier curves in R or Prism with their associated software packages. The study complies with all relevant ethical regulations.

**Tumor samples**

Resectable tumors were obtained from surgical specimens with confirmed PDA diagnosis. Eligible advanced cases (locally advanced or metastatic) required radiologic or histologic diagnosis of PDA. Cores from advanced PDA were obtained by image-guided percutaneous core needle biopsy. Fresh tumors were embedded in optimal cutting temperature compound and snap frozen in liquid nitrogen and nucleic acids were extracted as previously described 34. Only fresh tumors were used in this study.

**Tumor cell enrichment**

Biospecimens for whole genome sequencing (WGS) (n=330 samples from 314 patients) and RNA-sequencing (RNA-seq) (n=248 from 242 patients; 239 patients have paired WGS) underwent laser capture microdissection (LCM) for tumor enrichment as described previously 11 and 34. For single cell analysis, freshly resected tumors and fresh core biopsies were minced into fine pieces in ice-cold 10 cm tissue culture petri dishes using a razor blade in cold dissociation media (IMDM/2%FBS with 2X collagenase/hyaluronidase mix (Stem cell technologies) and 10 mg/ml DNaseI (Millipore)). Final dissociation volume was brought to 10ml and the sample was placed on a shaker at 4°C for overnight incubation. The next day, the sample was passed through a 100 μm nylon mesh, washed with 30 ml of IMDM/2%FBS, centrifuged and resuspended. The tumor cells were enriched through depletion of CD45+/CD90+/GlyA+ populations using MS columns (Miltenyi Biotec) following the manufacturer’s instructions. After enrichment, cells were resuspended in PBS/0.5% BSA and used for single cell RNA seq.

**RNA in situ hybridization (RNAish) for *GATA6* and *SNAIL2* detection**

RNA in situ hybridization was performed as described 25. Briefly, 4-5 μm tissue sections were obtained from OCT blocks, incubated at 100°C to 103°C for 15mnt with citrate buffer (10nmol/L, pH6) and treated with 10ug/ml protease (Sigma-Aldrich) at 40°C for 20mnt in a HybEZ hybridization oven (Advanced Cell Diagnostics). The probes for *GATA6* (Cat No. 603131) and *SNAI2* (Cat No. 554581) were purchased from Advanced Cell Diagnostics. RNAscope® 2.5 High Definition (HD)-RED Assay was used for further steps according to manufacturer’s instructions. *GATA6* and *SNAI2* signal was scored using a semiquantitative criteria: score 0, absent to rare discernable dots under 40x objective, score 1, few discernable dots at 20x; score 2, dots (4-9/cell) resolve at 10x; score 3, individual dots (more than 10 dots/cell) or cluster resolved at 5x. Samples containing less than a hundred tumor cells were excluded from analysis.

**Bulk RNA-seq**

From LCM tissue, RNA was isolated using PicoPure RNA Isolation Kit. RNA was treated with RNase-free DNase Set and quantified using Qubit dsRNA High Sensitivity kit. Quality measured using both RNA Screen Tape Assay and the 2200 TapeStation Nucleic Acid System. Only RNA with appropriate RNA integrity number (RIN) scores (on average >7) were used to prepare libraries. cDNA libraries were prepared using the TruSeq RNA Access Library Sample prep kit according to the manufacturer. Library pools were quantified on the Eco Real-Time PCR Instrument using KAPA Illumina Library Quantification Kits according to the manufacturer’s protocols. Paired-end cluster generation and sequencing of 2x126 cycles was carried out for all libraries on the Illumina HiSeq 2500 platform.

**Gene expression quantification**

RNA-seq was aligned using Kallisto 0.43.135 with default options and transcript sequences downloaded from Ensembl build 86 genome build**.**The transcript level expression was loaded using tximport36.  The lengthScaledTPM was used with tximport to calculate gene level TPM estimates. Finally, the TPM estimates were normalized using DEseq2 sizefactors and log2 + 1 transformed37. The genes that had ensembl biotypes not in the following list were removed from further analysis: 'antisense', 'lincRNA', 'protein\_coding', 'sense\_intronic', 'non\_coding', 'macro\_lncRNA', 'IG\_C\_gene', 'IG\_J\_gene', 'IG\_J\_pseudogene', 'processed\_transcript', 'IG\_V\_pseudogene' ,'IG\_V\_gene', 'IG\_D\_gene', 'IG\_C\_pseudogene'.

**NMF and consensus clustering**

To resolve expression signatures, non-negative matrix factorization (NMF) was used. To reduce the number of genes used for NMF, genes with a mean expression greater than or equal to 3 and a variance greater than or equal to 1.5 were retained for further analysis. The NMF algorithm used was implemented in the Python package scikit-learn version 0.18.238 was run with the following options: "max\_iter=400, tol=1e-4". For each K-value between 2 and 25, 1000 replicates with different random seeds were executed and the replicate with the lowest mean-squared error was retained for further analysis.  Exemplar genes for each component were chosen by building a N (number of genes) x K (number of NMF components) matrix where element (i, j) is equal to the Pearson correlation between the sample weights for component j and the gene expression for gene i.  Each gene was assigned to the component with the highest Pearson correlation with a minimum value of 0.2 and the top 25correlated genes from each component were used for exemplar genes from each component. To determine the number of optimal components, we used the cophenetic correlation coefficient. We used 12 components instead of 13 as there 13 added an ambiguous pattern that we expect is technical.

Consensus clustering was then performed on tumor expression signatures using the top 100 ranked genes from the 4 malignant related NMF components (1, 2, 6 and 10) using ConsensusClusterPlus v1.40.017. Using the Pearson correlation distance, clustering was performed using PAM (partitioning around medoids) with 10,000 re-samplings to generated the final consensus. Based on silhouette scores, the relative change between k values and the proportion of ambiguous clustering (PAC) score, 5 main clusters were identified.

**Differential gene expression**

Differential gene expression was performed using DESeq2 1.18.137 using recommended settings. Briefly, the kallisto alignments from the entire RNA-seq cohort were imported with tximport followed by a dispersion estimate and a quasi-likelihood negative binomial generalized log-linear model. Differential expression was run in pairwise fashion across the different subtypes as well as hybrid versus the rest.

**Gene set enrichment analysis (GSEA)**

Following differential gene expression, genes were ranked based on the p-value and the sign of the log2 fold change. Gene set enrichment was then performed using GSEA Prerank v3.039. GSEA was run against the hallmarks gene set with recommended settings.

**Additional bulk RNA datasets**

FPKM RNA expression data from 126 samples (TCGA), the normalized counts from 92 patients (ICGC) and counts from 42 advance cases (BC Cancer Personalized Oncogenomics Program (NCT02155621) and PanGen trial (NCT02869802) (POG/PanGen) were obtained and normalized using DESeq2 if applicable. In addition, the normalized RNA-expression data from forty-four malignant organoid samples were obtained from40. The expression profile of the top 100 genes from the 4 main NMF components were examined. The results published here are in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI (http://cancergenome.nih.gov).

**Single-Cell RNA-seq**

For each sample, 10,000 enriched tumor cells were loaded and separated into droplet emulsion using Chromium Single Cell 3’ v2 kit as per the manufacturer’s instructions (10X Genomics). Single cell RNA libraries were prepared using the Single Cell 3’ v2 reagent kit according to the user guide (10X Genomics) and quantified using the Qubit dsDNA High Sensitivity kit (Invitrogen). Library quality was assessed using the LabChip GXII Touch DNA High Sensitivity kit (Perkin Elmer) and were quantified using KAPA Illumina Library Quantification Kit (Roche) according to the manufacturer’s standard protocol. Cluster generation and sequencing was carried out for all libraries on the Illumina HiSeq 2500 platform using HiSeq PE Cluster Kit v4 and HiSeq SBS Kit v4 (Illumina) for High Output mode, or HiSeq PE Rapid Cluster Kit v2 and HiSeq Rapid SBS Kit v2 (Illumina) for Rapid Run mode.

**Single-cell RNA-Seq analysis**

Single cell data was aligned to hg19 using CellRanger v2.1.1 (10X Genomics). The data was further processed using Seurat v3.041. Cells with high mitochondrial expression (>30%), with high ribosomal expression (>50%) or those cells with fewer than 1,000 genes expressed were removed from further analysis. A loess curve was fit to the UMI by gene curve and those cells that were found greater than 2 standard deviations away were removed. Genes with fewer than 5 cells expressing it were removed, along with mitochondria and ribosomal genes. The data was log-normalized and the UMI count, percent mitochondria and cell cycle scores were regressed out. Using the top 2,000 variable genes, principal component analysis (PCA) was performed to identified the number of significant components before clustering using a shared nearest neighbor (SNN) graph. Clusters were identified using a resolution of 0.8. Finally, Uniform Manifold Approximation and Projection (UMAP) was performed

42. Stromal cells were identified using well-established linear markers from fibroblasts, liver, immune cells and endocrine/exocrine pancreatic cells19. Markers for each cluster were identified and compared with the top 100 genes in each of the NMF components. To confirm which clusters were malignant, we analyzed rare reads that captured exon 2 of *KRAS* from 10X scRNA-seq data. By tracing the *KRAS* mutation in those cells, the malignant clusters could be confirmed. The single cell data was further processed using only the clusters identified as malignant. Re-clustering was performed on the malignant cells in a similar fashion noted above. Each of the malignant clusters were scored for the top 100 ranked genes for each of the classical and basal related signatures as well as known EMT related genes 18 utilizing Seurat’s AddModuleScore functionality. Samples were then grouped in “Basal-like only”, “Classical only” and “Mixed” groups.

**Whole genome sequencing**

Whole genome sequencing was performed as described elsewhere11,34. DNA was extracted using Gentra Puregene Tissue Kit components or the Gentra Puregene Blood Kit depending on the type of sample (tumor and buffy coat respectively). Extracted DNA was quantified using Qubit dsDNA High Sensitivity kit. Illumina paired-end libraries were prepared using either the NEBNext DNA Sample Prep Master Mix Set, the Nextera DNA Sample Prep Kit, or the KAPA Library Preparation Kits, following the manufacturers’ protocols. Libraries were quantified on the Illumina Eco Real-Time PCR Instrument using KAPA Illumina Library Quantification Kits according to the manufacturer’s standard protocol. Paired-end cluster generation and sequencing was carried out for all libraries on the Illumina HiSeq 2000/2500 platform using high-throughput (A) 2X101 cycles with TruSeq Cluster kit v3 (Illumina Inc., San Diego, CA, USA Cat #PE-401-3001/FC-401-3001) and (B) 2X126 cycles with HiSeq Cluster kit v4 (Illumina Inc., San Diego, CA, USA Cat #PE-401- 4001/FC-401-4001), combined with rapid run (C) 2X101 cycles HiSeq SBS kits (Illumina Inc., San Diego, CA, US Cat #PE-401-4002/FC-402-4023). Samples were sequenced with the number of lanes predicted to yield a collapsed coverage of 50x and 35x for tumor and normal samples, respectively.

**Germline and somatic genomic read alignment and variant calling**

WGS alignment and variant calling was processed as described elsewhere9,11. Briefly, the data was aligned against human genome build 19 using bwa 0.6.243 and post-processed using picard v1.90 {http://broadinstitute.github.io/picard}. Germline variants were using the Genome Analysis Tool Kit (GATK4 v1.3.16) according to the GATK “best practices” for that version44. Somatic variants were called using 2 tools - Strelka v1.0.745 and MuTect v1.1.446 both of which were run on default settings. The final somatic single nucleotide variants used was the intersect of “Tier 1” SNVs from Strelka and the “PASS” variants from MuTect. In addition, Strelka identified somatic indels. Copy number segments, tumor cellularity and ploidy were obtained by using Celluloid v0.11.411. Somatic structural rearrangements were called as the union of filtered calls from two tools, CREST alpha47 and DELLY v0.5.548 .

**MMR and HRD**

MMR and HRD samples were identified using standard hallmarks as described in9. MMR samples were identified if the somatic SNV load was greater than 35,000 and the somatic indel load was greater than 3000. HRD samples requires at least 4 of the following criteria: SNV loads > 12,000, SNV C>T ratio is < 0.3, 4bp + deletion load is > 200, 4bp+ deletion ratio > 0.4, structural variant loads is > 200, 100 - 10,000 bp deletion load is > 50, 100 to 10,000 bp deletion ration is > 0.45, 10kb to 1mbp duplication load is > 75 and gene hits in *BRAC1*, *BRAC2*, *PALB2* and *RAD51C*. These samples were removed from downstream genomic analysis.

**MutSigCV**

To identify recurrently significant variants, 303 WGS samples were processed using MutSigCV v1.4149 using default parameters. Once recurrent mutations were identified from our whole genome cohort, mutations in those genes were mapped back to each RNA subtype. Fisher tests were performed for each of those genes between Basal-like-A/B and Classical-A/B and Basal-like A vs rest.

**HOTNET2**

To identify groups of mutated genes, points mutations and structural variants from 303 WGS samples were processed using Hotnet2 v1.2.150. Data was processed against irefindex9 with 100 network and 100 heat permutations. All other parameters were set to default values. A final delta of 0.000496943 was used with gene sets of size 3 or greater being significant.

**Genomic variations**

Mutant copies were calculated as previously described in11. Mutant *KRAS* imbalance was defined when mutant copies were greater than wildtype copies + 1. Copy number amplifications were defined as a copy number greater than ploidy + 2.5.

**Inferring evolution timing of CN segment containing KRAS**

Detailed method of inferring evolutionary trajectory of *KRAS* mutation in the context of CN changes in each tumor genome is described in Supplementary Figure 9. Briefly, allele ratios and mutant copies were obtained from combined usage of GATK4 v1.3.1644 and Celluloid v0.11.411. Somatic mutations, called by Strelka v1.0.745 and MuTect v1.1.446, were compiled from a uniform CN segment containing *KRAS*, and mclust51 was used to cluster them by mutant copies. Based on the principle of parsimony, we deduced the model of CN evolution from segment allele CNs, and built an i x j matrix of number of alleles that are present at phase “j” that can give rise to “i” number of final mutant copies. Using this matrix and the number of mutations in each mutant copy cluster, we estimated the amount of relative molecular time spent in each evolutionary phase.

**Statistics**

Details of the statistical analysis used to analyze the data is provided in their respective results section, methods and associated figure legends.

**Data availability**

Raw data is free available from EGA: EGAS00001002543.

**Code availability**

No unique code was developed for this study. R scripts or functions used have been indicated in the relevant results sections, figure legends or in the methods.

**Methods-only References**

33. Puleo, F. *et al.* Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features. *Gastroenterology* **155,** 1999–2013.e3 (2018).

34. Connor, A. A. *et al.* Association of Distinct Mutational Signatures With Correlates of Increased Immune Activity in Pancreatic Ductal Adenocarcinoma. *JAMA Oncol* **3,** 774–783 (2017).

35. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34,** 525–527 (2016).

36. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4,** 1521 (2015).

37. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 550–21 (2014).

38. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

39. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–15550 (2005).

40. Tiriac, H. *et al.* Organoid Profiling Identifies Common Responders to Chemotherapy in Pancreatic Cancer. *Cancer Discovery* **8,** 1112–1129 (2018).

41. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).

42. McInnes, L., Healy, J., 1802.03426, J. M. A. P. A.2018. [1802.03426] UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arxiv.org*

43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25,** 1754–1760 (2009).

44. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20,** 1297–1303 (2010).

45. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics (Oxford, England)* **28,** 1811–1817 (2012).

46. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31,** 213–219 (2013).

47. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature methods* **8,** 652–654 (2011).

48. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)* **28,** i333–i339 (2012).

49. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

50. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics* **47,** 106–114 (2014).

51. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* **8,** 289–317 (2016).