3. Methods

3.1 Pancreatic specimens were obtained and prepared for dissection

A single biopsy, from the tail of the pancreas, was obtained from patient 290B, a 60-year-old female donor with confirmed brainstem death. There was no significant past medical history reported. This anonymous donor was enrolled in the Cambridge Biorepository for Translational Medicine (REC15/EE/0152) and pancreas, bladder and spleen specimens were obtained with full, informed consent. All samples were handled and processed in line with Human Tissue Authority guidelines.

The biopsies were then placed in PAXgene Tissue FIX (PreAnalytiX GmbH, Hombrechtikon, Switzerland), a formalin-free tissue preservative. After 24 hours, the specimens were transferred to PAXgene STABILIZER solution (PreAnalytiX GmbH) and stored at -20 °C. The specimens were then paraffin-embedded by a trained histologist. An Accu-Cut SRM 200 microtome (Sakura Finetek, Leiden, Netherlands) was then used to cut 16 μm thick sections. Consecutive sections were mounted on Arcturus polyethylene naphthalate (PEN) membrane glass slides (Thermo Fisher Scientific, Waltham, MA, USA). These slides were kept at 4 °C until staining.

3.2 Slides were stained with haematoxylin and eosin

Staining with haematoxylin and eosin was carried out in a fume cupboard. Fresh ethanol 70% was prepared prior to starting. All equipment was rinsed in water prior to the staining and each aliquot was freshly made up for this individual staining process, before being appropriately disposed. Each step uses a different aliquot of the reagent, to ensure no contamination occurred from other samples. The staining procedure and timings was as follows:

Removal of paraffin wax and rehydration

1. Mount slides in a slide rack.
2. Place slides in xylene for 2 minutes.
3. Repeat the previous step in a second xylene aliquot for 2 minutes.
4. Place slides in ethanol 100% for 1 minute.
5. Repeat the previous step in a second ethanol 100% aliquot for 1 minute.
6. Place slides in ethanol 70% for 1 minute.

7. Place slides in de-ionised water for 1 minute.

Staining with haematoxylin and eosin

1. Place slides in haematoxylin for 15 seconds.
2. Place slides in tap water for 20 seconds.
3. Repeat the previous step in a second tap water aliquot.
4. Place slides in eosin for 10 seconds.
5. Place slides in a third tap water aliquot for 20 seconds.
6. Place slides in ethanol 70% for 20 seconds.
7. Repeat the previous step in a second ethanol 70% aliquot.
8. Place slides in ethanol 100% for 20 seconds.
8. Repeat the previous step in a second ethanol 100% aliquot for 20 seconds.
9. Place slides in xylene for 20 seconds.
10. Repeat the previous step in a second xylene aliquot.
11. Store samples in a protective box at 4 °C.

3.3 Slides were imaged using the Leica LMD7 Microscope (Leica Microsystems GmbH, Wetzlar, Germany)

Once stained, the sections had a temporary coverslip mounted prior to being imaged, as this produced superior images to unmounted, dry slides. This was performed in a fume hood and involved submerging the PEN membrane slides in Neo-Clear xylene substitute (Merck KGaA, Darmstadt, Germany), and then carefully placing a plastic coverslip over the section ensuring minimal bubbles were formed.

The Leica LMD7 (Leica Microsystems GmbH) was cleaned using Kimtech (Kimberley-Clark Professional, USA) wipes, DNase and 70% ethanol. The mounted slides were then loaded upside down, as this is how they will be positioned during LCM. Images of each individual section were obtained using the proprietary Leica LMD7 software (Leica Microsystems GmbH), at a 10X magnification. These images were invaluable in keeping records of the sections dissected and in retaining the spatial location of each islet excised.

The coverslips were then removed from each slide, again by submerging them in Neo-Clear (Merck KGaA) and gently sliding the coverslip off. The coverslip was promptly

disposed of in the sharps bin. Excess fluid was then removed using Kimtech (Kimberley-Clark Professional) wipes before the slides were placed in a protective box to store at 4 °C.

3.4 Laser capture microdissection was used to excise pancreatic islets

The unmounted, dry slides were loaded onto the Leica LMD7 (Leica Microsystems GmbH) with the PEN membrane (Thermo Fisher Scientific) side facing the ground. An Eppendorf twin.tec LoBind 96-well skirted PCR plate (Eppendorf AG, Germany) was then sterilised with UV radiation for 20 minutes, using the UVP Crosslinker (Analytik Jena AG, Germany). The sterilized plate was then loaded onto the Leica LMD7 (Leica Microsystems GmbH). The laser settings, on 10X magnification, were defined (Table 1) and laser calibration carried out.

Table 1 – The laser settings used in the LCM process

Set 1 is the primary setting and should be used first to appropriately excise the sample from the tissue. If the excised tissue fails to drop into the well, the more powerful set 2 can be used.

Laser setting	**Set 1**	**Set 2**
Power	35	35
Aperture	2	20
Speed	1	20
Line spacing	12	12
Head current	100%	100%
Pulse frequency	120	120
Offset	50	50
Specimen balance	0	0

Using the images obtained in the previous step, individual islets were then demarcated and labelled with the well number that they would be cut into, using the touchscreen interface on the Leica LMD7 (Leica Microsystems GmbH). LCM was then performed using the proprietary Leica LMD7 software (Leica Microsystems GmbH).

Laser capture microdissection (LCM)

1. Ensure the desired well is chosen before beginning the laser microdissection.
2. Select the “draw and cut” option and outline the islet to be excised using the touchscreen interface of the Leica LMD7 (Leica Microsystems GmbH).
3. Click “cut” to apply the laser to the outlined region. This will excise the islet.
4. Often due to static forces or incomplete tissue penetration by the laser, the dissected islet may not initially drop into the well, instead remaining attached to the slide. In this case, perform the following:
    a. Freehand cutting of any tissue that appears to be holding the cut tissue in the specimen, using the “set 1” laser setting.
    b. Individual brief pulses of the laser on loose parts of the cut tissue using the “set 2” laser setting.
5. Repeat this process in the same well, over several z-slices, to increase DNA yield per well.
6. For duplicates and triplicates, excise the same islet into different wells.

3.5 Excised tissue underwent protein digestion prior to whole-genome sequencing

Once the islets have been excised, protein digestion was then carried out to lyse the cells, allowing DNA extraction. This used the Arcturus PicoPure DNA Extraction Kit (Thermo Fisher Scientific) and is detailed below.

Protein digestion and DNA extraction

1. Prepare the Arcturus PicoPure DNA Extraction Kit (Thermo Fisher Scientific):
    a. Briefly spin down the tubes containing 150 µg of Proteinase K (Thermo Fisher Scientific), in a microcentrifuge at full speed (4000 g).
    b. Add 150 µL of the provided reconstitution buffer (Thermo Fisher Scientific) to each tube to produce a 1 µg/µL solution.
    c. Pipette the buffer-enzyme solution (Thermo Fisher Scientific) up and down gently.
    d. Vortex the buffer-enzyme solution (Thermo Fisher Scientific).
    e. Centrifuge the buffer-enzyme solution (Thermo Fisher Scientific) at full speed for 5 seconds.
    f. Add 20 µL to each well, keeping wells covered where possible with a sterile foil card.

g. Place strip caps on the wells.

2. Load the Eppendorf twin.tec LoBind 96-well skirted PCR plate (Eppendorf AG) into a centrifuge for 1 minute at 1500x.
3. Place the Eppendorf twin.tec LoBind 96-well skirted PCR plate (Eppendorf AG) onto the thermocycler using the following program (Table 2):

Table 2 – The thermocycler program used during protein digestion with Arcturus PicoPure DNA Extraction Kit (Thermo Fisher Scientific)

This program was modified from the manufacturer's recommendations, found at https://assets.thermofisher.com/TFS-Assets/LSG/manuals/cms_086062.pdf. The alternative program reduced the high temperatures recommending for inactivating proteinase K and instead used a longer inactivating step at a lower temperature.

Step	Temperature	Duration
1	65°C	3 hours
2	75°C	30 minutes
3	4°C	Hold

4. Store the cell lysate at -20 °C until library preparation.

3.6 Whole-genome sequencing of the pancreatic islets

DNA libraries were then generated from the low amounts of DNA using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA). This involved a shearing stage, generating a mean insert size of ~350 base pairs, followed by end repair and adaptor ligation. This process avoids the need for whole-genome amplification. The DNA libraries then went through 12 cycles of polymerase chain reaction (PCR) and the concentrations were quantified with the Qubit fluorometer (Thermo Fisher Scientific).

Two criteria were used to decide which samples would be whole-genome sequenced. The first was that the DNA concentration exceeded 20 ng/μL. This was to ensure there was sufficient DNA to produce a complex library of genomic DNA and avoid excessive

PCR duplicates. The second criterion was the confirmation of the islet histology by a trained clinical histopathologist.

Whole-genome sequencing (WGS) was then performed with the Illumina HiSeq X™ Ten system (Illumina Inc, San Diego, California, USA), through sequencing-by-synthesis. Paired-end reads of 150 base pairs in a single lane run were used, with the aim of achieving an effective coverage of 30X. Upon completion of sequencing, BWA-MEM (v0.7.16, https://github.com/lh3/bwa) (Li & Durbin, 2009) was used to align reads to the GRCh37 (hg19) build of the human genome. All genome coordinates described relate to this build. Duplicates were identified using biobambam (v2.0.54, https://github.com/gt1/biobambam) (Tischler & Leonard, 2014).

3.7 A somatic variant caller was used to generate the matched and unmatched calls

Somatic variant calling was undertaken with CaVEMan (v1.11.2, https://github.com/cancerit/CaVEMan) (Jones et al., 2016) using default parameters. Key conditions were to only accept a variant if it is present in greater than three reads at that site. The matched data was a whole-genome sequence, obtained from the bladder urothelium of the same patient (sample PD37726b_lo0071). Performing variant calling against a matched normal sample is the traditional way of identifying somatic mutations, as this removes mutations shared between both samples as germline. However, in doing so, matched variant calling removes those early embryonic, mosaic mutations that are present in both the pancreatic islets and the matched bladder sample. Since embryonic mutations are of particular interest for this project, we also performed an unmatched variant calling using a synthetic, unrelated normal sample as the comparison, retaining germline mutations as well as embryonic mutations.

In-house filters were then applied to both data sets to remove artefacts known to occur during the LCM pipeline. This filtering step was designed by Mathijs Sanders. LCM-related artefactual variants tend to co-occur with additional nearby variants and have been shown to arise in reads containing inverted repeats with similar alignment start positions. The origin of these variants has been modelled *in silico* and attributed to mismatched base pairing in DNA hairpin loop structures. Detecting these variants is

based on proximity of the variant to the alignment start site as well as the standard deviation and the median absolute deviation of the variant position, within the supporting reads. These statistics were calculated separately for positive and negative strand aligned reads. With sufficient supporting reads that have similar alignment starts, variants were retained if other reads demonstrated strong measures of variance.

In silico re-genotyping was then undertaken in the unmatched data using CGPVAF (part of vafCorrect, v.5.3.8, https://github.com/cancerit/vafCorrect). Ten matched bladder urothelium samples were also included in this re-genotyping. Variants that had passed CaVEMan in some samples, but not others, then had their VAF calculated in each sample they were present in, even if based on one read.

3.8 Copy number analysis was performed to assess for losses and gains

To ensure no copy number changes or loss of heterozygosity events, copy number analysis was performed on all 32 samples. ASCAT (Allele-Specific Copy number Analysis of Tumours, v4.0.1, https://github.com/cancerit/ascatNgs) and Battenberg (v.3.0.1, https://github.com/cancerit/cgpBattenberg) were both used (Nik-Zainal, Van Loo, et al., 2012; Raine et al., 2016; Van Loo et al., 2010). The bladder urothelium sample, PD37726b_lo0071, was used as the matched normal. While ASCAT depends on single nucleotide polymorphisms (SNPs) to calculate allele-specific copy numbers, Battenberg uses haplotypes (phased SNPs) to determine allelic ratios, making it preferable in sub-clonal populations (Nik-Zainal, Van Loo, et al., 2012; Raine et al., 2016; Van Loo et al., 2010). Together, the two complement each other and provide a more complete copy number analysis.

3.9 A mean VAF filter was applied to remove the germline variants

Computational analysis was undertaken with the R programming language (v.3.5.0, http://www.R-project.org) (R Core Team, 2018) and RStudio (v1.1.453, http://www.rstudio.com/) (RStudio Team, 2016). The first filter applied was to retain only variants with a mean VAF less than 0.4. This was applied to both the matched and unmatched data, with the motivation being to remove most germline SNPs, since these will be expected to have VAFs tightly clustered around 0.5.

3.10 The beta-binomial distribution identified over-dispersed somatic variants

Removing variants with a mean VAF across all samples equal to, or higher than, 0.4 should remove most germline SNPs, while retaining early embryonic mutations. However, some germline SNPs and low-frequency artefacts will be retained with this filter. To distinguish between these and genuine somatic variants, we developed a novel approach based on fitting a beta-binomial distribution to the number of reads supporting a mutation across samples. A given germline SNP or low-frequency artefact will be expected to affect all libraries similarly, with variation in the number of supporting reads mostly reflecting binomial sampling. Instead, genuine somatic mutations will be expected to vary considerably in their contribution to different areas of tissue. This can be quantified using the over-dispersion parameter of the beta-binomial distribution, with genuine somatic mutations expected to show a large degree of over-dispersion across libraries. This analysis was undertaken with the R package VGAM (v1.0-5, https://www.stat.auckland.ac.nz/~yee/VGAM/) (Yee, 2015), in collaboration with Tim Coorens.

The estimation range of the over-dispersion parameter, ρ, for each variant was bounded between 10^{-6} and 0.89, using a grid search with 0.05 intervals to obtain approximate maximum-likelihood estimates. The resulting distribution of ρ values across candidate mutations was then plotted as a histogram revealing a clear separation between highly over-dispersed variants and lowly over-dispersed variants. A cut-off ρ value was then chosen following manual inspection of the histogram, to retain over-dispersed variants as those likely to be somatic.

3.11 A depth filter ensured sufficient read numbers supported variants

A depth filter was subsequently applied to both data sets, after the mean VAF and beta-binomial filter. The purpose of filtering by coverage was to reduce the chance of a sampling bias being the reason a variant was called as somatic. Only variants with a mean coverage greater than 20X, across all samples, were retained.

3.12 Estimation of the observable mutational burden per cell

To estimate the average number of detected mutations per cell in a sample, the equation below can be used (Martincorena et al., 2015). This equation uses the allele

frequencies of each detected mutation to estimate the fraction of cells that carry the mutation, assuming a diploid copy number. Summing these fractions across all mutations produces an estimate of the observed mean mutational burden per cell, rather than per islet (Martincorena et al., 2015).

$$B = 2M\bar{f} = 2\sum_{i=1}^{M} f_i$$

Where B = Mutational burden, M = Total number of detected mutations, f *= VAF*

It is important to note that this estimate is restricted to observed mutations. In highly polyclonal samples, only a small fraction of all mutations may reach sufficiently high VAFs to be detectable and thus this calculation represents a lower bound estimate of the true number of somatic mutations present in each cell of a sample.

3.13 Mutational signature analysis identifies distinct mutational processes active in a sample

Identifying a mutational signature requires first preparing the data to a standardised format. By convention, the base substitutions refer only to the pyrimidine base (C and T) and each base substitution is displayed in the context of the 5' and 3' base on either side of it. This produces a matrix of 96-trinucleotide combinations, across six substitution types. Through non-negative matrix factorisation, the distinct mutational patterns can be extracted and fitted using prior knowledge of the 49 known single base substitutions (SBS), identified by the Pan-Cancer Analysis of Whole-Genomes Network (PCAWG) (Alexandrov et al., 2018). A multiple linear regression model can then weigh each signature against each other, to reveal their proportional influences in each sample, using the R package deconstructSigs (v1.8.0, https://github.com/raerose01/deconstructSigs) (Rosenthal, McGranahan, Herrero, Taylor, & Swanton, 2016).

3.14 Assessment of clonality using variant allele frequencies

The clonality of a sample can be studied using histograms to visualise the VAF distributions. Monoclonal and polyclonal samples can then be differentiated by their

respective distributions and mean values. A clonal sample, one where each cell within the islet derives from a recent common ancestor, is characterised by a binomial distribution centred around 0.5. Samples with large subclones can take on multimodal distributions while highly polyclonal samples tend to be dominated by rare variants.

3.15 A phylogenetic tree of pancreatic islet development can be reconstructed using data clustering algorithms

An n-dimensional hierarchical Dirichlet process (n-HDP) was used to cluster variants (Appendix 8.1) (Gundem et al., 2015; Teh, Jordan, Beal, & Blei, 2006). This algorithm was written by Peter Campbell. The reasoning behind using the n-HDP is that mutations which have occurred in the same embryonic cell will have a consistent VAF across different samples. Clusters, or groups, of mutations can then be identified by clustering the VAF profiles of all the mutations across samples, over numerous iterations. The optimal solution will be the one that places the most mutations, with the highest probabilities, into clusters. Each cluster can then be represented in each islet as a proportion of cells carrying the mutations found in that cluster. The pigeonhole principle can then identify whether the clusters within the islets are mutually exclusive or nested. From this, the branches on a phylogenetic tree can be drawn depicting the relationship between the inferred clusters or lineages (Gundem et al., 2015; Nik-Zainal, Van Loo, et al., 2012).

The visualisation of the individual phylogenetic trees for each sample was performed using the R package ggtree, (v1.12.0, https://github.com/GuangchuangYu/ggtree) (Yu, Smith David, Zhu, Guan, & Lam Tommy, 2016). This was done in collaboration with Tim Coorens. By overlaying the phylogenetic lineages onto the spatial locations of the islets in the section, the distribution of embryonic lineages in the tissue can be seen.