Somatic mutations in the pancreatic islets

Supervisor: Iñigo Martincorena

Pantelis Andreas Nicola

King's College

University of Cambridge

July 2018

This dissertation is submitted for the degree of
Master of Philosophy

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

It does not exceed the prescribed word limit prescribed by the Degree Committee for the Faculty of Biology. The word count is 19,956 exclusive of tables, footnotes, bibliography and appendices.

Abstract

The endocrine pancreas is composed of the islets of Langerhans. These micro-organs play a crucial role in glucose homeostasis by producing and regulating insulin and glucagon secretion. Dysfunction of these islets is part of the pathogenesis of diabetes mellitus. The global health burden of diabetes mellitus is growing, with an estimated one in eleven adults affected worldwide. A better understanding of the development and maintenance of the pancreatic islets could prove crucial to reversing this trend.

Whilst somatic mutations have been studied extensively in tumours, the exploration of normal tissue is still in its infancy. Here I present a novel workflow, using laser capture microdissection, whole-genome sequencing and innovative bioinformatics to accurately identify somatic mutations in healthy pancreatic islets. By sequencing 32 islets, all from a single individual, this work reveals islets to be polyclonal units formed by multiple embryonic founding lineages that often come to be dominated by one or two major lineages. The very low mutational burden observed here suggests these islets expand very early in embryogenesis and do not undergo large clonal expansions later in life. This is consistent with the islets being a slowly dividing tissue during adulthood, making it less likely that islet neogenesis or replenishment by a small number of progenitors, occurs in adults.

The 32 islets sequenced here all share a single common ancestor, one that also gives rise to the bladder urothelium. This is presumably the first cell that gave rise to all adult tissues. Spatially, nearby islets are more genetically similar than distant islets. This pattern demonstrates that different embryonic lineages contribute disproportionately to the islets in different areas of the pancreas, the implication being that this emerges during embryonic pancreatic development, or in adulthood through hypothetical islet fission events.

The translational potential of this line of work is substantial. Using this approach to understand the maintenance of islets in diabetes could yield a greater understanding of the pathogenesis, and whether somatic mutations could play a role in the disease. Applications to other normal tissues could similarly refine our knowledge of their development, maintenance and disease, with exciting prospects for clinical application.

Acknowledgements

This work could not have been possible without the kind generosity of the transplant donor, who so altruistically donated their tissues for research. The surgical team, led by Kourosh Saeb-Parsy at Addenbrooke's Hospital, Cambridge, collected the donor specimens and Yvette Hooks processed and sectioned them. The low-input sequencing pipeline was designed by Pete Ellis while the laser capture microdissection pipeline was developed by Luiza Moore, who also gave an expert histological opinion on all samples used in this project. Laura O'Neill was responsible for managing the sequencing submissions, which was undertaken by the Sanger Sequencing Core. The Cancer, Ageing and Somatic Mutation (CASM) IT department were central to the data analysis used here, as was Mathijs Sanders, who crafted the in-house variant filter used in this work.

CASM, and the Wellcome Sanger Institute, has been a stimulating environment full of friendly colleagues and I thank all those to whom I've spoken. Alex Cagan and Andrew Lawson have both been instrumental throughout this work and have shown themselves to not just be exceptional scientists, but natural-born mentors. Federico Abascal proved to be an encouraging teacher and I thank him for his help with the clustering analysis (section 4.10 and 8.4). Tim Coorens has been the backbone of this project and his computational input is omnipresent throughout this work, particularly in the design of the beta-binomial filter (section 3.10) and the production of the phylogenetic trees (section 3.15 and 4.13).

My thesis committee, made up of Peter Campbell and Kourosh Saeb-Parsy, provided critical insights for which I'm grateful. Additionally, I thank Peter Campbell for writing the clustering algorithm used here (section 3.15 and 8.1). My supervisor, Iñigo Martincorena, deserves the highest praise for taking me on with such passion and nurturing my development as a budding clinician-scientist. The lessons I've learnt from him, not just in research, will stay with me through the years.

Finally, I am humbled by the generous support that King's College, Cambridge and the Isaac Newton Trust have provided me with this year. To my family, friends and loved ones, you have all been unparalleled in your unconditional love and support and I dedicate this work to you all, with the hope that one day I can repay the sentiment.

Highlights

1. The whole genomes of 32 pancreatic islets were sequenced from a single human donor.

2. An unmatched analysis proved efficient in the removal of germline variants and artefacts, as well as accurately calling somatic mutations, including those occurring in early embryonic development.

3. The observed somatic mutation burden in the pancreatic islets is low and is driven by intrinsic mutational processes.

4. Almost all somatic mutations identified appeared to have no functional impact.

5. Islets are polyclonal units.

6. Pancreatic islets and bladder urothelium share the same most recent common ancestor.

7. Multiple embryonic founders establish each pancreatic islet but islets develop major and minor lineages.

8. Pancreatic islets do not appear to be maintained by a rapidly-dividing stem cell population, but whether there are multiple stem cell populations, or self-duplicating islet cells, needs further study.

9. The spatial distribution of islets and their embryonic lineages reveals their founding cells are non-randomly distributed.

10. There are potential applications of this work to the fields of tissue development, maintenance and disease. The possible role of somatic mutations in diabetes mellitus is a target of future research.

Contents

1. Introduction

1.1 Somatic mutations are acquired throughout life

Mutations are changes in the deoxyribonucleic acid (DNA) sequence. These can occur in the germline or the soma, to differing effects. Germline mutations are those mutations present in the haploid genomes of the gametes, that go on to fuse at conception. As a result, these mutations are inherited from the parental generation and are present in all the descendant cells of the totipotent zygote. Somatic mutations are those that occur any time after zygote formation and are not inherited from parental DNA. From the moment of conception, the zygote is under mutagenic pressure from intrinsic mutational process, one example being DNA replication errors. As the first cells undergo rounds of cleavage, each division is an opportunity for further mutations to occur. With time and exposure, extrinsic mutational pressures such as ultraviolet (UV) radiation and tobacco smoking, come to play a part in the development of somatic mutations (Stratton, Campbell, & Futreal, 2009).

Depending on their functional impact, somatic mutations can be classified as drivers or passengers (Figure 1) (Stratton et al., 2009). Drivers are a very small minority of somatic mutations that provide a phenotypic benefit to the cell. Often, these drivers are non-synonymous coding mutations, although some in non-coding regions of the genome can also act as drivers (Horn et al., 2013; Huang et al., 2013). The phenotypes bestowed upon the cell overlap with the hallmarks of cancer and include sustained proliferative signalling, enabling replicative immortality and resisting cell death (Hanahan & Weinberg, 2011). Due to the conferred growth advantage, cells with drivers have a relative gain of fitness over their neighbours, leading to positive selection and clonal expansions, by Darwinian evolution. This represents a critical step in carcinogenesis, as these drivers become causally implicated in the emergence of a future tumour (Stratton et al., 2009). Passengers on the other hand do not result in a growth advantage for the cell. These mutations include nearly all non-coding variants and the vast majority of coding mutations in genes not implicated in cancer. Passengers essentially “hitchhike” with those drivers that power a clonal expansion, as they too are present in the genome that is being positively selected for (Stratton et al., 2009).

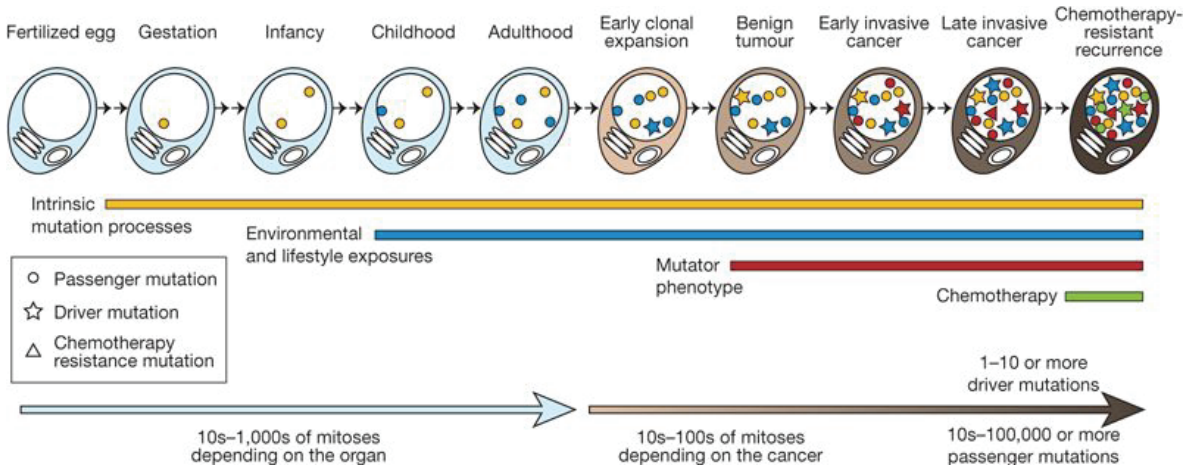


Figure 1 – Somatic mutations occur throughout life

Driver and passenger mutations accumulate throughout life. Drivers lead to clonal expansions that can emerge as tumours. There are several different mutagenic processes that contribute to these somatic mutations (Stratton et al., 2009).

1.2 The somatic mutational burden in normal tissues is comparable to some tumours

The mutational burden is the observed number of mutations that have been progressively acquired by a cell population. With increasing numbers of sequenced cancer genomes, the mutational burden across cancer types has become increasingly well-studied. Across 2,583 donors in the Pan-Cancer Analysis of Whole Genomes network (PCAWG), 43,778,859 single nucleotide variants (SNVs) have been detected across an array of tumour types, revealing an unprecedented insight into the typical numbers of mutations per tumour. These mutational burdens range from 10,000 to 100,000, in some of the most highly mutated cancer types such as UV-associated melanoma and tobacco-induced squamous cell lung cancer, to as low as 100 mutations per genome, in some bone and brain cancers (Campbell, Getz, Stuart, Korbel, & Stein, 2017).

Understanding the somatic mutations that arise early in cancer development, perhaps before the “mutator” phenotype exists, requires an in-depth look at healthy tissue. Prior to neoplastic transformation, healthy tissue would be expected to be harbouring somatic mutations and possibly the first driver, that can eventually lead to cancer (Stratton et al., 2009). Using deep, targeted sequencing of 74 known cancer genes, in 234 biopsies of healthy skin samples, Martincorena et al., (2015) revealed mutational

burdens averaging two to six base substitutions per megabase. This equates to genomes of a normal skin biopsy harbouring up to 30,000 mutations. Remarkably, in these histologically normal skin samples, the mutational burdens across all four patients were comparable to those seen in some skin cancers and several solid tissue malignancies (Martincorena et al., 2015).

1.3 The mutational signatures in normal tissues and cancer give insight into mutational processes

Elucidating the processes that drive the accumulation of somatic mutations provides insights into carcinogenesis. Mutations can be classified according to different features with different mutational processes often inducing distinct patterns of somatic mutations. These patterns can be used as fingerprints, or signatures, of mutational activity. The key features involved in modelling signatures were set out by Alexandrov et al., (2013):

1. The type of mutations observed, such as single base substitutions, insertions/deletions or chromosomal rearrangements;
2. The local sequence context, such as the bases that precede and follow a base substitution;
3. The location of the mutations throughout the genome, such as in particular regions susceptible to a certain mutagenic process or spatial clustering of mutations;
4. DNA damage repair mechanism involvement, as this leaves tell-tale marks on the DNA sequence and contributes to mutagenesis itself.

With increasing amounts of data and new analytic methods, the list of mutational signatures has continued to grow, from an initial 22, to the COSMIC-30 and recently the PCAWG-65 (Alexandrov et al., 2018; Forbes et al., 2017; Nik-Zainal, Alexandrov, et al., 2012). The PCAWG-65 mutational signatures are based on 84,729,690 somatic mutations, with 49 of these signatures relating to single base substitutions (SBS) (Alexandrov et al., 2018). Many different aetiologies, occurring in numerous cancer types, have been assigned a mutational signature including smoking tobacco and defective DNA damage repair due to BRCA 1/2 mutations (Alexandrov et al., 2018).

Despite these signatures having been defined in cancers, normal tissue also displays evidence of mutational signatures. For example, normal sun-exposed skin displays a high burden of C>T mutations at dipyrimidine sites caused by transcription-coupled repair of UV-induced DNA damage (Martincorena et al., 2015). This results in the high prevalence of the mutational signature SBS7a being found in histologically normal tissue, having previously been well-documented in UV-associated melanoma, as well as head and neck squamous cell carcinoma (Figure 2) (Alexandrov et al., 2018). The implication being that analysis of mutational signatures in normal tissues can shed light on the mutational processes driving precancer evolution.

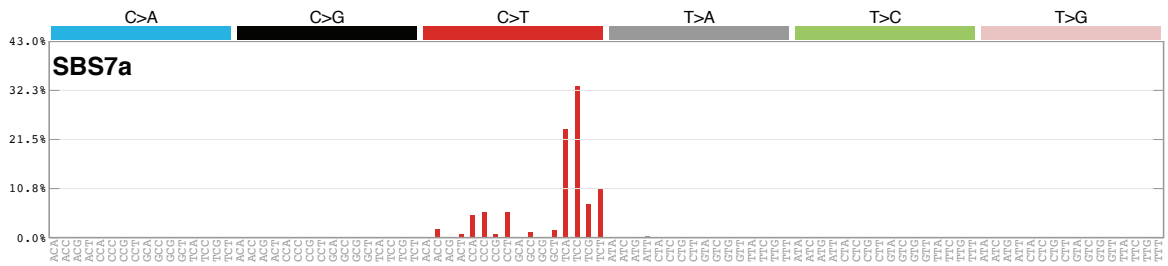


Figure 2 – Mutational signatures each have a unique profile of mutations
A 96-trinucleotide bar plot showing mutational signature SBS7a, one of the PCAWG-65 (Alexandrov et al., 2018). Each single base substitution is shown in the context of the pyrimidine bases involved, with the 5' and 3' bases included to make a trinucleotide. SBS7a is associated with UV light shows an excess of C>T substitutions, particularly in the context of TpCpA and TpCpC (Alexandrov et al., 2018).

1.4 The clonality of a tissue sample can be estimated using the somatic mutations present

The fraction of DNA molecules, within a sample, that harbour a given mutation is termed the Variant Allele Frequency (VAF). For example, inherited germline heterozygous mutations present in all diploid cells of the body will show VAFs around 0.5. This is because one of the two copies of the genome in every cell contains the mutant allele (Figure 3A). In contrast, somatic mutations occur once the zygote has been formed and are only present in a fraction of all somatic cells in an individual (Figure 3B). Somatic mutations occurring in the first few divisions of the embryo can appear in a considerable fraction of cells in the adult and are often termed mosaic

mutations, while late occurring mutations are typically constrained to small clones within a tissue.

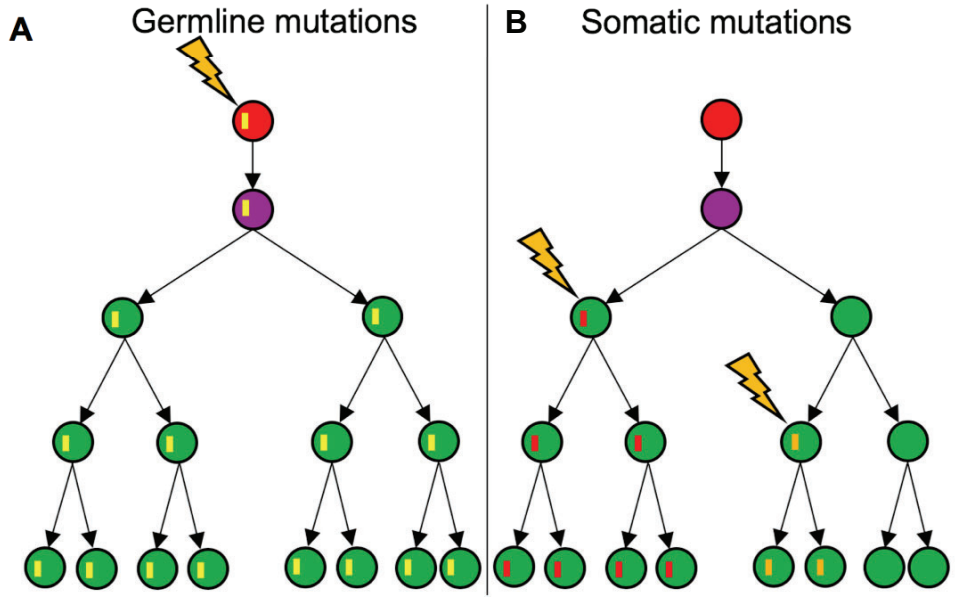


Figure 3 – The VAF of a mutation can be used to assess clonality

Phylogenetic trees displaying the ancestry of a cell population. Mutations (yellow lightning strike) are passed onto progeny. The parental lineage is shown by the red circle and the fertilised egg is the purple cell. Somatic descendants are shown by the green circles. Heterozygous variants are shown by the coloured bars in circles.

(A) A heterozygous germline mutation results in a VAF of 0.5 as they are present in all cells.

(B) Heterozygous somatic mutations will produce a variety of VAFs, according to how early they occur in development and to what degree a tissue is composed of lineages carrying the somatic variant. The exception to this is a somatic mutation occurring in the fertilised egg itself, which would give a VAF of 0.5.

The clonality of a sample can be studied by analysing the VAF distribution of all the mutations within a sample. Heterozygous variants in a clonal sample, one where all cells carry the same mutations and are thus closely related, would be expected to show binomial variation around a VAF of 0.5, assuming a diploid genome. Colonic crypts are a well-known example of a clonal tissue. Although each crypt contains multiple stem cells, by mere drift, single stem cells frequently take over a crypt

(Snippert et al., 2010). This leads to all cells of a crypt recently deriving from the same stem cell and manifests as a VAF distribution centred around 0.5. Contrasting this, a polyclonal sample would have fewer mutations at a high VAF as different cells carry different mutations, each representing a small proportion of the sample. Within the sample VAF distribution, these subclonal populations may then be represented as multiple peaks each with mean VAFs less than 0.5 (Figure 4) (Nik-Zainal, Van Loo, et al., 2012).

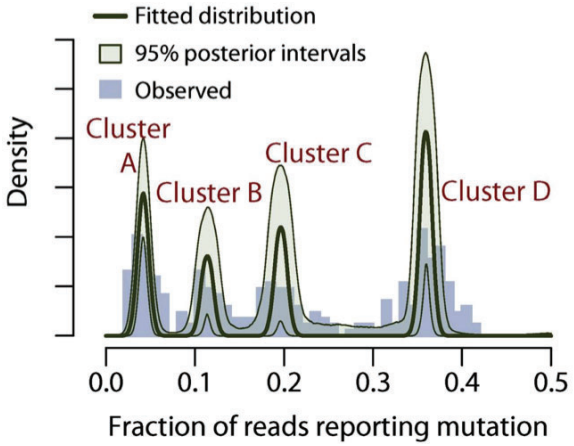


Figure 4 – The VAF distribution reflects the clonality of a sample
A polyclonal sample has numerous clusters of VAFs that represent subclones in the cell population (Nik-Zainal, Van Loo, et al., 2012). A Bayesian Dirichlet process has been used to model the VAF distributions with 95% posterior confidence intervals displayed in green. Four subclones are present, with cluster D being the dominant lineage, as it has the highest VAF (Nik-Zainal, Van Loo, et al., 2012).

1.5 Phylogenetic tree reconstruction of early embryogenesis has been demonstrated in clonal organoids

Utilising somatic mutations to reconstruct a phylogenetic tree, in a clonal tissue, has previously been demonstrated in mice (Behjati et al., 2014). Clonal organoids derived from the stomach, small and large bowel and the tail of two mice, were sent off for whole-genome sequencing and the variant caller, CaVEMan (Cancer Variants through Expectation Maximization) was used to identify somatic mutations (Behjati et al., 2014; Nik-Zainal, Van Loo, et al., 2012). Initially, this variant calling was performed with a matched tail sample to ensure efficient removal of germline variants. A subsequent unmatched run captured the entire complement of germline and detectable somatic

mutations and by comparing this to the matched run, the germline mutations could again be removed, leaving behind those variants exclusive to the unmatched run. After capillary sequencing of these exclusive variants, 35 were confirmed. As they are shared between the organoids and the matched tail sample, they likely occur early in embryonic development.

Maximum parsimony was then used to reconstruct a phylogenetic tree detailing the hypothetical order of mutation acquisition (Figure 5) (Behjati et al., 2014). This totalled 23 cell divisions across two trees, one from each mouse. Both were resolved to a single ancestral origin and although this first cell may be the zygote, the possibility of silent cell divisions and lack of statistical power in distinguishing real differences in read counts, means it isn't certain that this is the case. The mutation rate in early embryogenesis was estimated at 1.5 mutations per cell division. Importantly, the reconstructed tree represents the earliest divisions in the embryo and pre-dates gastrulation, confirming that germ layers are polyphyletic in origin, formed by the spatial aggregation of cells from different lineages (Behjati et al., 2014).

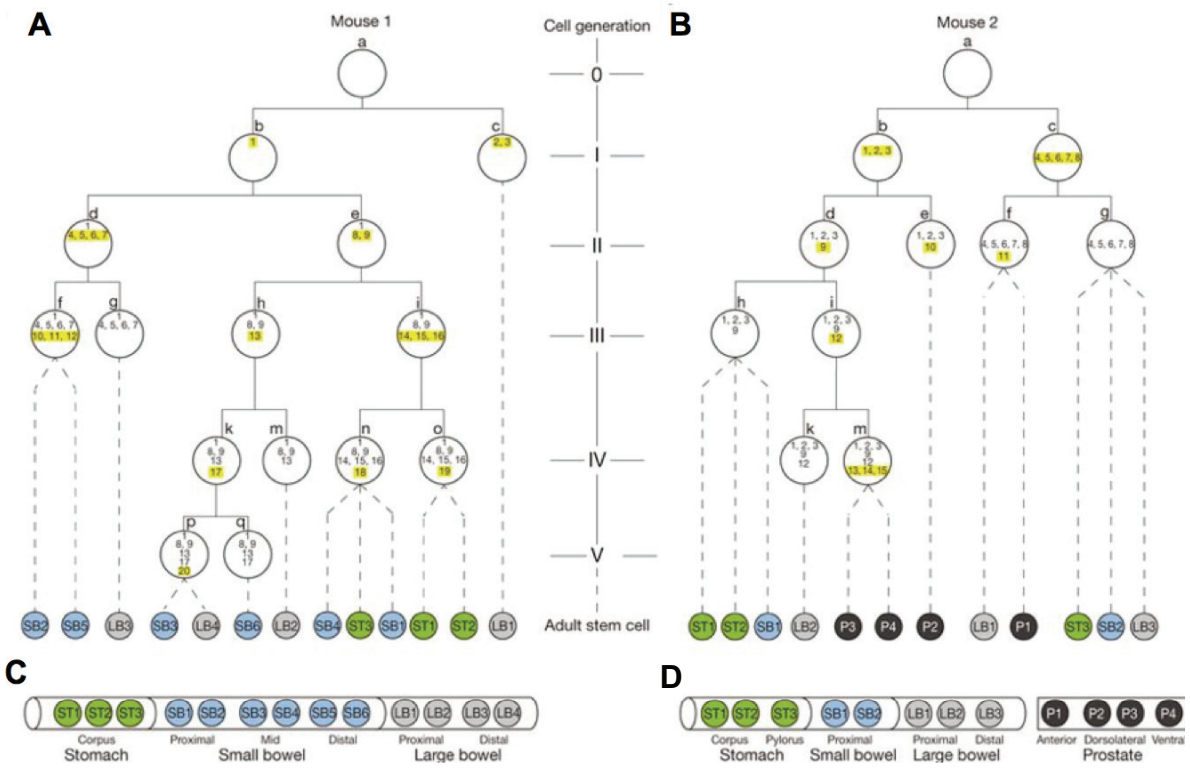


Figure 5 – Phylogenetic tree reconstruction of different tissues in mice
(A, B) Reconstructed phylogenetic trees for the two mice studied (Behjati et al., 2014). The numbers inside each node on the tree represent a unique mutation and by tracing the branches of the tree, the mutations can be seen accumulating in the most recent generations.

(C, D) The coloured circles at the tips of each tree branches indicate the tissue to which the lineage ultimately contributes.

1.6 Phylogenetic reconstruction with subclonal tissues requires a framework to identify cell populations

While single-cell derived clones, such as organoids, enable an easy reconstruction of phylogenetic trees, standard phylogenetic methods are not suitable when sequencing polyclonal populations of cells. In order to reconstruct phylogenetic trees from polyclonal populations, new methods had to be developed that first group mutations in discrete subclones and then build trees of the subclones. For example, Bayesian Dirichlet processes have been used in studies of breast and prostate cancer (Gundem et al., 2015; Nik-Zainal, Van Loo, et al., 2012). The premise of this is that by clustering variants together based on their VAFs, distinct subclones can be defined within the population.

This was demonstrated with 21 breast cancer genomes, whereby whole-genome sequencing, copy number analysis and somatic variant calling with CaVEMan produced a list of variant calls in each sample (Nik-Zainal, Van Loo, et al., 2012). Applying the Bayesian Dirichlet process based on the coverage of the variant read site and the VAF, the clustering of mutations reveal distinct subpopulations. Amongst these subpopulations in each sample, there was a dominant lineage, which accounted for more than half of the sample (Nik-Zainal, Van Loo, et al., 2012). Given the high numbers of shared mutations between clusters, these different populations appear to co-exist for a significant portion of their life history, before diverging into separate subclones. Applying the pigeonhole principle to these subclones enabled the order of mutation acquisition to be inferred and as such, a phylogenetic tree was reconstructed for each sample, the origin of which is the most recent common ancestor (MRCA) of all the identified subpopulations (Figure 6) (Nik-Zainal, Van Loo, et al., 2012). In this

way, somatic mutations, copy number information, mutation phasing and clustering methods (such as the Bayesian Dirichlet process) can be used for phylogenetic reconstruction in a single polyclonal sample.

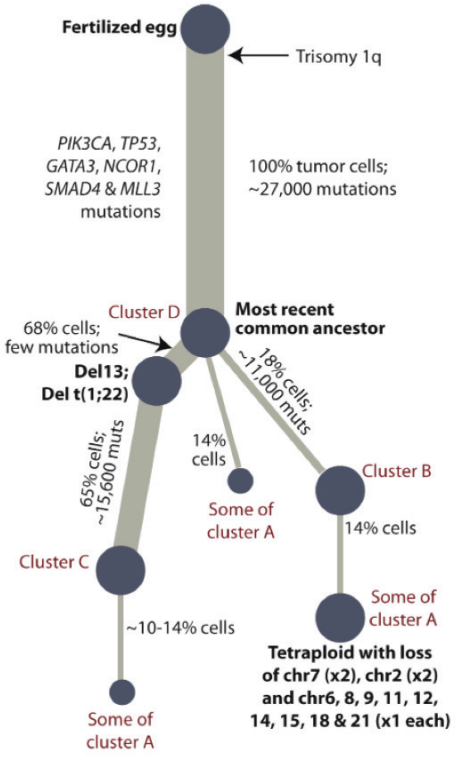


Figure 6 – Phylogenetic reconstruction of a single breast cancer sample
The somatic mutations identified within a tumour population can then be clustered together based on their shared variants with the Bayesian Dirichlet process. Applying the pigeonhole principle can then place these clusters in sequential order, tracing their phylogenetic lineage back to the MRCA (Nik-Zainal, Van Loo, et al., 2012).

While basic phylogenetic trees depicting the relationship between a few subclones can be inferred from a single sample, phylogenetic reconstruction from polyclonal samples is greatly helped by sequencing multiple related samples, such as sequencing multiple regions of a tumour. This is exemplified by a study that performed whole-genome sequencing and somatic variant calling in 51 tumour samples, obtained from ten patients with metastatic prostate cancer (Gundem et al., 2015). An n-dimensional Bayesian Dirichlet process was applied, enabling the identification of clonal and subclonal populations within each sample per patient. By retracing the phylogeny of

multiple tumour samples from the same patient, some including both the primary and secondary tumours, remarkable insights were gained into the metastatic process. Minor subclonal populations appeared to be responsible for the initiation of metastasis and in several cases, multiple subclonal populations from the same tumour appeared to independently achieve metastatic potential (Gundem et al., 2015). Furthermore, not only do metastases appear to *de novo* seed new metastases, but multiple metastases can seed a new metastatic deposit, forming a polyclonal foundation (Gundem et al., 2015). This rapidly diversifies the tumour populations in each secondary tumour and provides a new spatial dimension to the evolutionary history of cancer

Summarising, phylogenetic reconstruction requires accurate somatic variant calling, particularly for the reconstruction of early embryonic lineage trees, in which early branches can be supported by one or a few variants. The Bayesian Dirichlet process provides a framework within which the subclonal populations of a sample, and those between samples, can be identified. The relationship between samples can then be inferred using these clusters while the pigeonhole principle allows the deduction of the sequence in which these populations arose. By applying these principles to normal tissues, novel insights into embryological development, tissue maintenance and carcinogenesis can be sought.

1.7 The pancreatic islets perform endocrine functions

The pancreas is a glandular organ situated in the upper region of the abdomen, with dual exocrine and endocrine functions. The exocrine tissue forms the majority of the parenchyma of this organ and consists of acini and ducts (Figure 7). The acini produce and secrete pancreatic juice, an alkaline solution rich in digestive enzymes, into the branched ductal network which then drains into the duct of Wirsung and into the duodenum via the ampulla of Vater (Horiguchi & Kamisawa, 2010).

In contrast, the islets of Langerhans, or pancreatic islets, are spherical micro-organs distributed throughout the parenchyma of the pancreas that undertake numerous endocrine functions (Figure 7). Accounting for just over 2 cm^3 of tissue in an average adult human, the pancreatic islets are a mosaic of several cell types including α-cells, β-cells, δ-cells, ε-cells and PP-cells (Ionescu-Tirgoviste et al., 2015). The most

common cell type within the islet are β-cells, accounting for 60%, followed by α-cells making up 30% and the remaining 10% being δ-cells, ε-cells and pancreatic polypeptide cells (PP-cells) (Cabrera et al., 2006; Ionescu-Tirgoviste et al., 2015). Fundamental to glucose homeostasis, the β- and α-cells are locked in negative feedback pathways. In the presence of glucose, β-cells produce insulin, a peptide hormone with the primary aim of empowering tissues to utilise the glucose from the bloodstream. In contrast, glucagon from the α-cells acts to increase blood glucose levels from stores in the muscle and liver. Together these opposing functions form a tightly regulated system that promotes euglycaemia. It is the dysregulation of this homeostasis that results in diabetes mellitus (Zheng, Ley, & Hu, 2018).

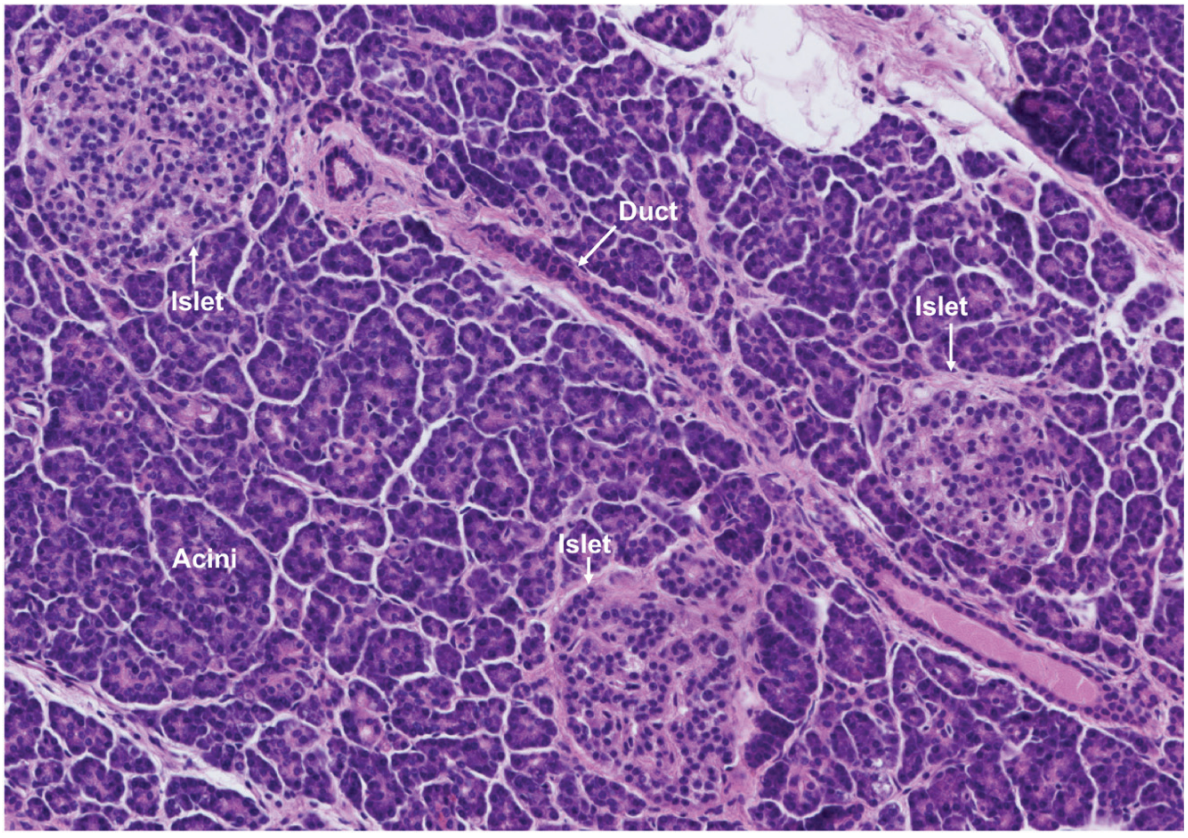


Figure 7 - An overview of pancreatic histology

Pancreatic section (5 μm thickness), taken from patient 290B. Exocrine tissues are the acini and ducts whereas islets are the endocrine component. Whilst exocrine and endocrine tissues are in close proximity, they greatly differ in their functions.

1.8 Primitive islets develop early in foetal development and numbers peak in the post-natal period

The pancreas is first evident as a developing endodermal embryological structure at human gestational day 26 (Piper et al., 2004). By 47 days post-conception, cells expressing the PDX1 transcription factor first appear (Jennings et al., 2013). These progenitors are capable of becoming either a ductal or endocrine cell. The expression of NGN-3 at weeks 7 and 8 diverts these PDX1 positive cells to the endocrine lineage and the first insulin expressing cells are seen around this time (Jennings et al., 2013). This marks a clear difference to the murine model whereby the initial hormone expressed is glucagon (Jennings et al., 2013; Rall, Pictet, Williams, & Rutter, 1973).

Over two more weeks of gestation, the other endocrine cell types emerge, although the β-cells remain the most prevalent endocrine cell during the first trimester (Piper et al., 2004). Key transcription factors that play a role in this continued endocrine development, through weeks 9 to 21, include NKX2.2, NKX6.1, ISLET1, NEUROD1 PAX4 and 6 (Jennings et al., 2013; Lyttle et al., 2008; Sarkar et al., 2008). In contrast, loss of SOX9 expression appears linked with the differentiation of progenitor cells into foetal β-cells (Jennings et al., 2013).

Clusters of endocrine cells first appear around week 12 of gestation. By week 14, these primitive islets develop a vascular network (Jeon, Correa-Medina, Ricordi, Edlund, & Diez, 2009; Piper et al., 2004). These clusters consist initially of more β-cells than α-cells, but this ratio balances out by week 16, remaining at 1:1 until birth (Gregg et al., 2012; Jeon et al., 2009; Riedel et al., 2012). It is the α-cells and δ-cells that show a greater proliferative index compared to the β-cells in the remaining pre-term period (Jeon et al., 2009; Sarkar et al., 2008).

In the neonatal period, β-cell numbers increase compared to the static α-cell population (Gregg et al., 2012). Whilst β-cell neogenesis, where ductal progenitors differentiate into β-cells, is more common in the pre-natal developing pancreas, this does not play a prominent role in the post-natal period. Instead, proliferation of existing β-cells accelerates, reaching a peak of 2% before ceasing by two years old (Gregg et al., 2012). This period also involves many immature islets taking on a more familiar

architecture, seen in the adult pancreas, and by six months all islets have reached this point (Gregg et al., 2012). This coincides with the nutritional shift that occurs at weaning and microRNAs, such as miR-17-5p and miR-29-3b, have been shown to play a key role at this critical time (Jacovetti, Matkovich, Rodriguez-Trejo, Guay, & Regazzi, 2015).

1.9 There are physiological, and pathological, causes for β-cell proliferation

After post-natal proliferation ceases, the proportion of β-cells proliferating, at any one time, drops to approximately 0.5-1% and continues to decrease further with age (Gregg et al., 2012). Even following a loss of endocrine tissue from a partial pancreatectomy, little evidence of β-cell proliferation has been observed (Menge et al., 2008). Human studies using *in vivo* thymidine analogue incorporation combined with radiocarbon, and lipofuscin accumulation, have both supported this, with the suggestion that final β-cell populations are defined before age 30 with little activity afterwards (Cnop et al., 2010; Perl et al., 2010). Only in rare, sporadic cases has β-cell neogenesis been observed in specimens obtained from donors older than five-years-old (Gregg et al., 2012). In light of this, the vast majority of β-cells appear to remain in a quiescent state through life.

The primary pathological cause of β-cell proliferation in adulthood is seen in diabetes mellitus type 2. Recognised by the World Health Organization as an important public health problem, estimates in 2015 put the age-standardised global prevalence of diabetes mellitus, both type 1 (DM1) and type 2 (DM2), at one in eleven adults (World Health Organization, 2016; Zheng et al., 2018). The majority of these are believed to be patients with DM2 (World Health Organization, 2016; Zheng et al., 2018). Primarily a disease driven by insulin resistance in the liver, muscles and islet cells, the dysregulation of glucose homeostasis that occurs in DM2 triggers compensatory β-cell hyperplasia (DeFronzo & Tripathy, 2009; El Ouaamari et al., 2016; Escribano et al., 2009).

While the initial compensatory proliferation and associated increased insulin secretion can help cope with the insulin resistance, the hyperinsulinaemia that results actually drives further insulin resistance and glucose production. This positive feedback

eventually overwhelms the compensatory mechanisms and hyperglycaemia prevails (Zheng et al., 2018). This produces the clinical symptoms often associated with DM2 including polyuria, polydipsia and fatigue. With DM2 established, significant changes then follow in the pancreatic endocrine tissue resulting in a reduced β-cell mass, altered β:α-cell ratios, co-expression of endocrine hormones and loss of β-cell identity (Butler et al., 2003; Enge et al., 2017; Mezza et al., 2014; Spijker et al., 2015). Disease progression often mandates insulin replacement therapy and significant macro- and micro-vascular complications become increasingly more prevalent (Fowler, 2008).

A more physiological cause of β-cell proliferation is seen in pregnancy. The introduction of placental lactogens and growth hormones drives hepatic gluconeogenesis and lipolysis, leading to hyperglycaemia and insulin resistance (Beck & Daughaday, 1967; Rieck & Kaestner, 2010; Sorenson & Brelje, 1997). In response, a 1.4-2.4-fold increase in β-cell mass has been demonstrated (Butler et al., 2010; Van Assche, Aerts, & De Prins, 1978). While in many women this is sufficient and entirely normal, if the insulin resistance is too great and there are other risk factors present, gestational diabetes can arise. The specifics of how the β-cells proliferate remains unclear, with both self-duplication of β-cells and islet cell neogenesis being hypothesised (Butler et al., 2010; Van Assche et al., 1978).

1.10 The maintenance of the pancreatic islets, through adulthood, is unclear

The maintenance of adult pancreatic endocrine tissue has been studied extensively, albeit mainly in model organisms, with numerous hypotheses generated. Mechanisms suggested for islet cell maintenance include self-duplication of existing differentiated β-cells, neogenesis of new islets through transdifferentiation of ductal cells and progenitor/stem cell replenishment.

Originally proposed many decades ago by Messier and Leblond (1960), self-duplication has been best demonstrated using a Cre/lox pulse-chase system in adult mice (Dor, Brown, Martinez, & Melton, 2004). β-cells were labelled and following the chase, the fraction of β-cells per islet was assessed. Over 12 months, self-duplication of pre-existing β-cells should not alter this fraction whereas stem-cell and progenitor renewal would. The results revealed little change in the fraction, but an increase in

endocrine tissue mass. This indicated self-duplication to be the main proliferative pathway. Given that the β-cells were observed to increase in number, this also challenged the notion that islet cells were post-mitotic (Dor et al., 2004). This has been supported by subsequent studies confirming that all β-cells retain the capacity to self-duplicate and that each cell appears to contribute equally to the maintenance of the islet (Brennand, Huangfu, & Melton, 2007).

However, there remain aspects of pancreatic islet maintenance that cast doubt on self-duplication being the only mechanism for islet proliferation. These are mostly focused on the potential that stem cells and progenitors have to differentiate into β-cells. Several different candidates have been suggested to exist, with the locations harbouring these stem cells and progenitors including the pancreatic ductal epithelium and the islet itself (Bonner-Weir, Baxter, Schuppin, & Smith, 1993; Zulewski et al., 2001). One such example supporting a progenitor hypothesis, involved the xenografting of human embryonic pancreases, with PDX1+ and Ngn-3+ progenitors, into immunocompromised mice. Whilst the PDX1+ progenitors differentiated into β-cells, the Ngn-3+ progenitors did not, suggesting differentiated endocrine cells were unable to self-replicate (Castaing, Duvillie, Quemeneur, Basmaciogullari, & Scharfmann, 2005; Castaing et al., 2001).

The definitive existence of pancreatic islet stem cells is proving difficult to confirm, with recent forays into the single-cell transcriptomics of pancreatic islets, failing to identify a single stem cell lineage (Muraro et al., 2016). This does not completely rule out the stem cell theory, as there may exist multiple, different stem cell populations that contribute to islet maintenance. However, given that these stem cells appear to be extremely rare within the islet, isolating even one population with single cell transcriptomics will require far larger data sets (Andrews & Hemberg, 2018).

Transdifferentiation of non-endocrine cells, such as the pancreatic ducts and acini, into endocrine cells, has also been suggested, particularly under injury. Given the translational potential of islet neogenesis for the treatment of diabetes mellitus, this hypothesis has garnered much attention. Researchers have transformed *in vitro* acini, islet cell precursors and even splenocytes into functioning β-cells (Guz, Nasir, &

Teitelman, 2001; Kodama, Kuhtreiber, Fujimura, Dale, & Faustman, 2003; Lipsett & Finegood, 2002; Socorro et al., 2017). Recent work has revealed peripheral regions of islets harbour immature β-cells that appear to be descended from nearby α-cells, the implication being that transdifferentiation may occur within the pool of different endocrine cells themselves (Chakravarthy et al., 2017; van der Meulen et al., 2017).

Finally, whether an entire islet unit can duplicate itself, in a "fission" event is debated. Fission has been well-proven in the colonic crypts, both in the post-natal period and in adulthood, as has crypt fusion (Bjerknes, 1986; Bruens, Ellenbroek, van Rheenen, & Snippert, 2017; Cheng & Bjerknes, 1985; Clarke, 1972). In the pancreatic islets, fission has been postulated using X-inactivation mosaic mice with lacZ insertion, on the X-chromosome (Seymour, Bennett, & Slack, 2004). Islets were identified that appeared to have an irregular morphology, whereby two small masses of endocrine cells appear to be linked by an isthmus of α-cells. These were named "dumb-bell" islets (Seymour et al., 2004). By comparing the X-inactivation status, and hence lacZ expression, of the masses on either side of the isthmus, it was deduced that the masses on either side of the isthmus were more related to each other than two randomly selected nearby islets were (Seymour et al., 2004). Further, comparing distinct islets to each other revealed this same measure of similarity decreased as the distance increased between them. The conclusions drawn were that these dumb-bell islets were in a state of fission, rather than fusion (Seymour et al., 2004).

1.11 Summary

Identifying somatic mutations has proved successful in cancer and the stage is set for studying normal tissue. The pancreatic islets represent a high-priority normal tissue to investigate, given the scale of the health burden that DM2 poses. While efforts have been made with single-cell RNA sequencing to decipher the somatic mutational landscape of the pancreatic islets, these methods continue to be burdened by a high false discovery rate (Enge et al., 2017). By establishing a workflow using whole-genome sequencing and laser capture microdissection (LCM), the somatic mutational profile of the islets can be examined and key questions regarding the development and maintenance of the pancreatic endocrine tissue can hopefully be answered, opening up the possibility of translational benefits in pancreatic islet disease.

2. Aims

2.1 Develop a robust workflow for analysing somatic mutations in normal tissue

Studying somatic mutations in normal tissue is still a new field. Previous studies have made use of targeted sequencing of a collection of known cancer drivers (Martincorena et al., 2015). For whole genome sequence data, a matched analysis has typically been favoured given the efficient removal of germline mutations that can be achieved. However, to capture the early embryonic mutations, an unmatched approach is necessary (Behjati et al., 2014). This comes at the cost of calling both germline and somatic mutations. Therefore, my first aim was to design a bioinformatics workflow that can confidently exclude germline variants from true somatic variants, with high sensitivity and specificity.

2.2 Characterise the landscape of somatic mutations in the normal pancreatic islets

Somatic mutations in healthy pancreatic islets have so far been investigated only with single-cell RNA sequencing (Enge et al., 2017). As such, this is limited to the exome and is affected by a high rate of errors introduced by the whole-genome amplification stage. This limits the insight into the mutational processes acting on them, particularly early in development. Using whole genome sequence data, I intend to obtain estimates for the mutational burden and use these to deduce what mutational processes the islets are subjected to. This is also of relevance to better understand the mutational processes that may be active in the normal cells that give rise to pancreatic neuroendocrine tumours.

2.3 Elucidate the early phylogeny of the pancreatic islets

By using somatic mutations, I hope to obtain new insights into the development of the pancreatic islets. The first question would be to determine clonality, confirming whether all cells in an islet derive from a single founder cell or lineage, or whether different lineages contribute to an islet. If islets are monoclonal or at least oligoclonal, dominated by one or a few major lineages, it might be possible to reconstruct an embryonic lineage tree (Figure 8). Integrating this with the spatial distribution of the islets would then provide a glimpse into the anatomical shaping of the pancreatic endocrine tissue, during early embryogenesis.

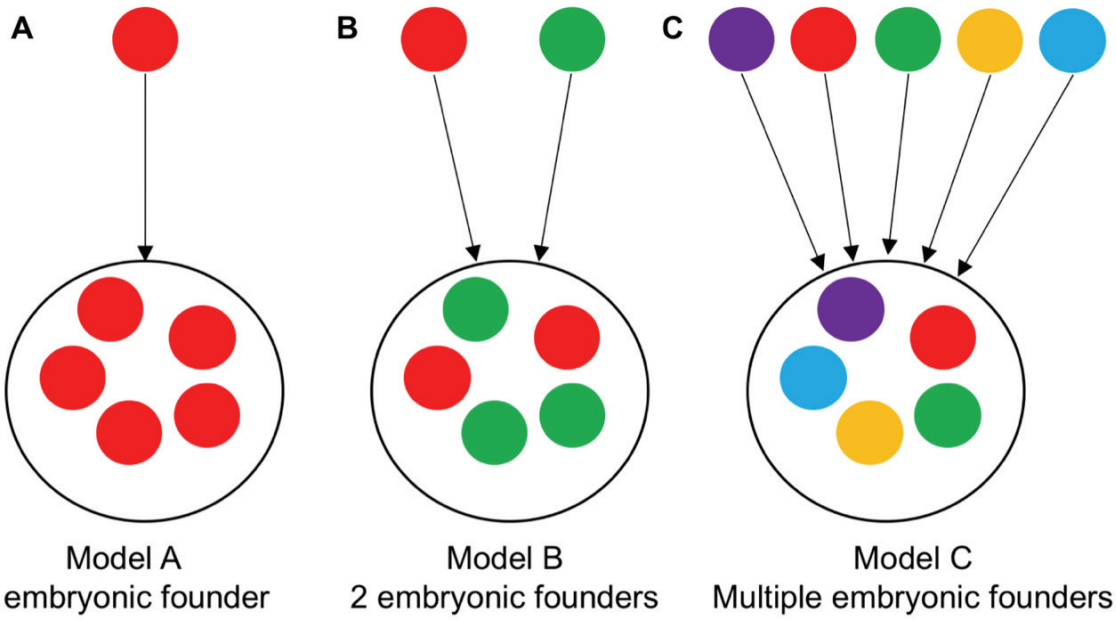


Figure 8 - Somatic mutations can be used to separate the embryonic lineages that contribute to the formation of the islets

(A) Model A shows a single cell founding an islet. The islet cells then go on to carry the same mutations as the founding cell. The VAF distribution of the entire islet would be centred on 0.5, in addition to the low VAF private mutations that have accumulated through life, in the individual islet cells. The MRCA of all the cells in the islet is the founding cell. Under this model, embryonic lineage trees can be obtained used standard phylogenetic methods such as maximum parsimony.

(B) Model B shows two different cells founding an islet. Each cell would carry ancestral mutations from either the red or green lineage plus their own private mutations. Mutations in each of the two founding lineages would have VAFs less than 0.5, but the sum should approach 0.5. The MRCA of the cells in the islet is the MRCA of the two founding cells. If each islet present has one or a few dominant lineages, embryonic lineage trees could be obtained using subclonal decomposition.

(C) Model C shows multiple cells founding an islet. Thus, the islet cells will share fewer ancestral mutations and have many more private mutations accounting for a smaller VAF each. If the contribution of different lineages does not vary across the islets, reconstructing the embryonic lineages under this scenario will not be possible.

2.4 Gain insight into the maintenance of the pancreatic islets through life

Some limited insights into the maintenance of the islet population throughout life might be obtained from the VAF distribution. If a large fraction of the cells of an islet was replenished by one, or a few, stem cells during adult life, this would be expected to manifest as peaks in the VAF distribution of the somatic mutations detected in the islet. A single stem cell could even take over the islet, much like those in the colonic crypts, and the resulting VAF distribution would show a single large peak at a high VAF. A similar distribution could also be obtained if islets are founded later in life by a single founding cell.

In contrast, if islets are maintained by the self-duplication of differentiated islet cells, or by a large number of slow cycling cells, or even if most islet cells are not replaced throughout life, clonal expansions within an islet would not be expected to reach detectable VAFs in adulthood. With this in mind, the VAF distributions may help differentiate between extreme models of tissue maintenance.

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).

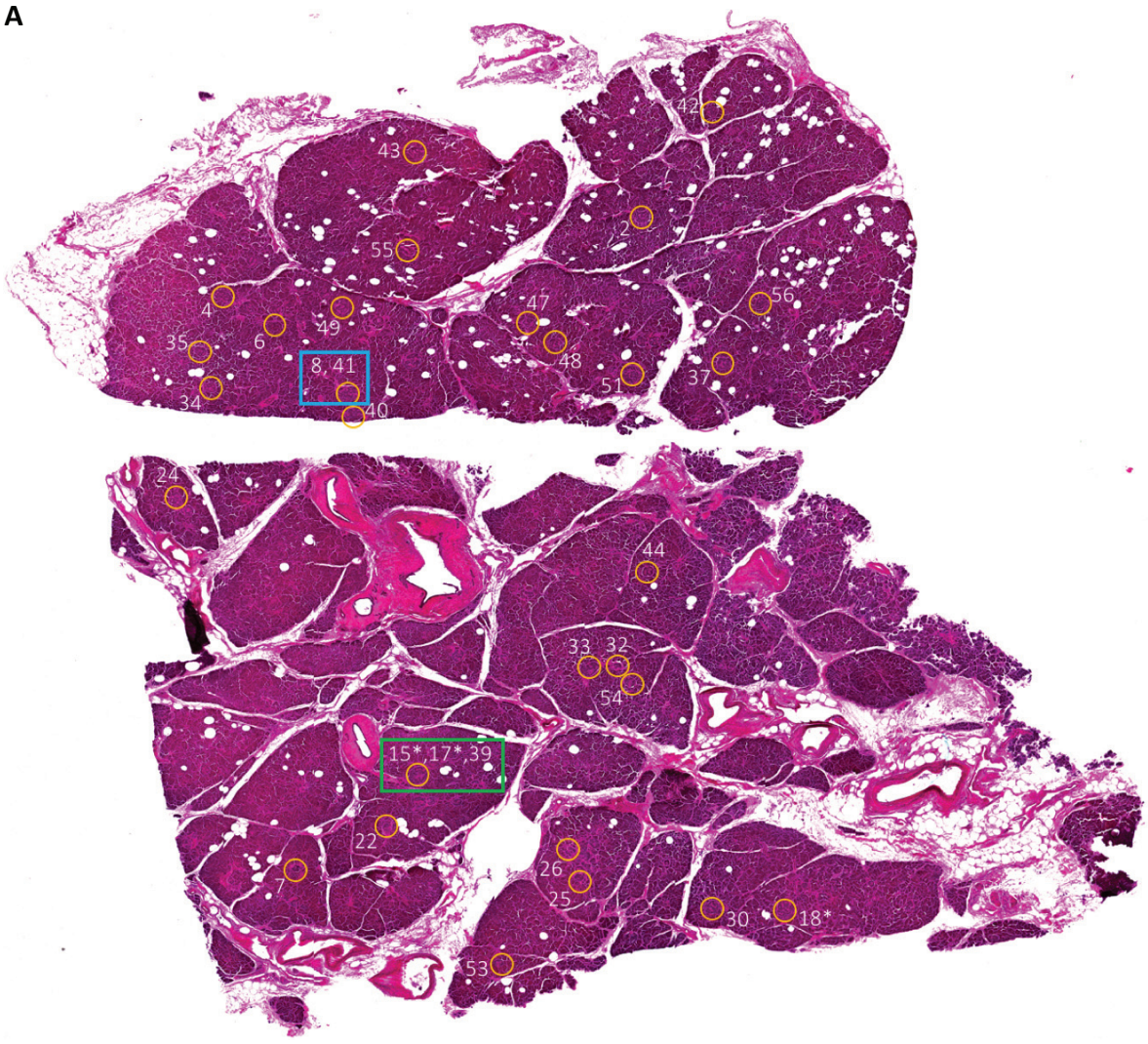
The visualisation of the individual phylogenetic trees for each sample was performed using the R package ggtree, (v1.12.0, https://github.com/GuangchuangYu/ggtree) (Yu, Smith David, Zhu, Guan, & Lam Tommy, 2016). This was done in collaboration with Tim Coorens. By overlaying the phylogenetic lineages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

4. Results

4.1 Whole-genome sequencing of 32 pancreatic islets from the same individual

Through LCM, 40 islets were obtained from a single pancreatic biopsy of patient 290B. From these 40 islets, 32 samples were sent on for WGS. The eight samples that did not go on to be sequenced all either had too little DNA and therefore failed library preparation, or they did not pass inspection by a trained clinical histopathologist. The reasons for failing histological review included contamination from nearby tissues, such as pancreatic acini, and incorrect identification of an islet. Included in the 32 samples sequenced was a biological duplicate and triplicate. Therefore, 29 unique islets in total were sequenced. An overview of the spatial location of the 29 unique islets sequenced is shown in Figure 9.

A



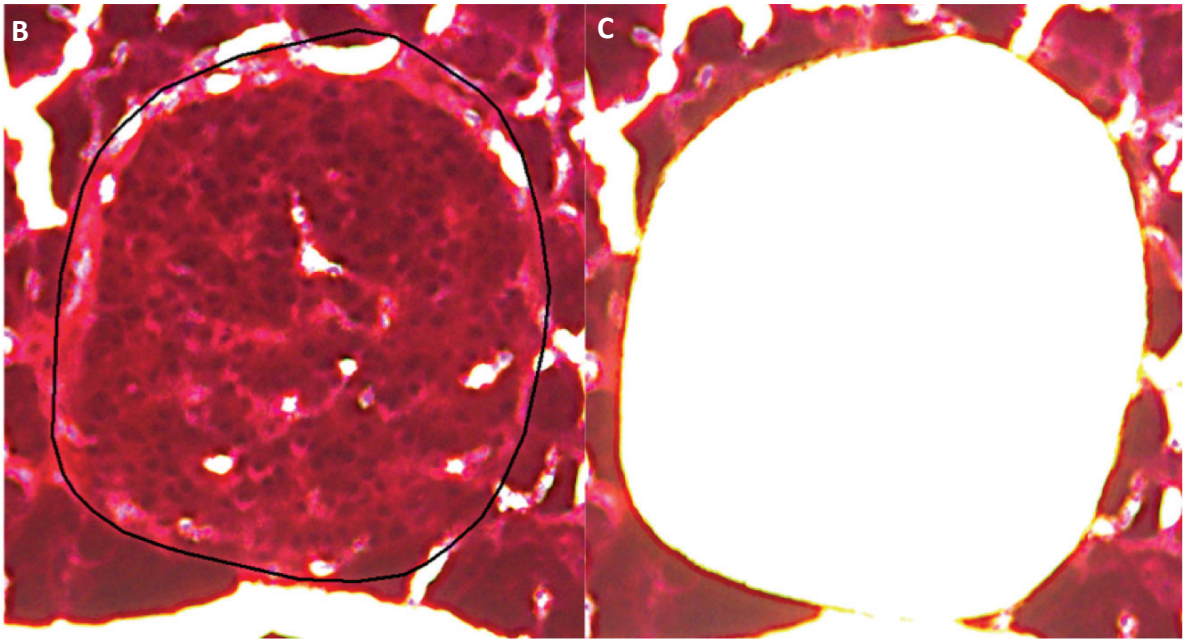


Figure 9 - Images from the LCM of the pancreatic islets

(A) An overview image of a single section of pancreas obtained during LCM with a 10X objective lens. The 29 unique islets are marked by orange circles. The number labelling these circles is the suffix of each sample ("PD37726d_lo00"). Islets labelled with a * are obtained from a z-slice 16 μm above or below this slice. The duplicate samples (blue box) are labelled as 8 and 41, while the triplicate samples (green box) are 15, 17 and 39.

(B), (C) A close-up of an islet excised during LCM, before and after dissection. The sample is PD37726d_lo0018 with an area of 38,659 μm^2.

The mean area per microdissection was 17,441 μm^2 while the mean number of z-slices was three. The total area excised per well was positively correlated with the DNA concentration obtained (Figure 10A). The mean DNA concentration per sample was 62 ng/μL, and coverage improved with increasing concentrations (Figure 10B). It appears enough DNA was obtained from these samples to provide a high library complexity that was not exhausted by the level of coverage achieved here.

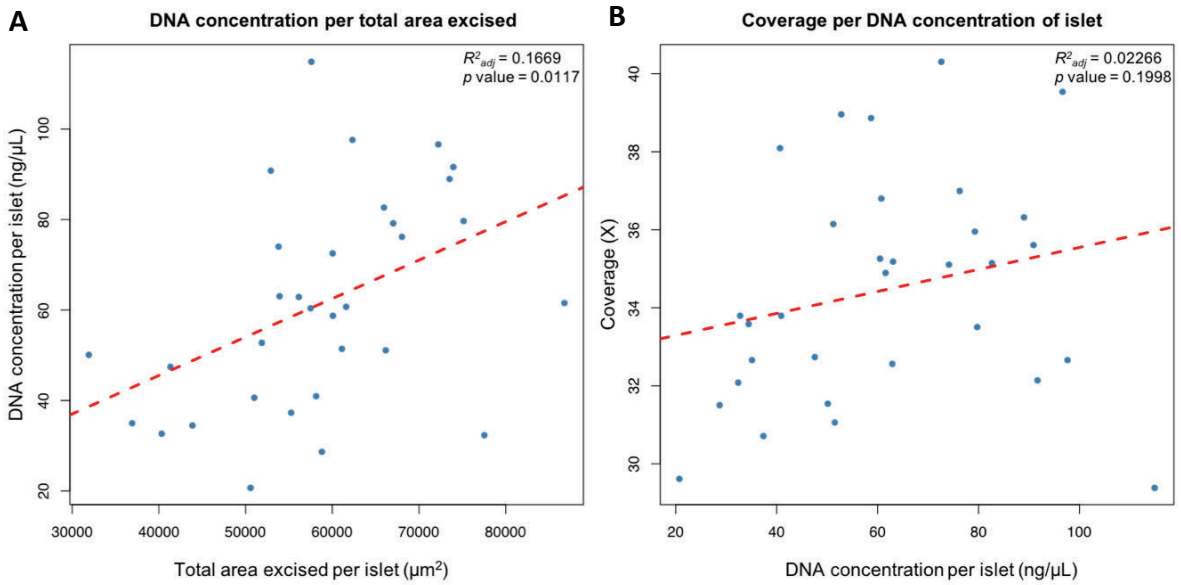


Figure 10 – The data metrics from the LCM workflow

The dashed red line in each graph indicates the linear regression with the adjusted R^2 and *p* value in the top right.

(A) A scatterplot showing a statistically significant increase in DNA concentration, as the total area excised per islet increases.

(B) A scatterplot showing the coverage achieved from the DNA concentration per islet. The correlation is not statistically significant, as shown by the *p* value exceeding 0.05.

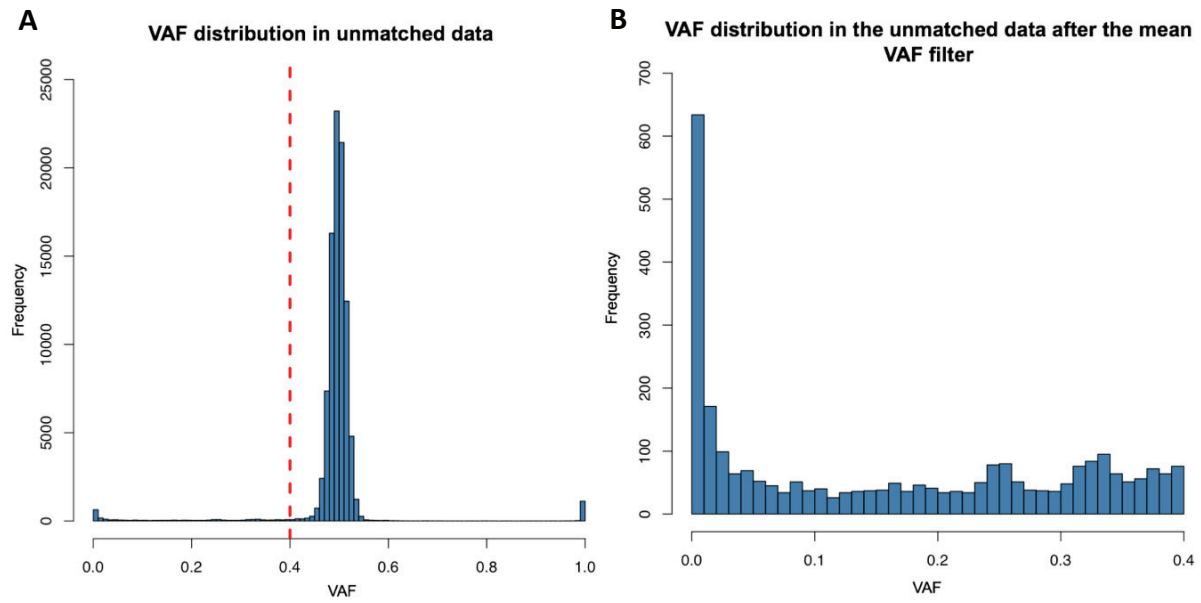
4.2 Successful identification of somatic mutations in individual islets

Traditionally, variant calling in cancer genomic studies relies on comparing a tumour sample to a matched normal to identify mutations exclusively present in the tumour sample, while removing germline mutations shared between both samples. Early embryonic mutations provide an additional challenge using these traditional methods as they are present in both the sample of interest and the normal matched sample.

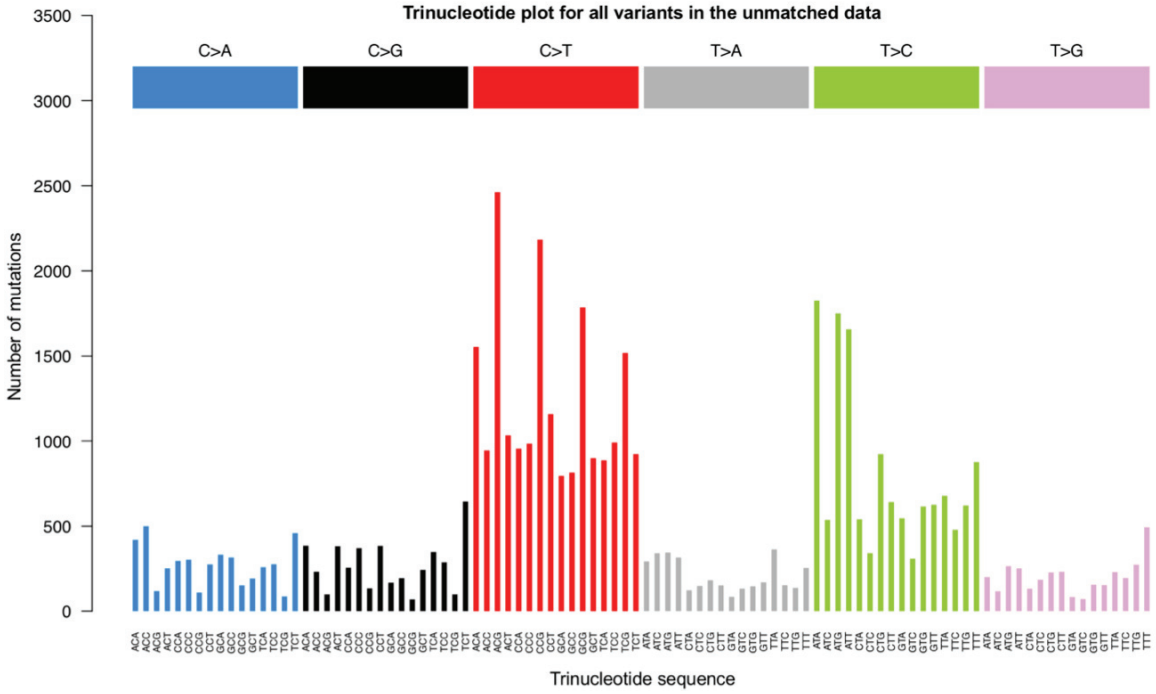
To approach this task, CaVEMan was run in two different ways: a standard run using a matched bladder urothelium (sample PD37726b_lo0071), and an unmatched run, using an unrelated WGS sample as a reference (section 3.7). The latter analysis results in the identification of both somatic and germline mutations, but also allows the

retention of those early embryonic mutations that are critical to phylogenetic reconstruction. With appropriate filtering of the unmatched data, and comparing the calls to the matched analysis for validation, it is hoped the germline mutations can be removed while still retaining the early embryonic mutations.

Sequencing artefacts introduced during the LCM pipeline were then removed using filters designed by Mathijs Sanders. This was followed by *in silico* genotyping with ten matched bladder urothelium samples. Copy number analysis was then performed and showed no significant gains or losses, with a mean ploidy of 1.97 (Appendix 8.2). In the unmatched run, the total number of variants identified was 1,978,687, with 95,317 being unique. The first step in removing the germline variants was to remove any calls with a mean VAF, across all samples, greater than 0.4. This left 79,465 variants, of which 2,799 were unique. The effect of this mean VAF filter can be seen in Figure 11.



C


Trinucleotide plot for all variants in the unmatched data
C>A
C>G
C>T
T>A
T>C
T>G
Number of mutations
0
500
1000
1500
2000
2500
3000
3500
Trinucleotide sequence

D

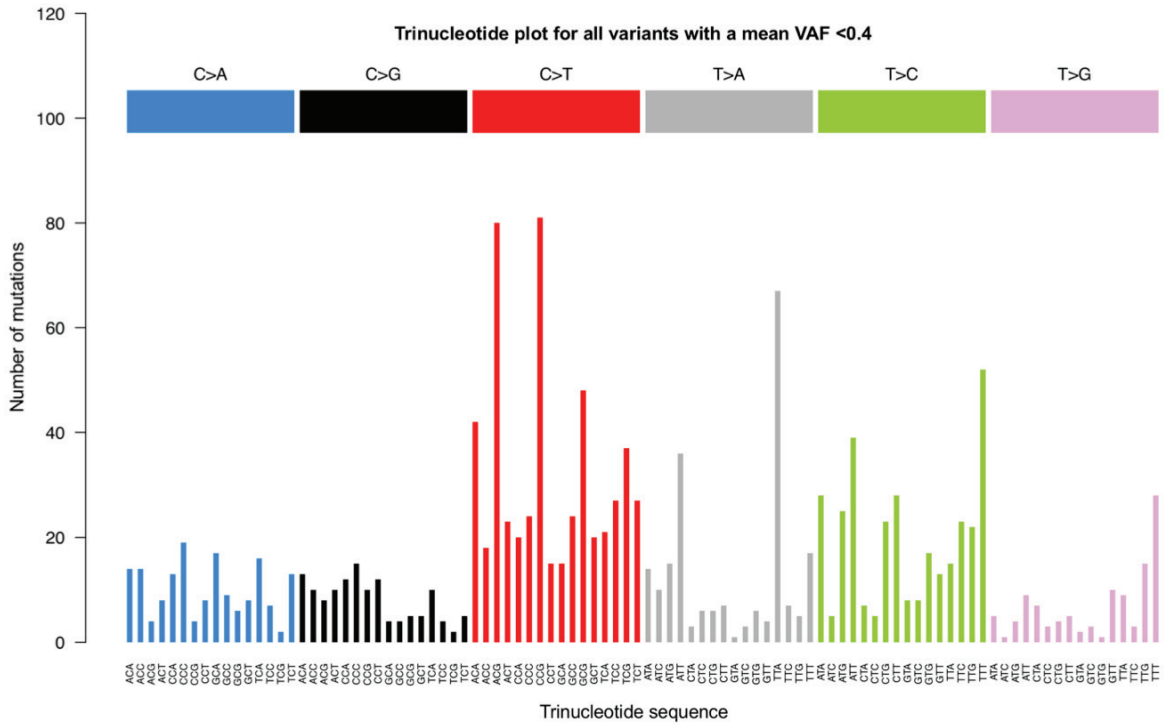

Trinucleotide plot for all variants with a mean VAF <0.4
C>A
C>G
C>T
T>A
T>C
T>G
Number of mutations
0
20
40
60
80
100
120
Trinucleotide sequence

Figure 11 – The initial variant filtering in the unmatched analysis significantly reduced the number of variants

(A) The VAF distribution prior to any filters being applied. The large number of mutations in the histogram are in a binomial distribution with a mean of 0.5. The red vertical line signifies the mean VAF cut-off of 0.4.

(B) The VAF distribution following removal of the variants with a mean VAF greater than or equal to 0.4.

(C) The 96-trinucleotide bar plot for all 95,317 unique variants, prior to any filters being applied. There is an excess of C>T mutations.

(D) The 96-trinucleotide bar plot following the application of the mean VAF filter. While C>T mutations still dominate, TpTpA and ApTpT mutations are significantly more prominent than before.

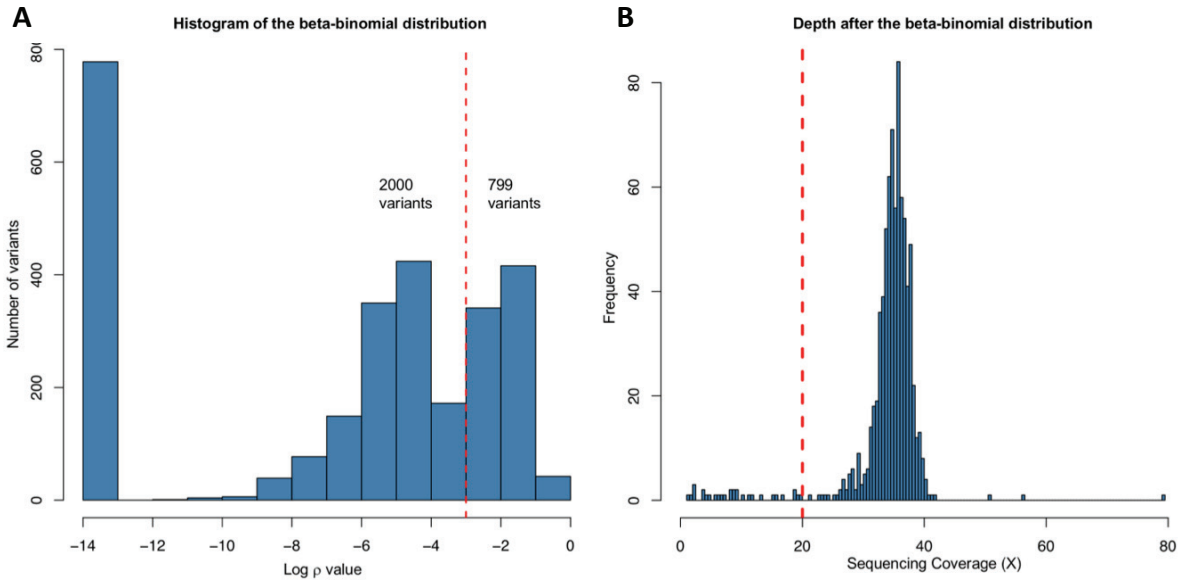
For consistency, the matched run was processed analogously and identified 1,318 variants initially, before this reduced to 1,284 after the mean VAF filter.

4.3 Use of the beta-binomial distribution to identify variable sites

Although the removal of variants, with mean VAF across samples greater than 0.4, is expected to remove the vast majority of heterozygous and homozygous variants, some can remain at lower frequencies, either by chance or owing to systematic mapping biases. Distinguishing those genuine somatic mutations relies on the hypothesis that their VAFs would vary considerably between samples, from the same individual, depending on the relative contribution of different lineages to different samples. This is helped by the availability of ten matched bladder urothelium samples which have previously been shown to be dominated by individual clones. In contrast, a germline variant or low-frequency artefact, would be expected to be more evenly distributed across libraries. In this way, somatic mutations would show a greater level of dispersion amongst sample, compared to germline mutations and artefacts.

To quantify the extent of the variation per variant across samples, while removing the stochastic noise from binomial sampling, a beta-binomial distribution was fitted to the

mutant counts from each sample (section 3.10). The over-dispersion parameter (ρ), representing this variation, took on a bimodal distribution across the mutations, separating those that genuinely vary across biopsies, from those that show an approximately constant error rate. Upon manual inspection of this distribution, the 799 unique variants with a log ρ greater than -3 were retained, and the remaining 2,000 were discarded (Figure 12A). A depth filter was then applied to both data sets specifying that all variants have a mean coverage greater than 20X (Figure 12B).



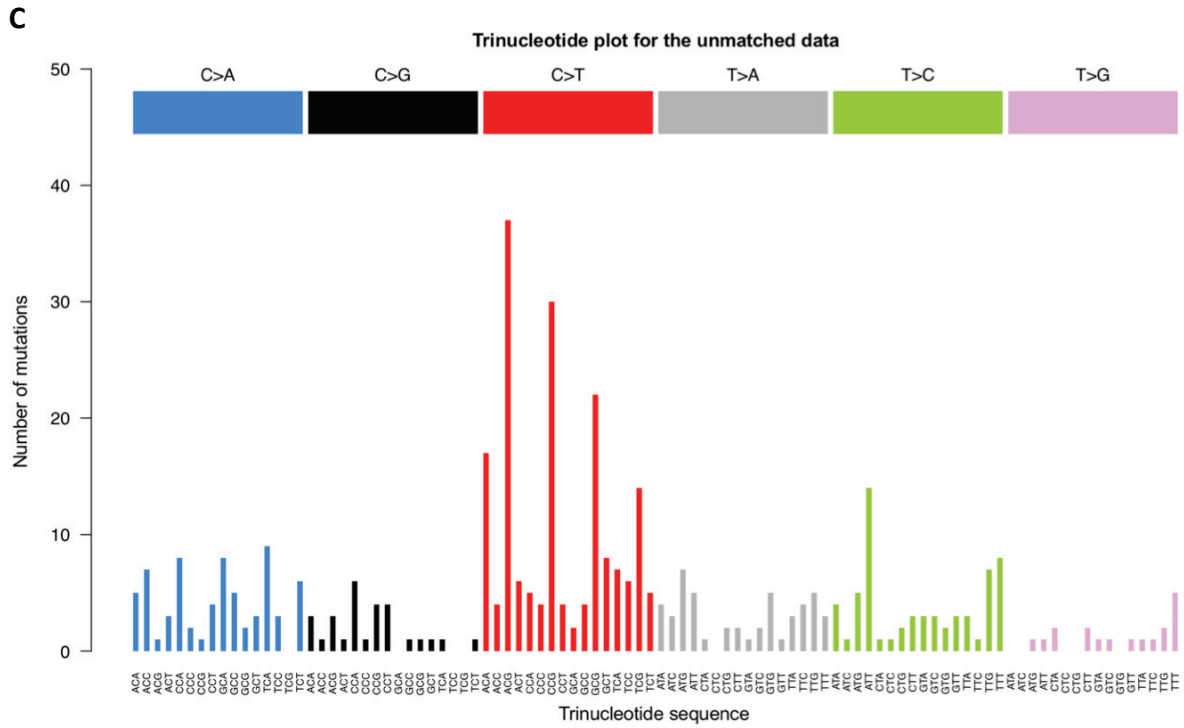


Figure 12 – The beta-binomial distribution identified over-dispersed variants
(A) The beta-binomial distribution applied to the 2,799 variants that passed all previous filters in the unmatched data. The vertical red line represents log ρ of -3, and 799 variants are shown exceeding the over-dispersion parameter.

(B) The mean coverage in the unmatched data following the beta-binomial filter. A depth filter was applied to the 799 variants that passed the beta-binomial filter, to retain those with a coverage >20X. This cut-off is shown by the dashed red line and excluded 30 variants.

(C) The 96-trinucleotide bar plot for the 769 variants remaining in the unmatched data, following the beta-binomial distribution and depth filter. There is still a clear prevalence of C>T mutations, although the T>A mutations have decreased.

Applying these two filters further reduced the number of unique variants in the unmatched data to 769. This appeared to remove a significant number of the T>A mutations that had been present. These transversions have been linked to an artefact generated by the fragmentase enzyme mix, during whole-genome library preparation (New England Biolabs, Ipswich, MA, USA). The filtered-out variants are further analysed in the Appendix 8.3. Similarly, this approach reduced the variant count in the matched data to 737 unique mutations. A summary of these filtering steps is shown in Figure 13.

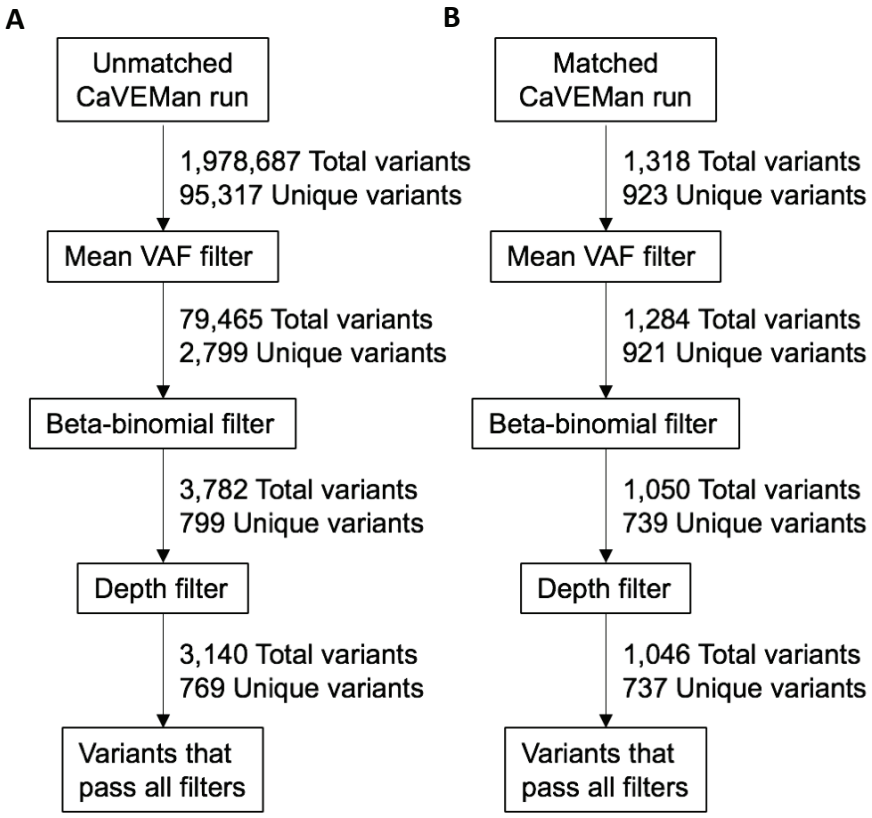


Figure 13 – The two workflows for the matched and unmatched data

(A) The unmatched workflow initially shows a larger number of variants due to the lack of germline filtering during the CaVEMan run. However, this is extensively reduced by the filters in the workflow.

(B) The matched workflow used a matched normal sample during CaVEMan. This was the bladder urothelium sample PD37726b_lo0071 and ensured removal of germline variants early on in the workflow.

4.4 The unmatched analysis provided comparable results to the matched analysis

With a final list of variants in each data set, comparing the two sets revealed a high degree of concordance (Figure 14). The 769 variants in the unmatched data includes all 737 variants that are present in the matched data, but also an extra 32 exclusive mutations.

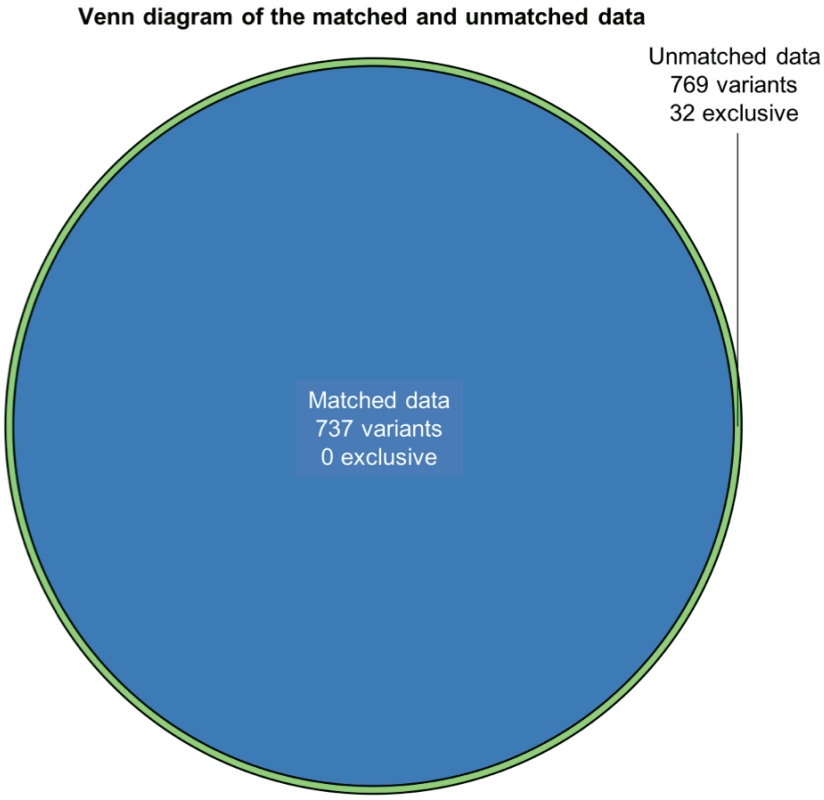


Figure 14 - A Venn diagram demonstrating the overlap between the matched and unmatched data

There are 737 shared mutations (blue circle) between the two data sets, and all of these are nested within the 769 variants in the unmatched data (green circle). The extra 32 mutations exclusive to the unmatched data set are highlighted by the green rim produced from the overlapping circles. Generated using the R package VennDiagram, v1.6.20, https://CRAN.R-project.org/package=VennDiagram (Chen & Boutros, 2011).

These 32 exclusive mutations were then manually checked using the genome browser, JBrowse (v2.2.0, https://github.com/GMOD/jbrowse) (Buels et al., 2016). Two variants were removed from the data due to poor read quality (Figure 15), leaving 767 variants, with 30 of these being exclusive to the unmatched data. All 30 variants were present in both the islets and the bladder samples, suggesting that these may either precede or occur in the MRCA of both tissues.

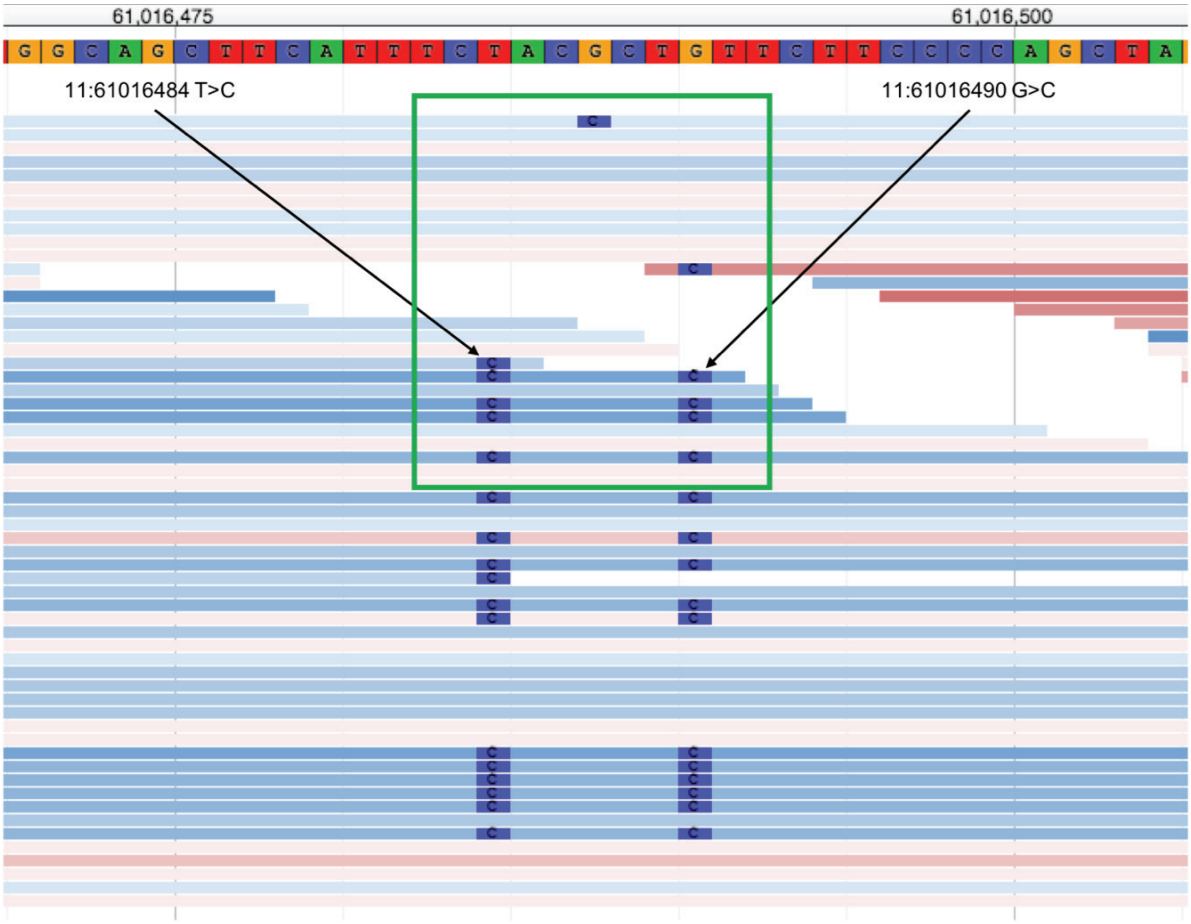


Figure 15 – Manual inspection of the two variants that were excluded from the unmatched data

JBrowse screenshot of sample PD37726d_lo0008 with the two variants highlighted by the arrows and labels (Buels et al., 2016). The reference sequence is present at the top of the image. Each horizontal bar is an individual sequencing read with red representing a forward strand and blue being a reverse. The read quality is represented by the intensity of the colour in each read. Darker intensities signify a poor read quality, compared to lighter shades.

The two variants appear to be present almost exclusively in poor quality reads. This stark contrast is shown best the green box. Additionally, almost all reads with the variants have been sequenced in the same direction.

The conclusions from comparing the unmatched and matched results are that the unmatched analysis not only identifies the same mutations as the matched data, but it also rescues key mutations that were removed by the germline filter. These appear

mostly to be genuine somatic mutations, with two potential artefacts recovered. Overall, this new filtering approach for unmatched variant calling appears to offer a powerful way of identifying somatic mutations and early embryonic mutations without a considerable loss in specificity, compared to traditional matched normal analyses. Therefore, the unmatched data alone will be used for all further analyses.

4.5 Early embryonic mutations are identified by unmatched variant calling

The 30 mutations found exclusively in the unmatched data included some at a particularly high mean VAF across all samples. This is consistent with these variants being early embryonic mutations that are mosaic in multiple biopsies, and common to both the pancreatic islet and bladder urothelium whole genomes (Figure 16).

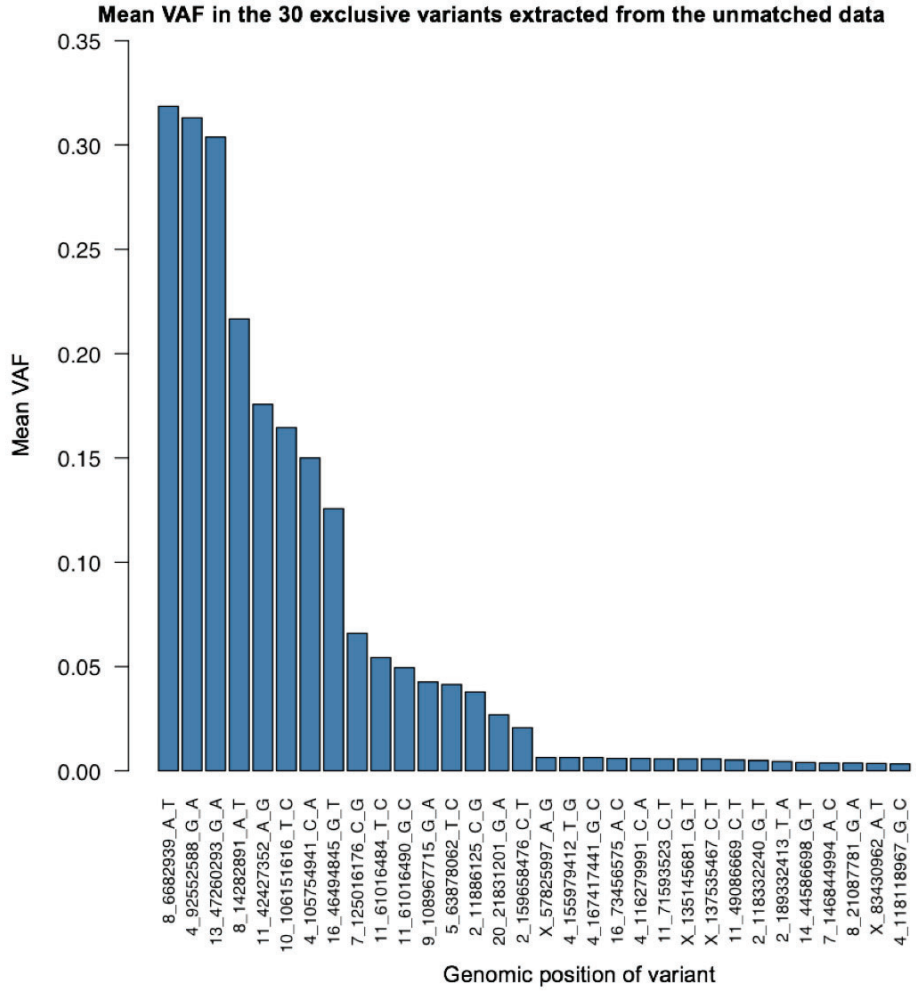


Figure 16 – The 30 exclusive variants showed a range of mean VAFs

A bar plot of the 30 variants found exclusively in the unmatched data, after all filters and manual inspection. The range of means is from 0.033 to 0.319, while the mean VAF across all variants is 0.070.

Within these 30 variants recovered through the unmatched analysis, eight had mean VAFs across all samples that exceeded 10%. This is consistent with what would be expected from those very early embryonic mutations. Three variants had global mean VAFs greater than 25%, accounting for over 50% of the cells in all islet and bladder samples. These may have occurred in the first cell division of the MRCA of the pancreatic endocrine tissue and the bladder urothelium. These include an A>T transversion at 8:6682939, a G>A transition at 4:92552588 and another G>A transition at 13:47260293 (Figure 16). The unmatched data therefore appears to have recovered mutations that could have occurred in the early developmental stages.

4.6 Almost all mutations identified had no apparent functional impact

From the list of 767 mutations in the unmatched data, the vast majority occurred in intergenic and intronic regions (Figure 17).

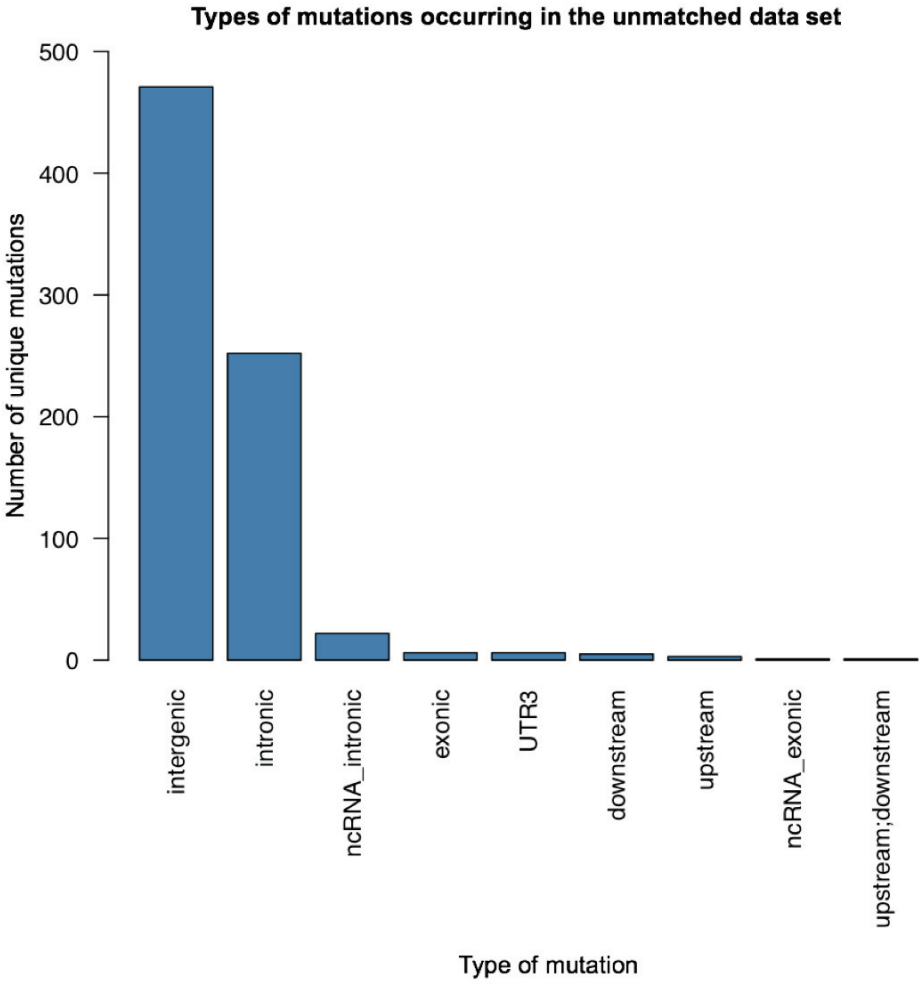


Figure 17 - The frequency of mutations in different regions of the genome
The vast majority of variants are found in intergenic and intronic locations, with very few in the exonic regions. ncRNA is non-coding RNA and UTR3 is the untranslated region in the 3' end of the gene.

Six non-synonymous mutations were identified among the 767 mutations (Table 3). Three of these had PolyPhen2 HDIV scores exceeding 0.85 and were classed as "deleterious" (Adzhubei et al., 2010).

Table 3 – Six non-synonymous mutations were identified
The non-synonymous mutations are shown with PolyPhen2 HDIV scores and VAFs. *In silico* re-genotyping with CGPVAF ensured that variants that would otherwise have been missed by CaVEMan, due to too few variant reads, were called.

Gene	Mutation	Amino acid change	PolyPhen2 HDIV Score	Islets with mutation	VAF
XRN1	3:142144304 C>A	p.Trp161Cys	1	PD37726d_lo0004	0.12
				PD37726d_lo0018	0.05
				PD37726d_lo0047	0.06
				PD37726d_lo0048	0.07
ATRIP	3:48501909 G>A	p.Val393Met	1	PD37726d_lo0007	0.12
				PD37726d_lo0037	0.04
LIPI	21:15561402 T>A	p.Thr150Ser	0.985	PD37726d_lo0002	0.03
				PD37726d_lo0008	0.02
				PD37726d_lo0035	0.03
				PD37726d_lo0055	0.03
				PD37726d_lo0056	0.45
NEK10	3:27333020 A>T	p.Asp477Glu	0.349	PD37726d_lo0006	0.14
				PD37726d_lo0022	0.02
OR2T12	1:248457927 C>A	p.Arg318Ser	0	PD37726d_lo0004	0.15
				PD37726d_lo0007	0.03
				PD37726d_lo0048	0.03
OR8H3	11:55890132 C>T	p.Thr95Met	0	PD37726d_lo0037	0.15

The three deleterious non-synonymous mutations occurred in the exonic regions of the genes *XRN1*, *ATRIP* and *LIPI*. *XRN1* is an exoribonuclease involved in the degradation of RNA transcripts carrying nonsense mutations (Gatfield & Izaurralde, 2004), while *ATRIP* plays a key role in the repair of single-strand DNA breaks alongside the ataxic telangiectasia and Rad3-related protein (Zou & Elledge, 2003).

LIPI codes for a lipase I, an enzyme involved in the metabolism of lipids has been associated with hypertriglyceridaemia (Wen et al., 2003). However, a variant affecting codon 150, as found here, has not previously been identified in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database (https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=LIPI) (Forbes et al., 2017). The variant detected here was common to five islets with four of them having a VAF less than 0.04. Given that PD37726d_lo0008 and PD37726d_lo0041 are duplicates, a variant present in one would be expected to be present in the other. The VAF being below the limit of detection makes it likely this is simply a missed variant in sample PD37726d_lo0041.

In sample PD37726d_lo0056, this variant carries a much higher VAF (0.46). Manual inspection of the reads using JBrowse supported this variant being a true somatic mutation (Figure 18) (Buels et al., 2016). Given that this passed all the filters and manual inspection, as well as being within the limits of detection, this would make an interesting candidate gene to investigate further for any phenotypic effects and possible selective pressures.

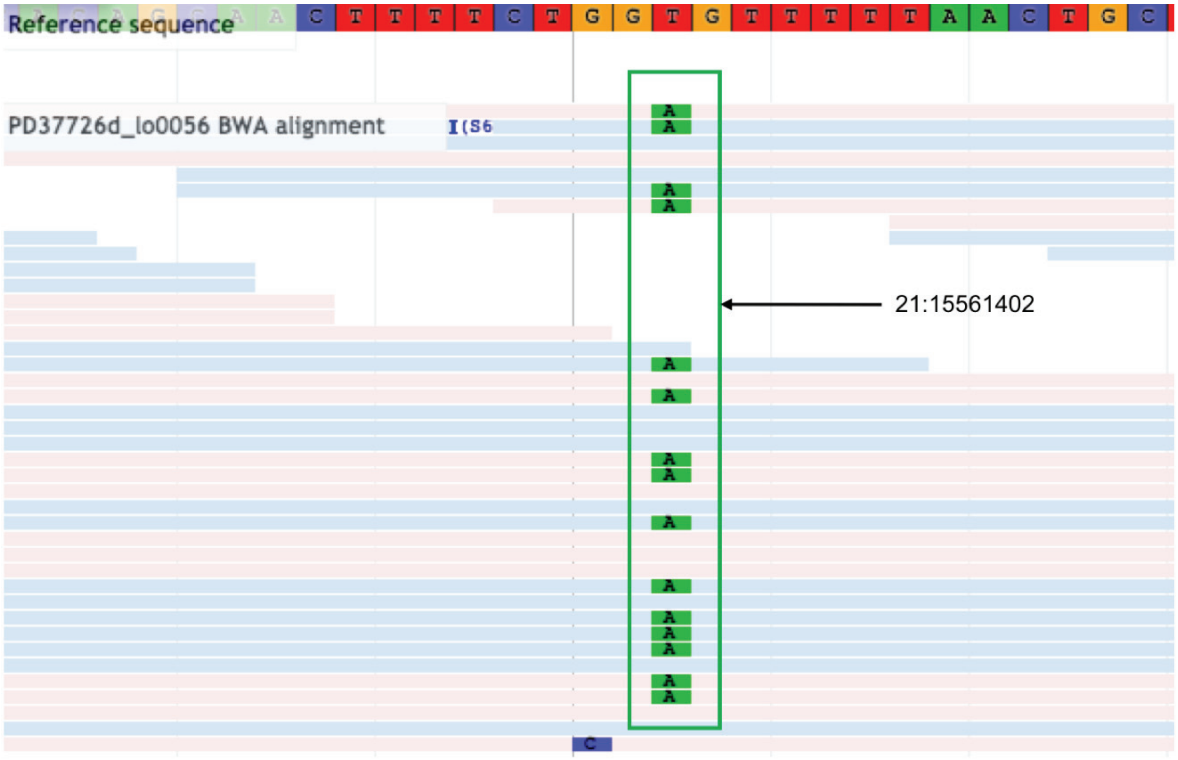


Figure 18 – The *LIPI* variant in sample PD37726d_lo0056

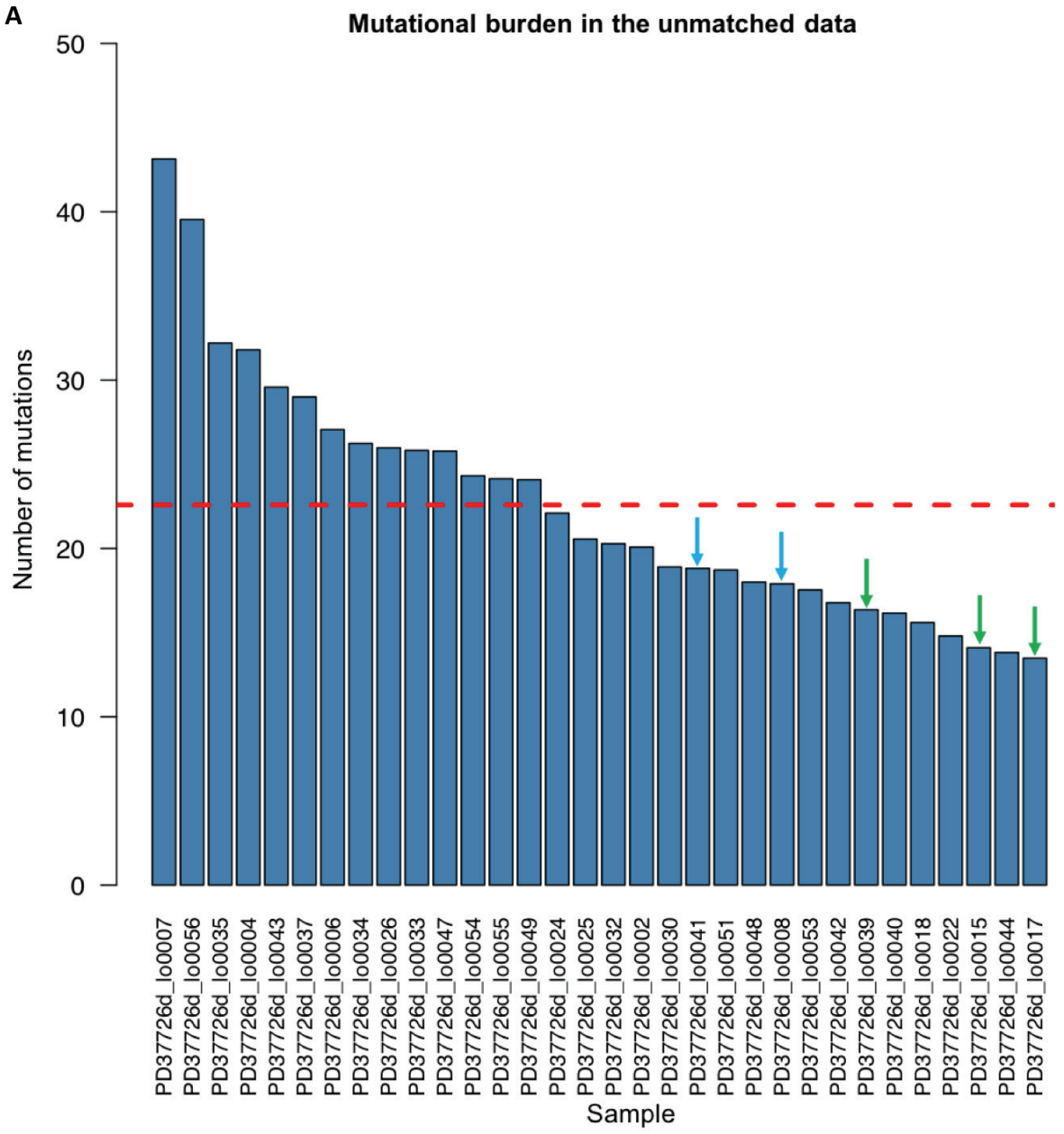
JBrowse screenshot showing the sequencing reads from sample PD37726d_lo0056 (Buels et al., 2016). Each horizontal bar is an individual read with the red representing a forward strand and blue being a reverse. The lighter the shade of these colours, the better the read quality.

The *LIPI* variant (21:15561402T>A) is highlighted by the green box. The large number of good quality reads, in both directions, carrying this variant support the notion that this is a genuine somatic mutation.

Nevertheless, analyses of somatic mutations from cancer genomes and healthy tissues suggests that most coding mutations accumulate effectively neutrally in somatic tissues, making it likely that these mutations are simple passenger events (Martincorena et al., 2017). In fact, given the exome represents 1-2% of the genome, the six non-synonymous mutations identified here are in keeping with the number of coding mutations expected by chance across 767 variants.

4.7 The observed mutational burden in the pancreatic islets is low

The mutational burden represents a snapshot of the detectable mutations in a sample. A high mutational burden would indicate a strong mutagenic process affecting the sample compared to a low mutational burden. In the unmatched data, 767 unique mutations were identified. In each islet, the number of mutations ranged from 13 to 43 mutations per cell, with a mean of 23 (Figure 19). Comparing the duplicates and triplicates, each have a similar mutational burden, as would be expected for identical samples.



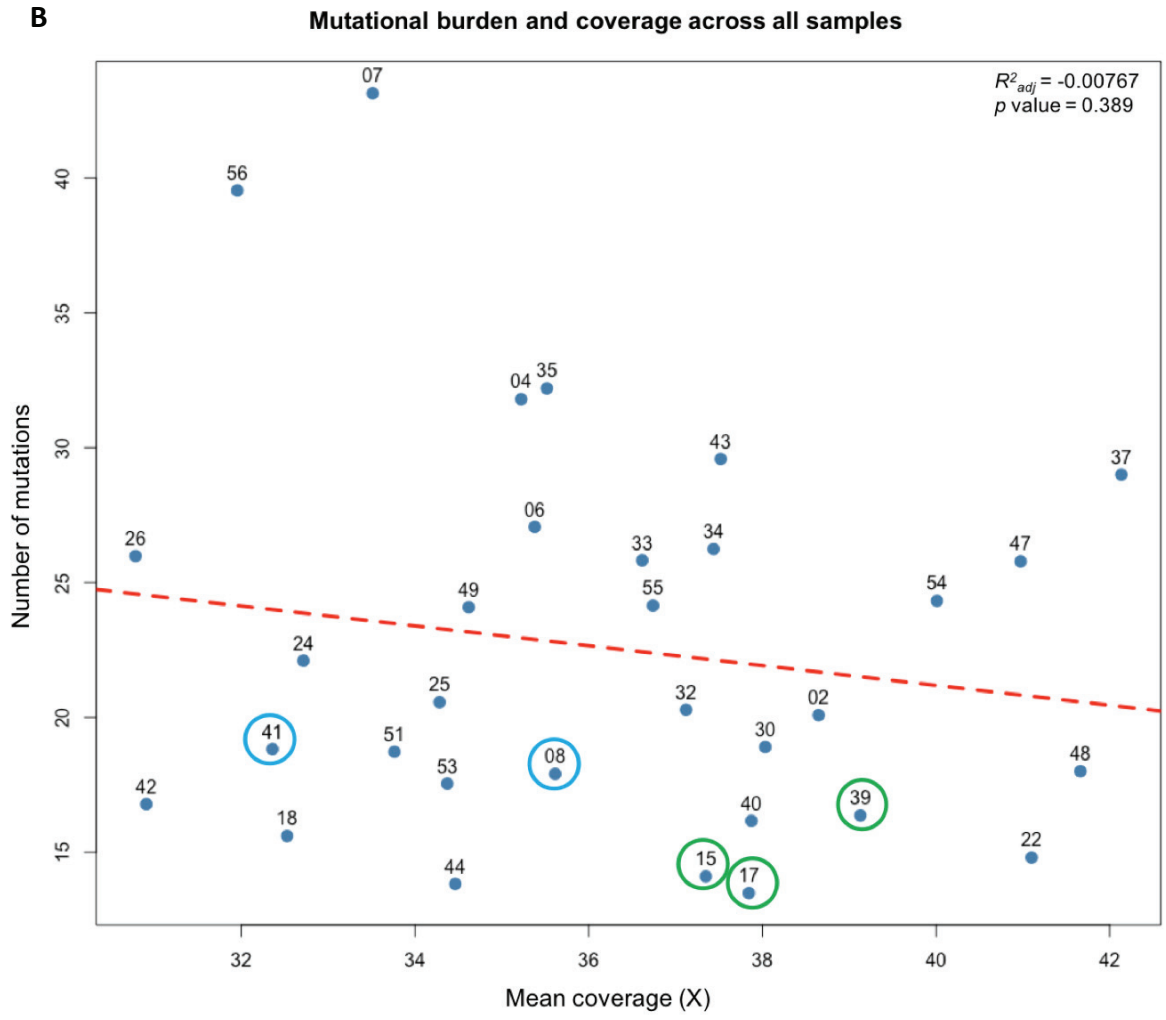


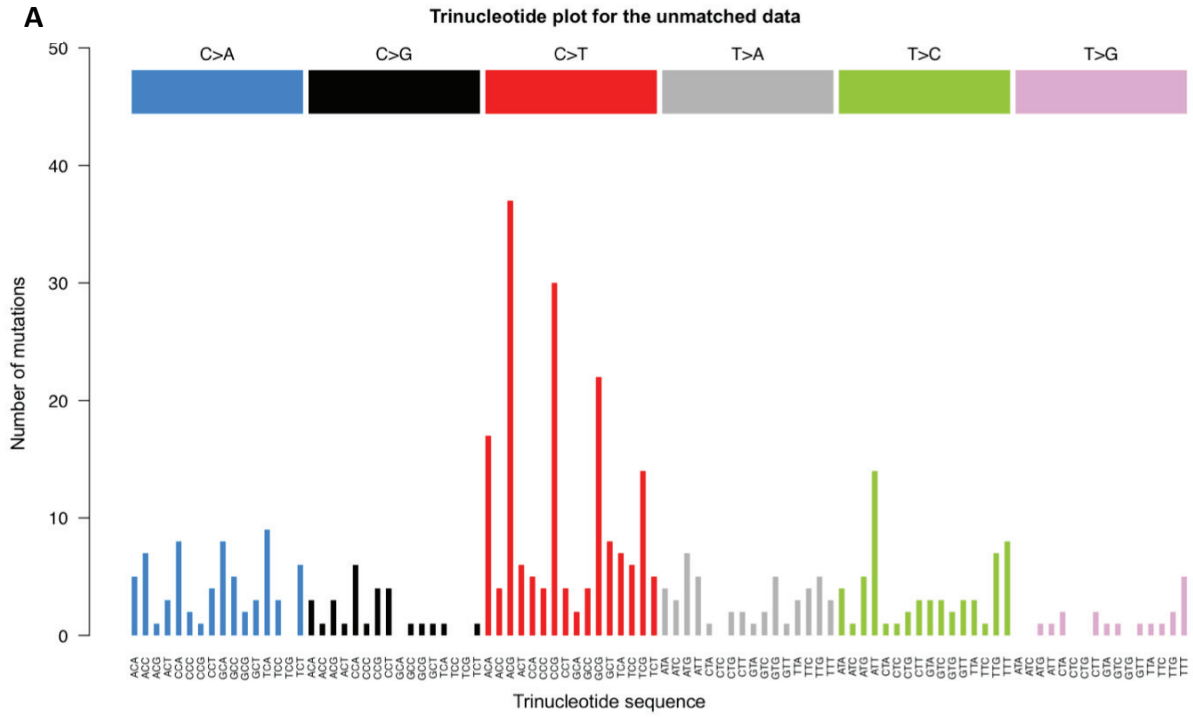
Figure 19 – The observed mutational burden in the pancreatic islets
Duplicates are highlighted in blue and triplicates in green.

(A) A bar plot highlighting the mutational burden (y-axis) across all 32 samples sequenced (x-axis). The mean (23) is shown by the dashed red line. The range is from 13 to 43 mutations. There appears to be a concordance between duplicate and triplicate samples.

(B) A scatterplot of the observed number of mutations (y-axis) compared to the mean coverage per sample (x-axis). Samples are labelled by the number that follows the "PD37726d_lo00" prefix. The dashed red line indicates the linear regression of this plot. There is no significant correlation between the mean coverage and the number of mutations (R^2_{adj} = -0.00767, *p* value = 0.389).

4.8 Intrinsic mutational processes dominate the pancreatic islets

In tandem with the mutational burden, eliciting the mutational signatures can help characterise the processes driving the accumulation of mutations. Amongst the 767 mutations in the unmatched data, there is a high proportion of C>T mutations (Figure 20A). These are particularly prevalent in CpG sites. Smaller numbers are seen in the T>C, C>A and T>A substitution classes. The mutational signatures extracted are dominated by signatures SBS5 and SBS1 (Figure 20B).



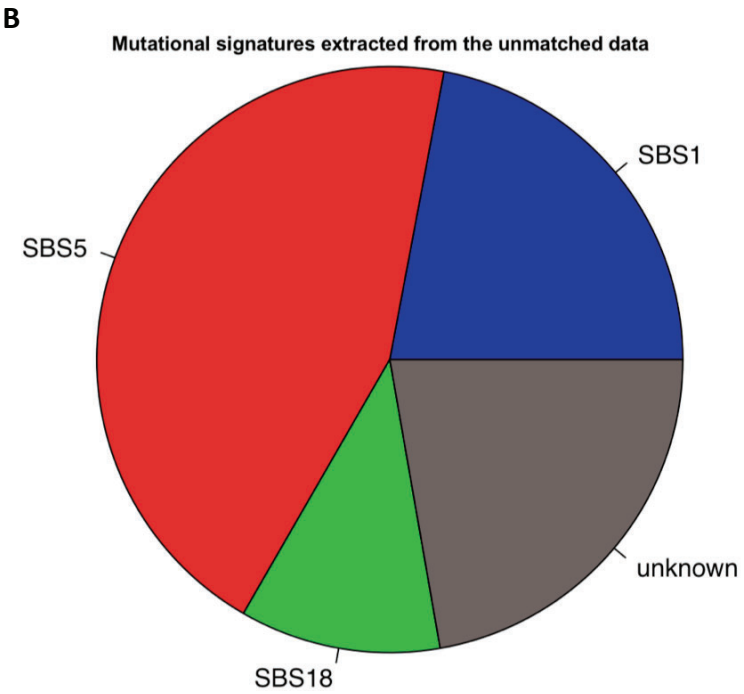


Figure 20 - The mutational signatures in the pancreatic islets
(A) The 96-trinucleotide bar plot for the 767 mutations in the unmatched data. C>T mutations are the most common base substitutions in the islets.

(B) Pie chart displaying the mutational signatures extracted from the 767 mutations in the unmatched data, using deconstructSigs (Rosenthal et al., 2016). Signature SBS5 is the dominant mutational signature, followed by signatures SBS1 and SBS18 (Alexandrov et al., 2018).

Both SBS5 and SBS1 have been found in all the cancer types analysed in the PCAWG data (Alexandrov et al., 2018). SBS1 is a well-understood mutational signature that arises from the spontaneous deamination of 5-methylcytosine at CpG sites, throughout the genome. As a result, it is made up of C>T mutations at CpG sites (Figure 21A) (Alexandrov et al., 2018).

Less is known about SBS5. The mutational profile of SBS5 is flatter, with all six pyrimidine substitution classes being affected, and C>T and T>C being the most common (Figure 21B) (Alexandrov et al., 2018). Studies of signature 1 and 5 from cancer genomes of different patients, across a range of ages, have shown they tend to increase with age. This suggests an ongoing process, occurring throughout life at a relatively constant rate (Alexandrov et al., 2015; Alexandrov et al., 2018). Considering the ubiquity of SBS5 and the similarities to the intrinsic process represented by SBS1, it is likely that SBS5 is also an intrinsic mutational process.

SBS18 appears to be an entirely different mutational process (Figure 21C). Found in many cancers, the C>A mutations are due to reactive oxygen species (Alexandrov et al., 2018). Given the low proportion represented here, it is possible that this signature was introduced during the processing and sequencing of the islet samples.

Figure 21 – The reference trinucleotide plots for the three mutational signatures extracted

The three reference signatures from the PCAWG data that match the extracted signatures from the pancreatic islets (Alexandrov et al., 2018).

(A) The 96-trinucleotide bar plot for signature SBS1 showing a high proportion of C>T mutations, in NpCpG sites.

(B) The 96-trinucleotide bar plot for signature SBS5 showing a flat mutational profile with higher numbers in the C>T and T>C substitution classes.

(C) The 96-trinucleotide bar plot for signature SBS18 showing isolated C>A mutations, particularly in the context of NpCpA/T.

4.9 The pancreatic islets are not clonal units

Each of the 32 pancreatic islets appears to be polyclonal (Figure 22). This is made clear by the VAF distribution being centred on means much lower than 0.5. Nevertheless, many of the islets also harbour a few mutations at high VAFs, some even approaching 0.5. Although some variation is expected due to binomial noise, some of these high VAF variants betray the existence of a dominant embryonic lineage in certain islets, as it is shown later. An example of this is seen in PD37726_lo0056. This sample hosts the previously described non-synonymous *LIPI* mutation at a VAF of 0.46.

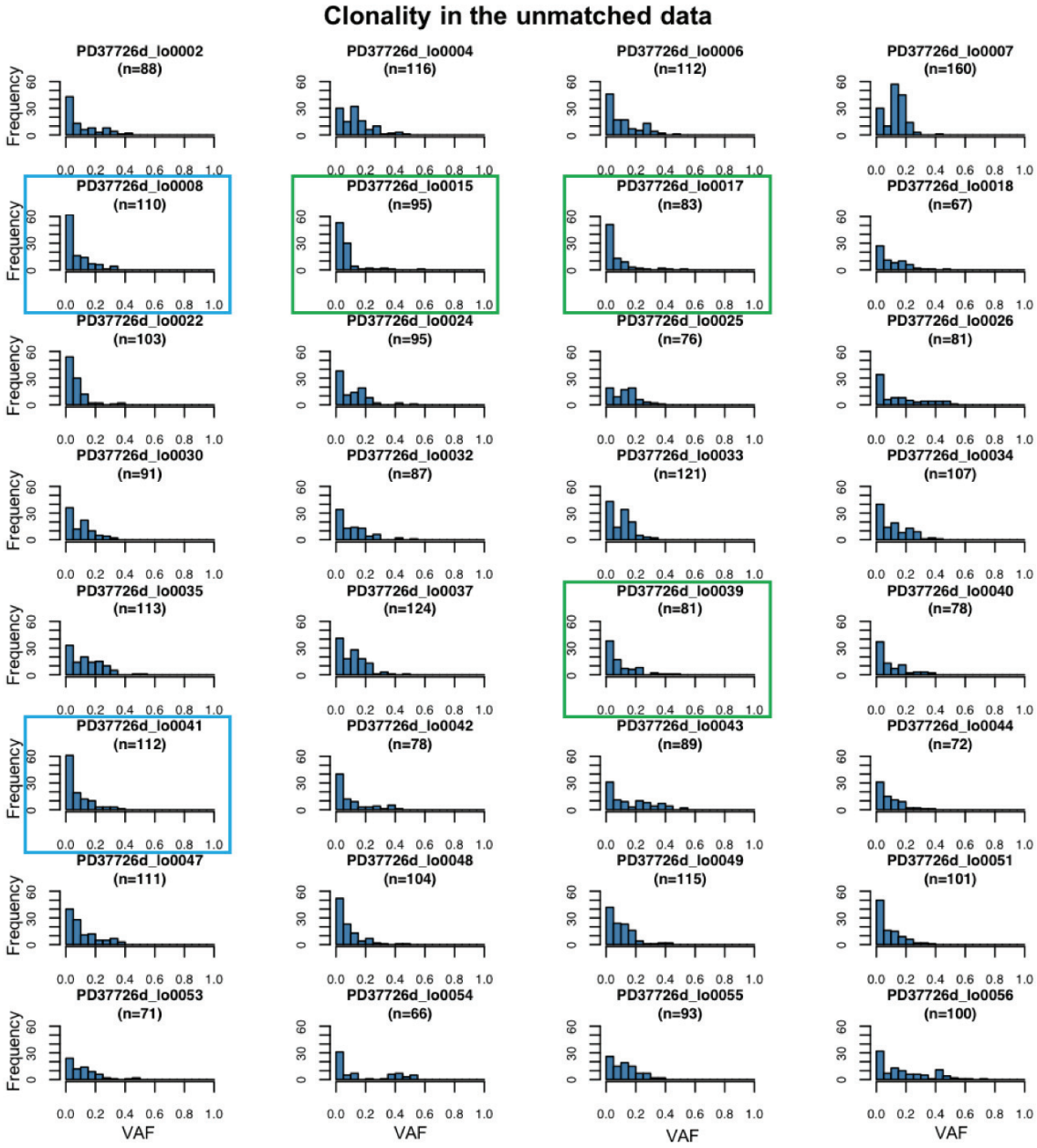


Figure 22 – The clonality of the 32 pancreatic islet samples

The VAF distributions in each of the 32 samples is shown with frequency on the y-axis and VAF on the x-axis. The duplicates and triplicates are shown in the blue and green boxes respectively. The number of mutations is noted underneath the sample name. Each islet appears to be polyclonal.

4.10 Phylogenetic reconstruction of the early embryonic lineage tree

To reconstruct the early splits in the phylogenetic tree, two additional criteria were applied to the 767 variants. The first was that the variants had to be shared in more than one sample. Secondly, each variant that was shared, had to be at a VAF>0.2 in each of the samples it was present in. The reasoning behind this is that an early embryonic variant would be expected to make up a significant proportion of the islet it is present in. Different combinations of these two criteria were trialled to identify the optimal combination and although relaxing them led to more variants being included, many of these were private mutations carried no additional phylogenetic information (Table 4).

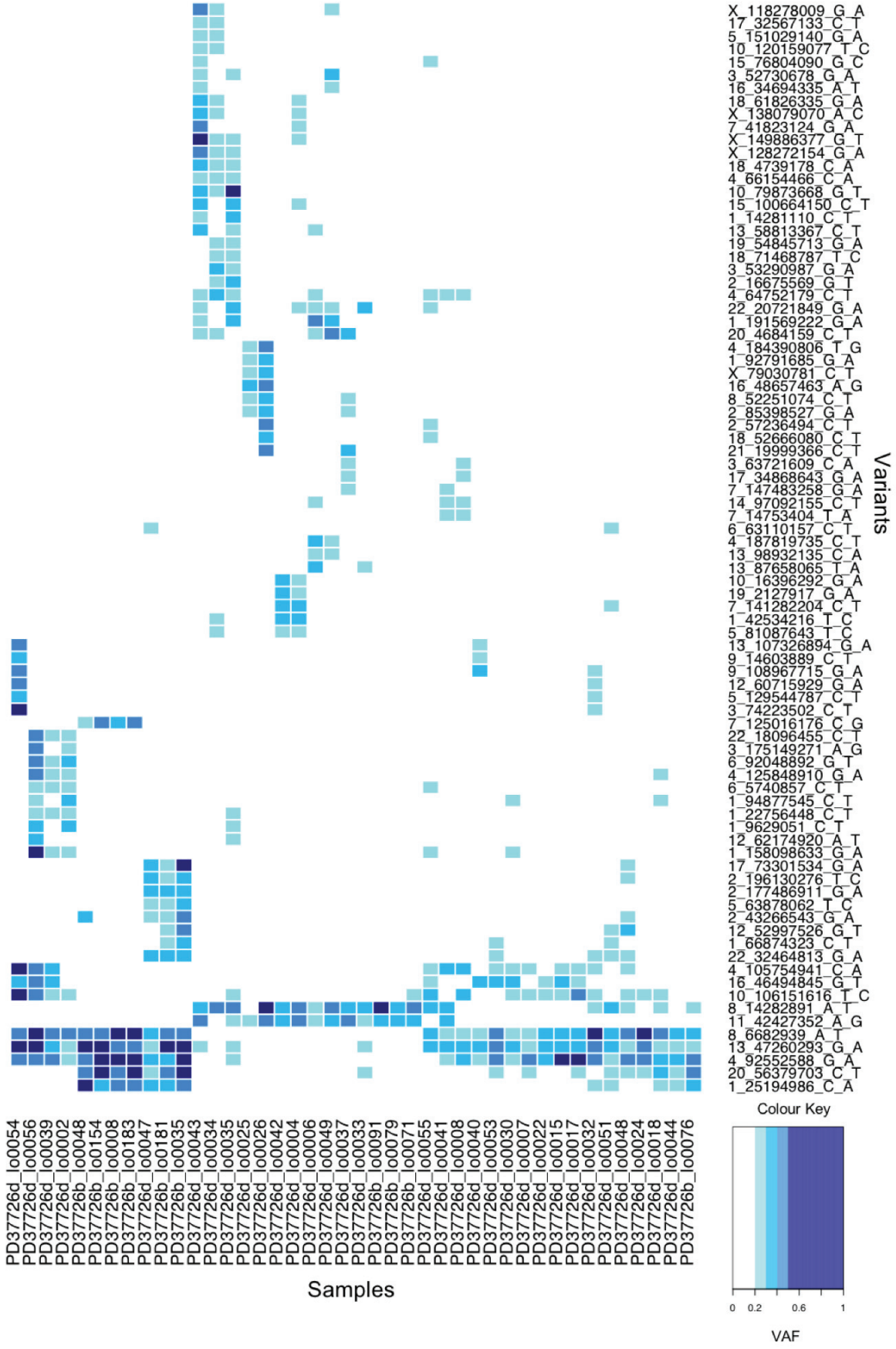
Table 4 – The different number of variants available for phylogenetic tree reconstruction, when including two additional criteria.

The permutations of these additional criteria did not significantly improve the phylogenetic tree reconstruction. This is because at VAFs nearing the limit of detection, the variants are harder to distinguish from each other, and from noise.

	VAF>0	**VAF>0.2**
All variants	767	261
Shared variants (in >1 sample)	623	84

Phylogenetic reconstruction was then undertaken with these 84 variants by examining the shared variants between samples. The ten bladder samples previously used for *in silico* re-genotyping were also included with the 32 islets in order to provide greater power in identifying clusters. The mutational spectrum of these 84 variants and the clusters identified between samples, are both shown in Figure 23.

A


Heatmap shared variants in >1 sample with a VAF>0.2
Variants
X_118278009_G_A
17_32567133_C_T
5_151029140_G_A
10_120159077_T_C
15_76804090_G_C
3_52730678_G_A
16_34694335_A_T
18_61826335_G_A
X_138079070_A_C
7_41823124_G_A
X_149886377_G_T
X_128272154_G_A
18_4739178_C_A
4_66154466_C_A
10_79873668_G_T
15_100664150_C_T
1_14281110_C_T
13_58813367_C_T
19_54845713_G_A
18_71468787_T_C
3_53290987_G_A
2_16675569_G_T
4_64752179_C_T
22_20721849_G_A
1_191569222_G_A
20_4684159_C_T
4_184390806_T_G
1_92791685_G_A
X_79030781_C_T
16_48657463_A_G
8_52251074_C_T
2_85398527_G_A
2_57236494_C_T
18_52666080_C_T
21_19999366_C_T
3_63721609_C_A
17_34868643_G_A
7_147483258_G_A
14_97092155_C_T
7_14753404_T_A
6_63110157_C_T
4_187819735_C_T
13_98932135_C_A
13_87658065_T_A
10_16396292_G_A
19_2127917_G_A
7_141282204_C_T
1_42534216_T_C
5_81087643_T_C
13_107326894_G_A
9_14603889_C_T
9_108967715_G_A
12_60715929_G_A
5_129544787_C_T
3_74223502_C_T
7_125016176_C_G
22_18096455_C_T
3_175149271_A_G
6_92048892_G_T
4_125848910_G_A
6_5740857_C_T
1_94877545_C_T
1_22756448_C_T
1_9629051_C_T
12_62174920_A_T
1_158098633_G_A
17_73301534_G_A
2_196130276_T_C
2_177486911_G_A
5_63878062_T_C
2_43266543_G_A
12_52997526_G_T
1_66874323_C_T
22_32464813_G_A
4_105754941_C_A
16_46494845_G_T
10_106151616_T_C
8_14282891_A_T
11_42427352_A_G
8_6682939_A_T
13_47260293_G_A
4_92552588_G_A
20_56379703_C_T
1_25194986_C_A
PD37726d_lo0054
PD37726d_lo0056
PD37726d_lo0039
PD37726d_lo0002
PD37726d_lo0048
PD37726b_lo0154
PD37726b_lo0008
PD37726b_lo0183
PD37726d_lo0047
PD37726b_lo0181
PD37726b_lo0035
PD37726d_lo0043
PD37726d_lo0034
PD37726d_lo0035
PD37726d_lo0025
PD37726d_lo0026
PD37726d_lo0042
PD37726d_lo0004
PD37726d_lo0006
PD37726d_lo0049
PD37726d_lo0037
PD37726d_lo0033
PD37726b_lo0091
PD37726b_lo0079
PD37726b_lo0071
PD37726b_lo0055
PD37726d_lo0041
PD37726d_lo0008
PD37726d_lo0040
PD37726d_lo0053
PD37726d_lo0030
PD37726d_lo0007
PD37726d_lo0022
PD37726d_lo0015
PD37726d_lo0017
PD37726d_lo0032
PD37726d_lo0051
PD37726d_lo0048
PD37726d_lo0024
PD37726d_lo0018
PD37726d_lo0044
PD37726b_lo0076
Samples
Colour Key
0
0.2
0.6
1
VAF

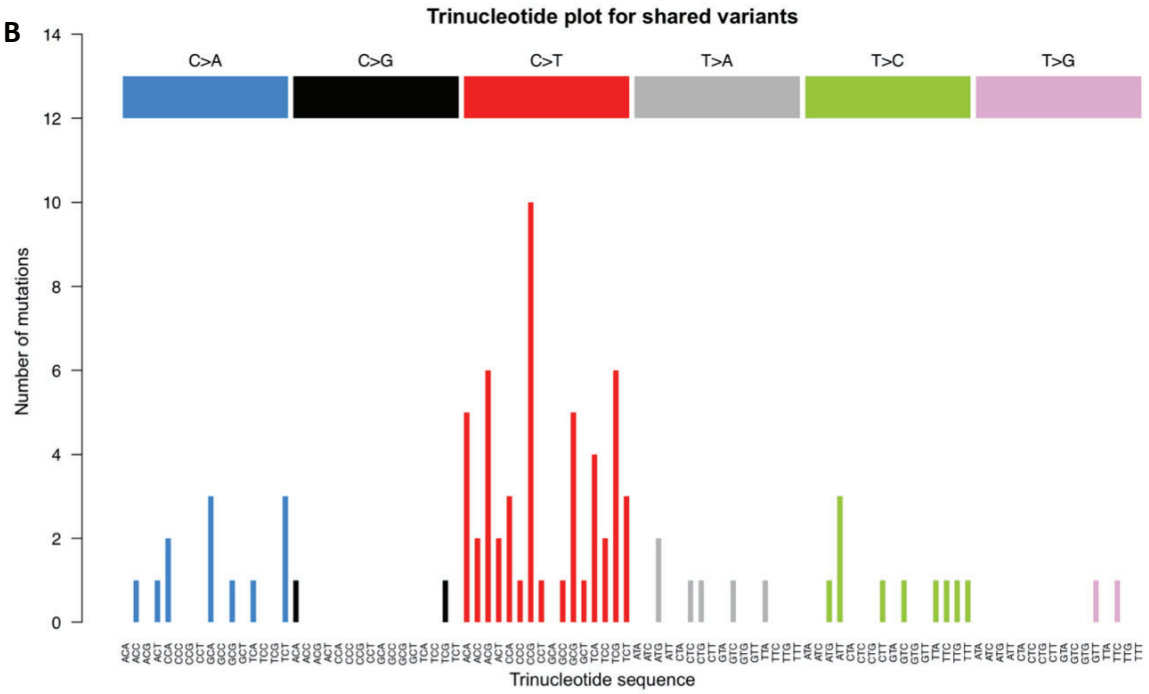


Figure 23 – The 84 variants used for phylogenetic tree reconstruction

A) Heatmap displaying the 84 variants (y-axis) across all 42 samples (x-axis). Islets are prefixed with "PD37726d" while bladder urothelium samples are labelled as "PD37726b". A VAF key is located in the bottom right. The shades of blue highlight the presence of a variant at a VAF greater than 0.2. Those variants with a VAF less than 0.2, and those absent from the samples, are in white. Clusters of samples that share variants can be seen by the groupings of blue. Generated using the R package gplots (v3.0.1, https://CRAN.R-project.org/package=gplots) (Warnes et al., 2015).

(B) The 96-trinucleotide bar plot shows few mutations but a clear dominance of C>T mutations, in the red.

The heatmap in Figure 23A reveals how different groups of mutations (rows) contribute to very different extents in individual islets (columns). The existence of clustering evidences the presence of different genetic lineages dominating each islet. The heatmap also reveals that, while some islets appear to have high VAFs from a single group of mutations, others show moderate VAFs from different clusters.

To formalise the observations above, an n-dimensional hierarchical Dirichlet process (n-HDP) was then employed to identify clusters of mutations with VAFs consistent across samples. These clusters were then used to reconstruct a phylogenetic tree (Appendix 8.1). The n-HDP algorithm ran for 15,000 iterations and the first 10,000 of these were discarded (Figure 24). Fifteen clusters were determined to be the optimal solution, of which three were very similar (Figure 25).

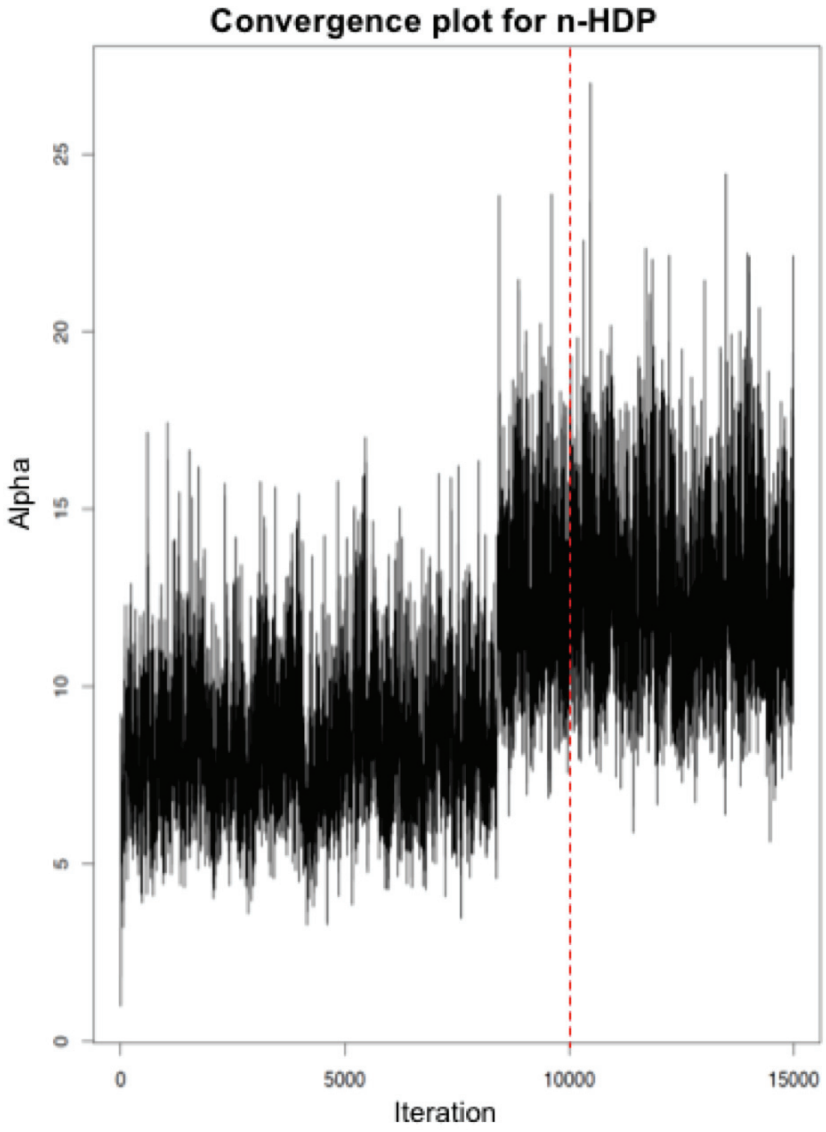


Figure 24 - The convergence plot generated by n-HDP

The number of iterations is seen along the x-axis while the y-axis is the number of clusters (alpha). The first 10,000 iterations were discarded, marked by the dashed red line.

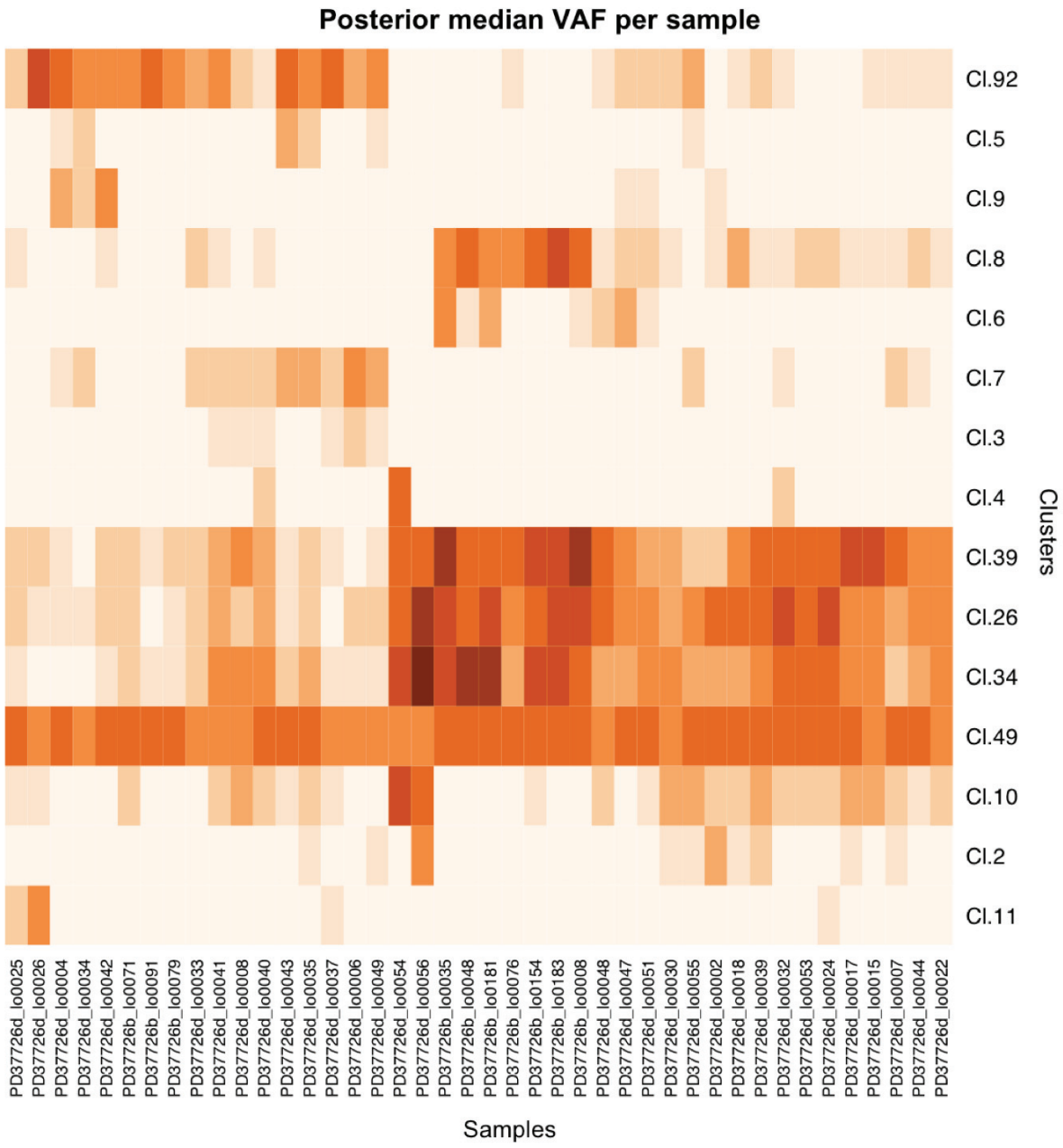


Figure 25 – The n-HDP clustering output

Heatmap of the clusters (y-axis) per sample (x-axis). Islets are prefixed with "PD37726d" while bladder urothelium samples are labelled as "PD37726b". Darker colours represent higher median VAFs. Shared clusters amongst samples can be seen.

Cluster 49 is striking as the two mutations that make up this cluster, appear to contribute equally to all samples. On manual inspection, these two variants appear have good read quality, adequate depth and presence in many samples. The

mutations are 22:32464813G>A and 1:66874323C>T with mean VAFs of 0.107 and 0.079 respectively. As this cluster did not segregate the samples, it was discarded when reconstructing the phylogenetic tree. The two mutations making up this cluster were manually inspected and showed a low VAF across many samples, in reads of good quality. It is likely therefore that cluster 49 reflects the inability by the n-HDP algorithm to appropriately assign a cluster to these variants, without violating the hierarchy. Running the n-HDP algorithm with relaxed criteria, to allow an increased number of mutations, will likely remove this cluster.

Analysing the clustering heatmap in Figure 25, it is clear that all samples show mutations from the cluster trio of 39, 26 and 34 or from cluster 92. The dichotomous nature of this implies these clusters represent the first split in the phylogenetic tree. This is supported by applying the pigeonhole principle to the fraction of cells carrying the variant across islets. This can be exemplified using the boxes in Figure 26, which show the estimated fraction of cells carrying the mutations in each cluster from two samples. The entire set of boxes for all 32 islets and 10 bladder samples is included in the Appendix 8.4 and these were generated in collaboration with Federico Abascal.

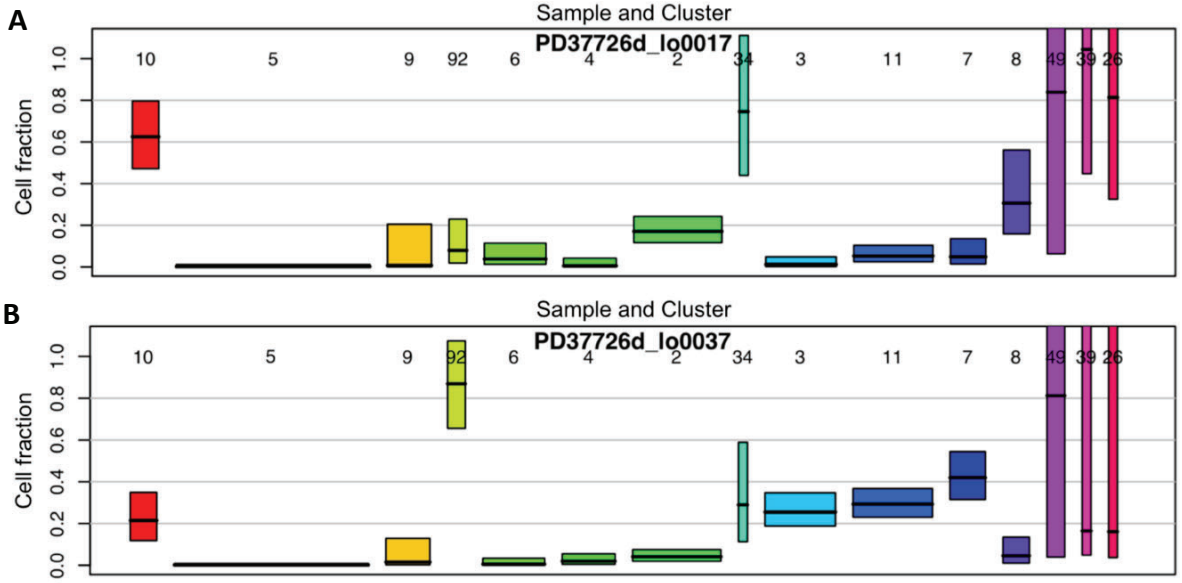


Figure 26 – The pigeonhole principle identified the phylogeny of clusters
Boxes each showing the cell fraction occupied by each cluster (y-axis). Cell fraction is equal to the VAF doubled. The numbers represent the cluster number assigned by n-HDP (x-axis). The sample name is at the top of each box. Box width is

proportional to the number of mutations while the length is the 95% credible interval. Cluster 49 (purple) was discarded as it appeared to be present in all cells.

(A) Sample PD37726d_lo0017 shows clusters 34, 39 and 26 are highly represented in the cell fraction, each contributing over 70%. These represent the first mutations in the ancestral lineage. Cluster 10 is the next largest fraction and is too large to be mutually exclusive, hence must be nested in the previous three clusters. Cluster 8 occupies 30% of the cell fraction and could be mutually exclusive to cluster 10 or nested within it.

(B) Sample PD37726d_lo0037 shows cluster 92 in nearly 90% of the cell fraction and cluster 7 at 40%. These represent the first two splits in the phylogeny. It becomes unclear for the third generation of the phylogeny, the exact nature of how cluster 10, 34, 3 and 11 are represented, given their low cell fraction does not distinguish between whether they are mutually exclusive or nested.

Looking at Figure 26, the higher the cell fraction is, the earlier this cluster occurred in the ancestry. The next largest fraction occupied by a cluster is then assessed by summing this with the first lineage fraction. If the sum exceeds 1, then this means both lineages cannot be present alongside each other (sibling clusters), but instead, the smaller lineage must be nested within the larger one. In this way, a second split can be defined. This can be confirmed by comparing the nesting to the shared clusters on the heatmap. For example, from Figure 26A, it is clear that clusters 34 and 10 are nested as their corresponding mutations account for approximately 70% and 60% of the cells of the sample (islet PD37726d_lo0017). By using this approach across all samples, multiple phylogenetic relationships can be identified. The smaller VAFs become harder to differentiate between lineages that exist alongside each other, and within each other. This limits the number of branches on a tree that can be reliably distinguished using subclonal decomposition and the pigeonhole principle.

By working through all 42 boxplots, 32 for the islets and 10 for the bladder, and using the heatmap generated from the n-HDP process, a conservative phylogenetic tree was reconstructed, with splits only drawn when the pigeonhole principle could confidently be applied (Figure 27).

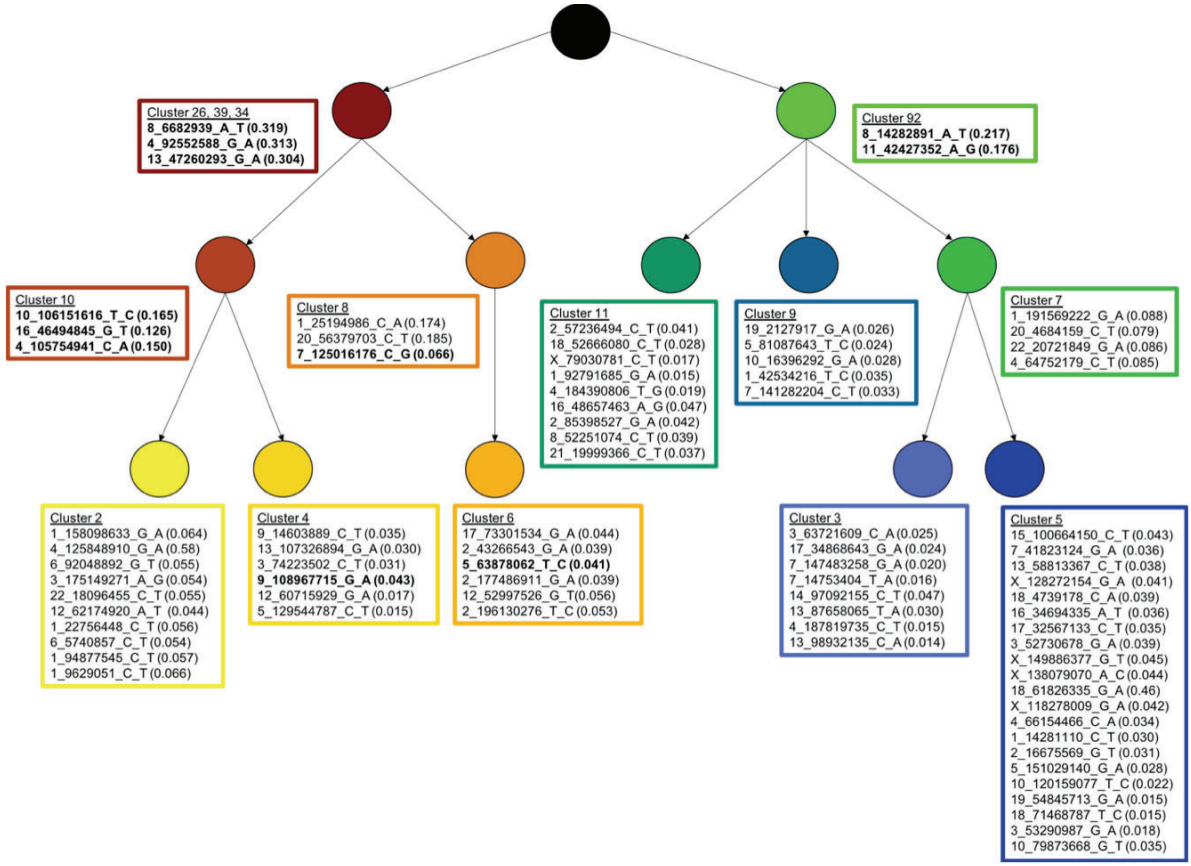


Figure 27 – The phylogenetic tree reconstructed with the n-HDP clustering
Each branch has an associated set of unique mutations with the mean VAF in brackets. The variants are assigned to an individual cluster, as per the n-HDP clustering. The variants in **bold** represent those recovered from the 30 exclusive mutations in the unmatched data set.

In the end, three generations were identified. A polytomy was also identified indicating that there are likely missing variants or silent divisions. The number of mutations per cluster varied from 2 to 21, with increasing numbers of mutations per cluster with each additional generation.

4.11 The early ancestry of the islets appears to be fully explained

As a validation step for this phylogenetic tree, the cumulative mean VAF for each level in the tree was then calculated, in the islets (Figure 28). This means summing the mean VAFs of the mutations from each branch of the tree, at each level. Per generation, or level in the phylogenetic tree, the cumulative mean VAF should sum to 0.5 if all lineages are accounted for. A shortfall here suggests a variant may have been

missed or placed in the wrong generation. For example, a given islet may be formed by 60% of cells derived from the first putative daughter cell on the left and 40% from the daughter cell on the right. If that is the case, we would expect the mean VAFs of the mutations in the left and right branches of the first level to be approximately 0.3 and 0.2 respectively, with both summing to 0.5, indicating no missing split in the tree.

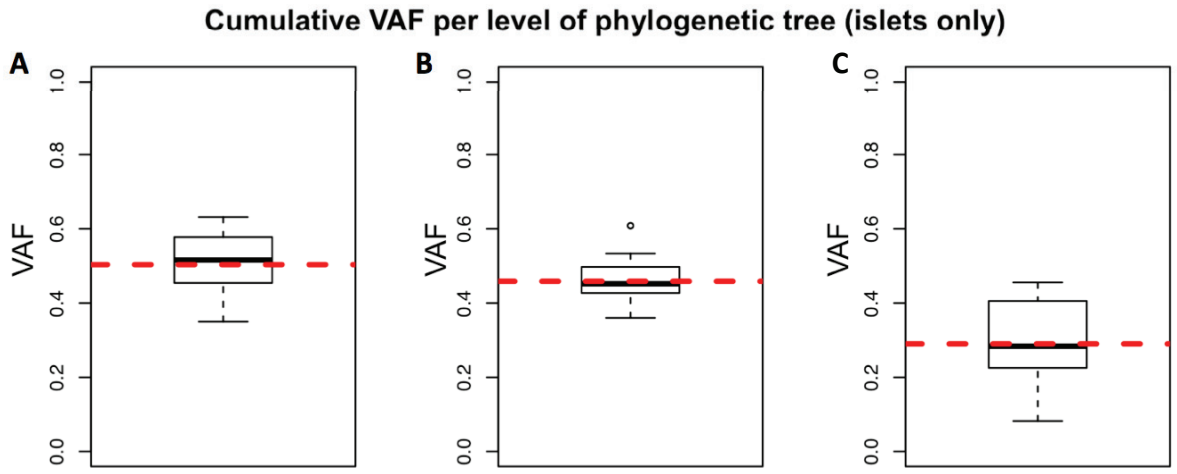


Figure 28 – The cumulative VAF for the islet samples

Boxplots detailing the distributions of all cumulative mean VAFs across samples from each of the three levels of the phylogenetic tree. Only the 32 islet samples are included. The median across all branches in the level of the tree is marked by the black line inside the box, while the interquartile range, between the first and third, is shown by the box margins. The upper whisker represents values that are greater than the third quartile, to a degree of 1.5x of the interquartile range. The lower whisker represents values that are less than the first quartile, to a degree of 1.5x of the interquartile range. The mean is shown by the dashed red line.

(A) The first split shows a cumulative mean VAF of 0.5.
(B) The second split is shown with a cumulative mean VAF of 0.46.
(C) The third split is shown with cumulative mean VAF of 0.29.

The early lineage of the islets appears to be well-explained by the n-HDP tree. The cumulative mean VAFs of the first and second levels are 0.5 and 0.46 respectively, suggesting that the key divisions involved in the early lineage, that relates all islets, are captured by the tree. The third level marks a clear change with a cumulative mean

VAF of 0.29. This level is likely under-described with regards to the branches and splits involved, with about 60% being represented. This is not unexpected and is a result of insufficient fractioning of mutations into different clusters by the n-HDP method. Incorrectly lumping multiple mutations into a single cluster will lead to clusters that conceal the different branches that would otherwise be arising here in the tree.

For example, cluster 8 may represent this. The C>G transversion at 7:12501676 carries a global VAF of 0.066. This is lower than the other variants found in the same cluster, both at 0.174 and 0.185. In fact, a mean VAF of 0.066 is more in keeping with the values found in the third level. Moving the variant from cluster 8 to a new cluster in level III, makes a significant impact on the cumulative mean VAFs, by increasing the second level cumulative mean VAF to 0.49 and the third level cumulative mean VAF to 0.32. Therefore, it is likely that the n-HDP clustering did not place this variant in the correct cluster, despite the seemingly correct identification of an early embryonic variant. Overall, the cumulative VAFs analysis suggests that the clusters in the third level of the new tree are not fully resolved.

4.12 The early embryonic lineages of the bladder appear incomplete

Compared to the islets, the ancestry of the ten matched bladder urothelium samples is less clear from these data (Figure 29).

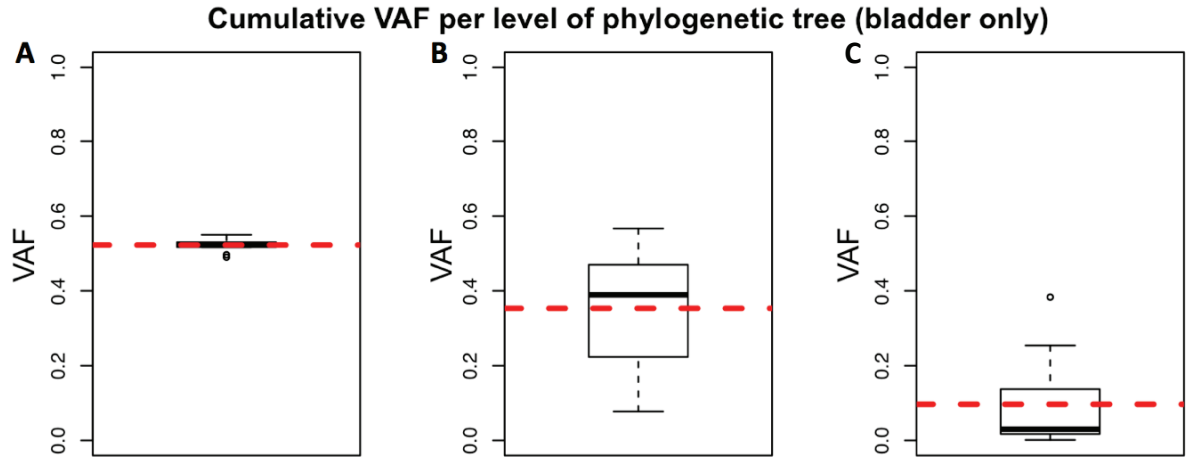


Figure 29 – The cumulative VAF for the bladder urothelium samples

Boxplots detailing the distributions of all the variants in each of the three levels marked on the phylogenetic tree. These are in the same format as those in Figure 28. Only the 10 bladder samples are included.

(A) The first split is shown with cumulative mean VAF of 0.52.
(B) The second split is shown with a cumulative mean VAF of 0.35.
(C) The third split Is shown with a cumulative mean VAF of 0.10.

Whilst the first level of the phylogenetic tree correctly accounts for all cells in the bladder samples, the second and third split appear to miss out key variants and branches. This is shown by the dramatic decrease in cumulative mean VAF. This results in 70% of the second level of the tree being explained and only 20% of the third split being accounted for. Further insights into these missing bladder variants and splits can be seen when combining both the bladder and islet samples (Figure 30).

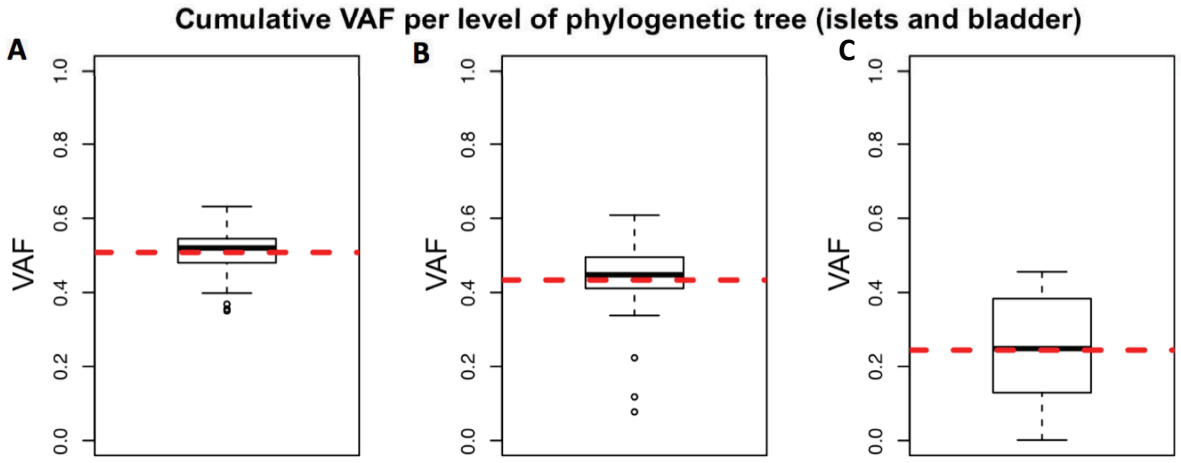


Figure 30 - The cumulative VAF for both the islets and bladder samples

Boxplots detailing the distributions of all the variants in each of the three levels of the phylogenetic tree. These are in the same format as those in Figures 28 and 29. All 42 islet and bladder urothelium samples are included.

(A) The first split shows a cumulative mean VAF of 0.5.

(B) The second split shows a cumulative mean VAF of 0.43. The three outlying samples at the lower end are the triplicate bladder samples PD37726b_lo0071, PD37726b_lo0079 and PD37726b_lo0091.
(C) The third split shows a cumulative mean VAF of 0.24.

The first level of the tree shows a cumulative mean VAF of 0.5, supporting the notion that this phylogenetic tree is rooted at the MRCA of the pancreatic islets and bladder urothelium. The second split is almost fully accounted for except for three outliers, with mean VAFs below the tail of the boxplot. These three bladder samples are actually a triplicate. Given their anomalous fitting in the boxplot, they were investigated further. This was done by analysing the VAF of each variant, from each cluster, in all three samples. Each branch in the level of the tree could then be compared to reveal the path taken by the samples through the phylogeny (Figure 31).

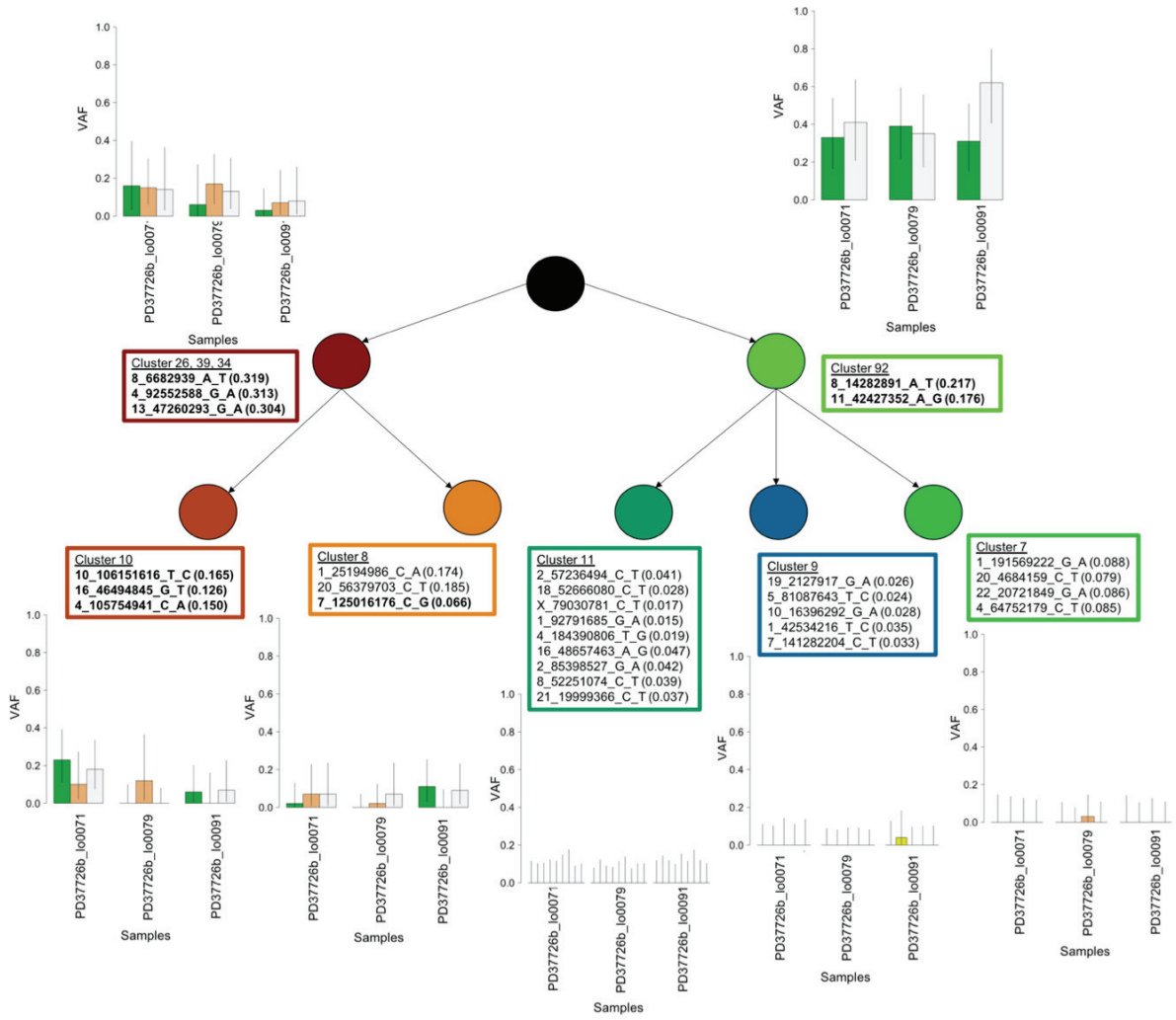


Figure 31 - The early phylogeny of the bladder triplicate

Each bar plot represents the associated branch of the tree, and hence one (or more) clusters of variants. The y-axis reports the VAF while the x-axis shows the three bladder samples (PD37726b_lo0071, PD37726b_lo0079, PD37726b_lo0091). Error bars depict 95% binomial confidence intervals.

The triplicate does not appear to be represented in the right-hand side of the tree in the second generation. This suggests a missing cluster that represents bladder urothelium, and not any of the islets sampled here.

Analysing Figure 31 shows that approximately two-thirds of the cells in this bladder triplicate derive from cluster 92. However, none of the three descendant lineages (clusters 11, 9 and 7) appear to contribute to this bladder sample. This confirms that there is at least one missing lineage in the phylogenetic tree that does not significantly contribute to any of the 32 islets, but instead give rise to most of the cells of this bladder urothelium sample. Therefore, the polytomy in Figure 27 must have at least one more branch.

4.13 Islets are composed of different embryonic lineages and are often dominated by one or two major lineages

Having demonstrated that the first two levels of the phylogenetic tree accurately reflect the embryonic lineages in the islets, the contributions that each individual embryonic lineage makes towards each islet can be studied in detail. A monoclonal foundation of the islet would be expected to have a single clear lineage while oligoclonal and polyclonal foundations, will produce a more heterogeneous picture with multiple variants, on different branches contributing to each islet (Figure 32).

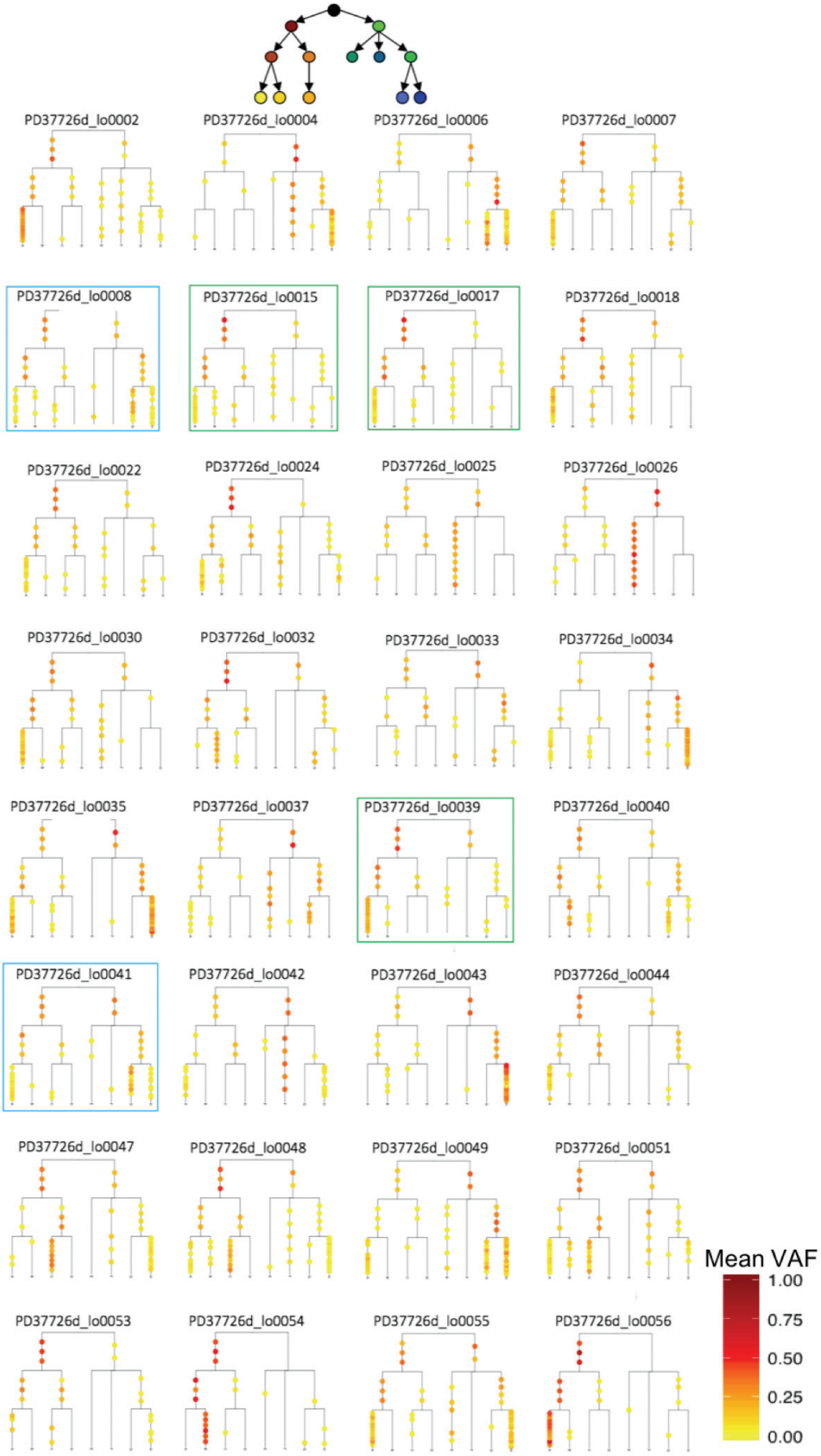

PD37726d_lo0002
PD37726d_lo0004
PD37726d_lo0006
PD37726d_lo0007
PD37726d_lo0008
PD37726d_lo0015
PD37726d_lo0017
PD37726d_lo0018
PD37726d_lo0022
PD37726d_lo0024
PD37726d_lo0025
PD37726d_lo0026
PD37726d_lo0030
PD37726d_lo0032
PD37726d_lo0033
PD37726d_lo0034
PD37726d_lo0035
PD37726d_lo0037
PD37726d_lo0039
PD37726d_lo0040
PD37726d_lo0041
PD37726d_lo0042
PD37726d_lo0043
PD37726d_lo0044
PD37726d_lo0047
PD37726d_lo0048
PD37726d_lo0049
PD37726d_lo0051
PD37726d_lo0053
PD37726d_lo0054
PD37726d_lo0055
PD37726d_lo0056
Mean VAF
1.00
0.75
0.50
0.25
0.00

Figure 32 – The phylogenetic trees of each of the 32 islets

The contribution of each lineage to each islet is depicted here using the R package, ggtree (Yu et al., 2016). These images were produced in collaboration with Tim Coorens. A schematic of the original reference tree is shown at the top of the diagram and a mean VAF key is in the bottom right. In all the trees, an extra branch has been added to cluster 8 (light brown node on the left side of the second level in the schematic) in order to show the silent division that distinguishes cluster 6, from cluster 8.

Each branch represents a new cluster and each coloured point is a specific, unique mutation, which has a fixed, consistent location across all the trees. The colour of the mutation corresponds to the mean VAF with red being 0.5 and yellow to white representing much lower values. Duplicates and triplicates shown in the blue and green boxes respectively.

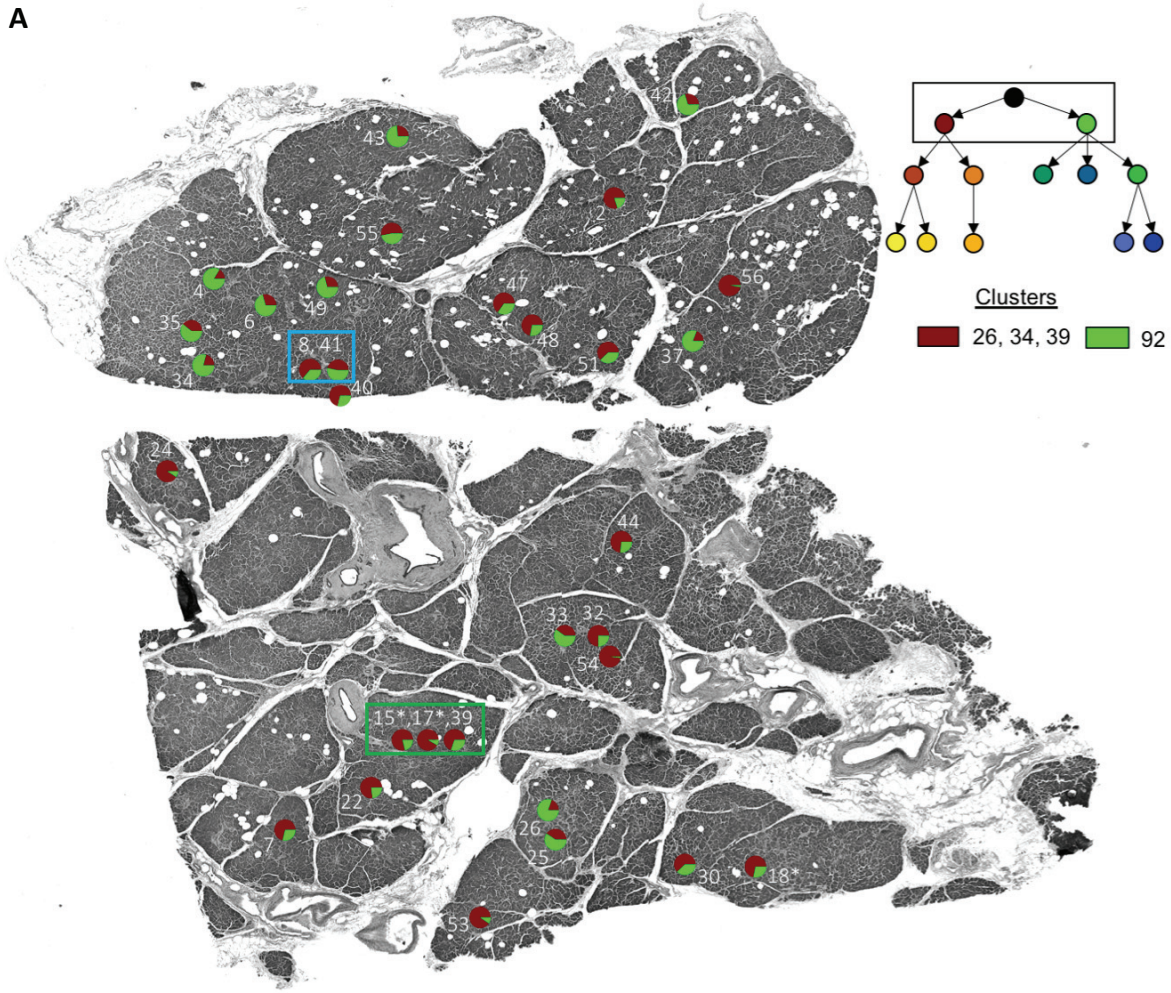
Figure 32 summarises how much the different embryonic lineages contribute to each islet. For example, sample PD37726d_lo0056 nearly entirely derives from the left most embryonic lineage of the tree. The VAFs close to 0.5 along the left-most branches show that the islet is almost entirely composed of cells descended from clusters 26, 39 and 34 in the first level, and clusters 10 and 2 lower down the tree. This is consistent with the high prevalence of the *LIPI* variant in this islet. While other lineages are present, shown as the yellow mutations on other branches, it is clear these explain only a small fraction of the cells in this islet. Contrasting this, other islets show a more even split between lineages such as PD37726d_lo0055. Thus, individual islets appear to be formed by multiple embryonic lineages but often have one or two dominant lineages contributing to most of the cells in the islet.

4.14 Integrating the spatial location of the islets with their lineages reveals a non-random distribution across the pancreatic tissue

By overlaying the contribution of different embryonic lineages onto the spatial locations of the islets, nearby islets can have their ancestry compared to each other. If the spatial distribution of islet embryonic lineages is random, this would suggest that early embryonic founding of each islet is independent of its neighbours. Contrasting this is would be a non-random distribution, whereby islets in the same spatial regions are

more similar to each other, than to distant islets. The latter scenario may suggest that different lineages preferentially seed the formation of islets in different areas of the developing pancreas, or alternatively, that islet fission is a common phenomenon in the formation or maintenance of the islets.

As the first two levels of the phylogenetic tree appear to give a near-complete picture of the early lineage tree, that gave rise to all 32 islets studied, the contribution to each islet by the lineages shown in the top two levels of the tree can be confidently portrayed, in combination with their spatial location (Figure 33).



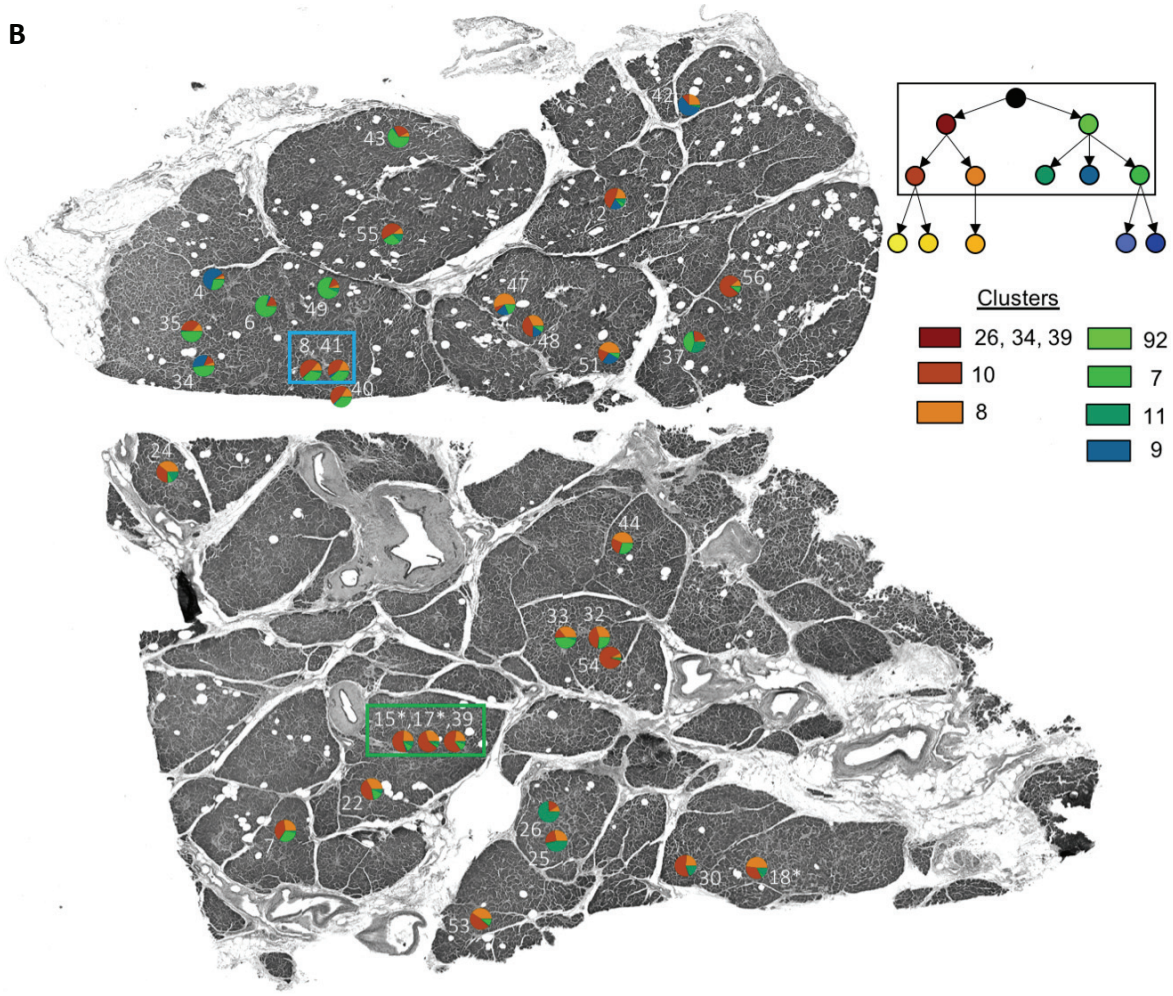


Figure 33 – The integrated spatial and phylogenetic information for each islet
Greyscale pancreas overview section used in LCM. Each pie chart represents the spatial location of an individual islet and the number labelling these pie charts is the suffix of each sample ("PD37726d_lo00*"). Islets labelled with a * are obtained from a z-slice 16 μm above or below this slice. Duplicates (8 and 41) and triplicates (15, 17 and 39) are displayed next to each other and are in blue and green boxes respectively. A schematic of the phylogenetic tree generated by the n-HDP is shown in the top right with a legend below, displaying the corresponding colours and names for each cluster.

(A) The proportions of the pie charts are related to the first split in the phylogenetic tree. As such, the fraction of the pie chart in brown (clusters 26, 34 and 39) represents the cell fraction descending from the left-sided branch of the first split, while green represents the right side (cluster 92).

(B) The proportions of the pie charts here relate to the branching of the second split in the phylogenetic tree. This includes clusters 10 and 8 on the left side of the tree, in shades of brown, and clusters 7, 9 and 11 on the right, in shades of green.

Figure 33 shows that the founding cells of the islets, and hence the embryonic linages that make up the islets, appear to be non-randomly distributed. Islets from the same area of the section show a more similar contribution of different linages than pairs of distant islets. For example, the top left region of Figure 33 shows several islets that share the same embryonic lineages, in similar proportions. These relationships are generally preserved across both levels of the phylogenetic tree. Demonstrating the precision of this approach, the duplicate and triplicate samples are consistent amongst themselves. This non-random distribution suggests that nearby islets are founded by the same population of ancestral cells, or that once founded, an islet can undergo a fission event and duplicate itself. Evidence of islet fission however is very limited and the current data is unable to distinguish between the two hypotheses (Jo et al., 2011; Seymour et al., 2004).

5. Discussion

5.1 LCM with an unmatched analysis may prove to be a reproducible workflow in other normal tissues

Capturing the landscape of somatic mutations within normal tissue is a growing field. At the Wellcome Sanger Institute (Cambridge, UK), a new pipeline using LCM of small areas of tissue and low-input DNA sequencing has been developed. Being able to take such precise biopsies, drawn freehand, makes few histological structures off limits. Whilst the spherical nature of the islets does pose an additional challenge in obtaining a complete three-dimensional sample, the use of multiple z-slices taken within the same islet, allows an approximation of the entire spherical islet to be made.

The study of somatic mutations has typically relied on matched data with the acknowledgement of the limitations this carries with regards to early embryonic mutations and phylogenetic reconstruction. However, the unmatched workflow presented here supersedes this matched approach. Germline mutations were confidently removed, identical somatic variants were called and, importantly, early embryonic mutations were recovered with minimal introduction of artefacts. Utilising this workflow, both prospectively and retrospectively, to similar data sets from other tissues could help decipher their somatic mutational profiles and early embryonic phylogenies. As the field grows and more normal tissues are investigated, the work here could prove pivotal in directing future somatic mutation research.

5.2 Novel insights have been gained into somatic mutations of the pancreatic islets

The almost unprecedented very low number of somatic mutations identified in each whole genome, as well as the pattern of mutation sharing across islets and across the pancreas and bladder, strongly suggests that many of the mutations detected in this analysis occurred during early embryogenesis. Signature analysis of the mutations detected confirmed that the majority of them can be assigned to intrinsic mutational processes without clear evidence of mutagen-induced mutations. This is perhaps to be expected for mutations of embryonic origin.

As expected given the very low mutational burden, there were few non-synonymous mutations identified. One notable variant in the *LIPI* gene stands out as a high VAF coding mutation (sample PD37726d_lo0056). Its high VAF in the absence of aberrant

copy number changes confirms that the mutation is present in approximately 90% of the islet cells (95% confidence interval: 56%-100%). Consistently, this islet appears to be mostly derived from a single branch of the phylogenetic tree. The high VAF of this variant could be consistent with it being an early passenger mutation present in an embryonic cell that gave rise to most cells in this islet, or with a later clonal expansion by drift or positive selection. Nevertheless, the low mutational burden in this islet, not dissimilar from other more polyclonal islets, suggests an early embryonic origin.

An interesting observation from this study is that the MRCA of all cells in the 32 islets of Langerhans also appears to be the MRCA of all cells in the ten matched bladder urothelium samples. This resembles the observation from a previous study in mice that showed the MRCA of the endoderm is also the MRCA of the ectoderm and mesoderm, with the suggestion this is likely to be the zygote (or at least the cell that gave rise to seemingly all cells in the adult) (Behjati et al., 2014). From the results presented here, it is unclear whether the MRCA of the pancreatic islets and bladder urothelium is the zygote, or whether it is simply the first cell that gave rise to all adult tissues.

5.3 The founding model of the pancreatic islets is still only partially understood

The presence of so many different embryonic lineages in each islet indicates that multiple embryonic founding cells of different lineages, come together to seed each islet. Given that many islets have all five branches of the second generation of the phylogenetic tree represented in their phylogeny, this suggests that at least five embryonic founding cells established these islets. This is consistent with a polyclonal founding model (Model C, Figure 8) for the pancreatic islets. However, the phylogeny of most islets suggests the existence of dominant lineages disproportionally contributing to each islet (Figure 34). This is in keeping with previous studies showing asymmetric contributions being made to adult tissues, from embryonic ancestors (Behjati et al., 2014; Ju et al., 2017).

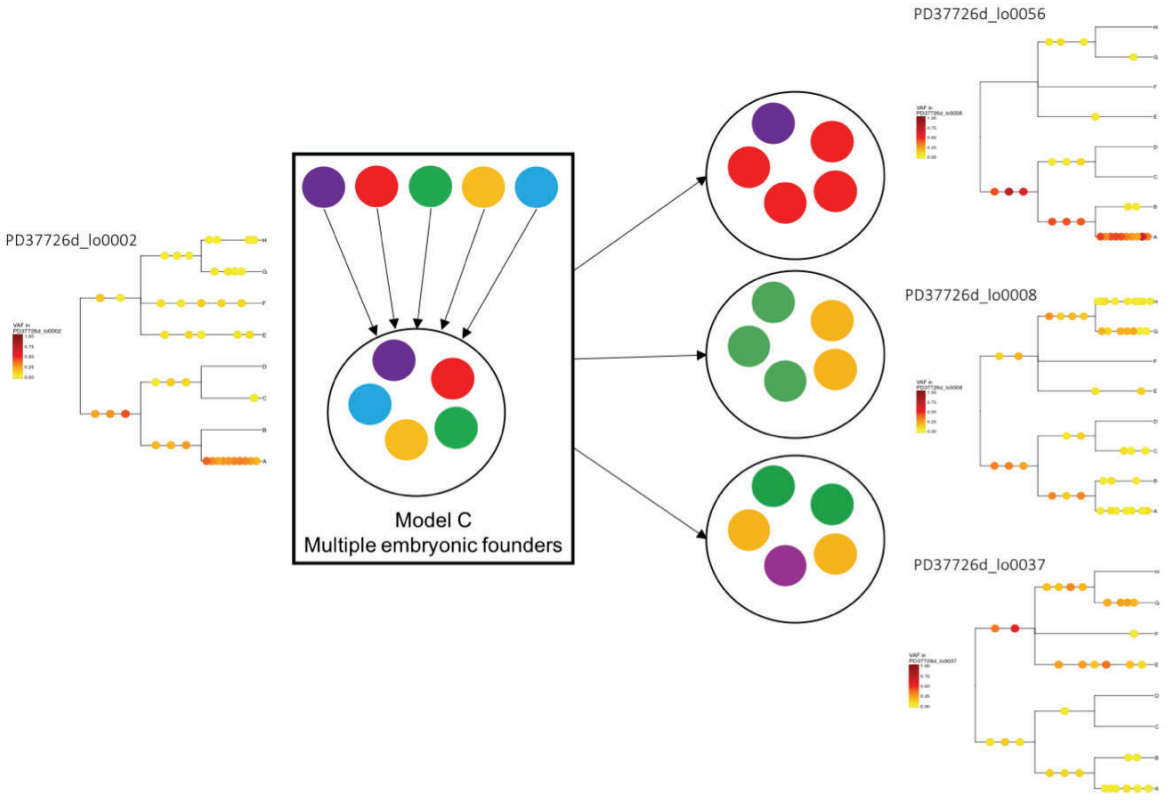


Figure 34 – Multiple embryonic founders may seed an islet

One interpretation of the data is that multiple founding cells (coloured circles) appear to make up the islets (black circle) (Model C, Figure 8). However, the proportion of lineages seeding the islet may vary. Some islets may be entirely polyclonal, whereby all five lineages occupy an equal proportion of the population (such as that represented in the box) or they may be oligoclonal, with examples shown on the right of the diagram. Indeed, some of the islets here appear to be more clonal than others, with PD37726d_lo0056 a notable example.

Since the pancreatic islets also contain other minor cell types, such as non-endocrine cells like endothelial cells, it remains to be shown whether the existence of a dominant lineage in many islets is due to a monoclonal origin of all endocrine cells, in a given islet, or whether endocrine cells in a single islet truly arise from multiple embryonic progenitors. Future studies combining information from adjacent areas of exocrine pancreas or additional phenotyping information from immunohistochemistry or single-cell RNA sequencing, may shed light on whether different lineages contribute to different cell types.

The results shown here are in stark contrast to the somatic mutations that can be detected in fast dividing clonal tissues like the colonic crypts. The patterns seen here in the islets instead suggest the islets are formed by a few founder cells early in development and do not undergo any subsequent clonal sweep later in life, as this would be expected to come with an increased burden of clonal mutations. This is consistent with the current belief that islets are maintained by the infrequent division of many cells, such as self-duplication of differentiated β-cells, and seem inconsistent with islets being replenished, or formed, in adulthood by one or a few cells (Bonner-Weir et al., 2012; Dor et al., 2004). Nevertheless, the findings here relate only to the 32 islets sampled.

Spatially, the founding of these islets appears to have been a non-random process, with islets nearby sharing similar ancestry. Further statistical methods would be necessary to quantify the extent of this with one option being a permutation approach with the islet positions and relatedness. The implication of this non-random distribution is that the same ancestral founding cells, occurring early in development, may seed regions of the pancreas during embryogenesis. Alternatively, once formed, islets may then divide into two identical islets through islet fission (Seymour et al., 2004). Several pairs of islets are seen that are close to each other with a high degree of resemblance in Figure 32 and these may represent examples of either the same founding cells or fission (Figure 35). Indeed, the two may not be mutually exclusive.

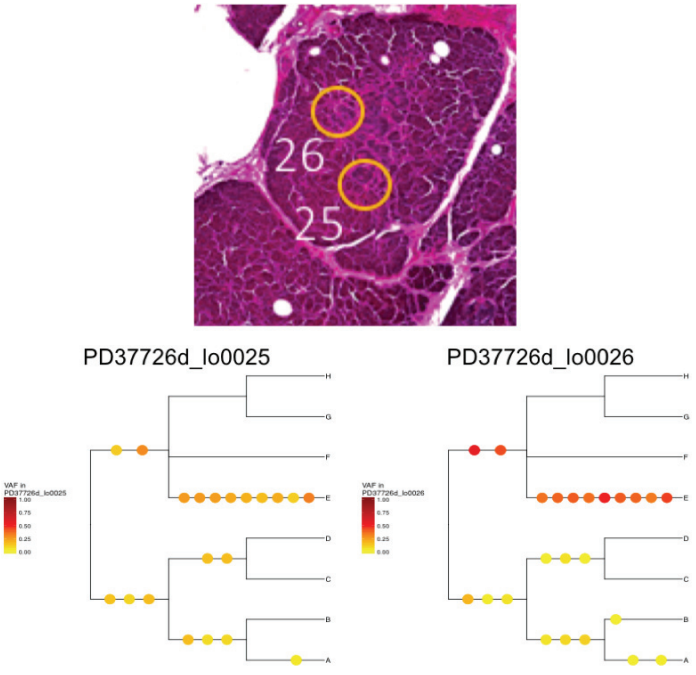


Figure 35 – Possible islet fission events are difficult to differentiate from similar founder populations

In light of their proximity to each other, samples PD37725d_lo0025 and PD37726d_lo0026 appear remarkably similar in their early phylogenetic ancestry. This may be due to an identical founder population or a fission event. Distinguishing the two will be difficult with the current method.

5.4 Limitations of the current methodology

There are several domains where improvements could be made. While LCM appears to be well-suited to the task, there are risks of contamination between plate wells. This could be from other samples during the cutting process, or when preparing the plate. In this study, this was closely monitored during the LCM stage through proper precautions and cleaning steps. The concordance of the duplicates and triplicates provides further support that there was little contamination of samples.

One key limitation was that all the islets were excised from a single biopsy, in one region of the pancreas, from a single patient. Whilst deep sampling of a single donor is necessary for phylogenetic reconstruction, multiple donors would be needed to make generalizable conclusions. To improve on this, many more islets would be required, from numerous biopsies across a range of donors.

Mutational signature analysis is a fast-growing field and has revealed numerous insights in cancer types. However, current knowledge of the mutational processes active in normal tissues is in its infancy. Whilst the same mutational processes may be occurring in normal tissue, there might be mutational signatures that are specific to certain normal tissues, and therefore not contribute sufficient numbers to malignancies for them to be detectable. By restricting the current analysis to signatures found in cancer genomes, the contribution of other signatures may have been overlooked. A *de novo* approach to mutational signature extraction may prove helpful in ascertaining the mutational processes specific to normal tissue, particularly when working with multiple patients. The R package HDP (https://github.com/nicolaroberts/hdp) (Roberts, 2018; Teh et al., 2006), could be used to do this. Still, one practical issue with *de novo* extraction is the need for a large number of mutations to work with, from

many different samples or patients with different contributions of these mutational signatures. Given the low mutational burden detected here, a *de novo* signature extraction would require a much larger number of islet whole genomes.

The minimum detectable VAF here was approximately 0.1, based on the WGS coverage and the CaVEMan default parameters. For a polyclonal tissue like the islets, the detectable variants were mostly ancestral mutations, most of which might have been present in the founder cells of an islet. Somatic mutations acquired through life by islets cells are unlikely to be present in a sufficient fraction of cells of an islet to be detectable. As a result, the mutational burden estimated in this study is expected to heavily underestimate the mutational burden of individual islet cells at the time of death of this donor, and instead likely represent the mutational burden of the founding cells.

A much greater coverage would enable the detection of more recent variants, and this would also be advantageous when looking at phylogenetic reconstruction. This would still be hindered by the polyclonal composition of the islets and the intrinsic errors introduced in sequencing. A possible solution to improve the recall of rare, or even private, somatic variants may be bottleneck sequencing (BotSeqS), whereby molecular barcoding combined with a dilution step prior to whole-genome library preparation, can dramatically increase the ability to identify those low VAF variants (Hoang et al., 2016). Additionally, the use of single-cell genomic and transcriptomic sequencing (G&T-seq) plus single-cell derived organoids, may play a role in the future somatic mutation workflow, particularly in polyclonal tissues (Enge et al., 2017; Jager et al., 2018; Macaulay et al., 2015).

Summarising, the sequencing of 32 pancreatic islet whole genomes, all from a single donor, has shown an unmatched analysis to be superior to a matched approach. The observed somatic mutational burden in these islets appears to be low and driven by intrinsic processes. Further, the pancreatic islets seem to be polyclonal units, established by multiple embryonic founders, with major and minor lineages. They do not appear to be maintained by a fast-dividing stem cell population and their spatial distribution is non-random, suggesting regions of the developing pancreas are seeded by the same populations of founding cells. Finally, the islets also appear to share a MRCA with the bladder urothelium, going back likely to the fertilised egg.

6. Future Directions

6.1 Single-cell derived splenocyte colonies may help enrich the phylogenetic tree

The phylogenetic reconstruction completed here approximates the early embryonic lineage of the islets. In order to obtain a more precise and complete phylogeny of the early development in this donor, single-cell derived splenocyte colonies are being cultured, in collaboration with Elisa Laurenti at the Wellcome-MRC Stem Cell Institute (Cambridge, UK). In single-cell derived clonal populations, heterozygous variants in the original founding cell take on VAFs of 0.5 in the larger clonal population, irrespective of its original VAF in the tissue. Rare variants can therefore become detectable under current sequencing protocols, at moderate depths, allowing higher-quality phylogenetic trees to be reconstructed using standard phylogenetic approaches (Behjati et al., 2014).

6.2 Immunohistochemistry could play a role in explaining the lineage proportions

Most islets appear to be a polyclonal unit, derived from at least five embryonic founders. Often the islets appeared to have a major lineage accounting for over half of the population, along with several more minor sub-populations (Figure 32). Given that β-cells make up about 60% of the islet, with α-cells being 30% (Ionescu-Tirgoviste et al., 2015), it would be interesting to examine whether the proportions of the islet cell types correlates with the lineages present in the population.

It is unclear whether different lineages could reflect different cell types. It is likely that non-endocrine cells within an islet partially explain the presence of multiple embryonic founders, but this is difficult to assess with the current data. It is also unclear whether different subpopulations of endocrine cells in a given islet may derive from different founder cells, including differences between β-cells and α-cells. With a view of investigating further, immunohistochemistry can be performed on the sections for the markers expressed by each cell type. Key targets for this would include chromogranin, insulin, glucagon, trypsin and CD31 (Campbell-Thompson, Heiple, Montgomery, Zhang, & Schneider, 2012; Lin, Chen, & Wang, 2015; Pusztaszeri, Seelentag, & Bosman, 2006). Their relative fractions could then be compared to those obtained from the whole-genome sequencing data. In preparation for this, 5 μm sections have been obtained; from directly above and below the 16 μm sections cut for LCM. As a

tissue thickness of 5 μm represents less than a single cell layer, the sub-populations of different cell types in the islets excised can be experimentally identified. Building on this, G&T-seq may provide a superior option to identifying cell types within the islet.

6.3 Targeted genotyping of pancreatic tissues may reveal more detailed insights into development and maintenance

One key question is whether transdifferentiation of non-endocrine cells into endocrine cells, particularly β-cells, could be achieved (Bonner-Weir et al., 2008; Kim & Lee, 2016). This possibility has attracted considerable attention for its translational potential, in an age where DM2 has become more widespread, with a growing health burden (World Health Organization, 2016). To investigate transdifferentiation further, as well as more generally assessing the contribution of different lineages to the exocrine pancreas surrounding the islets, LCM has been used to obtain pancreatic ducts and acini from the same patient (290B) (Figure 36).

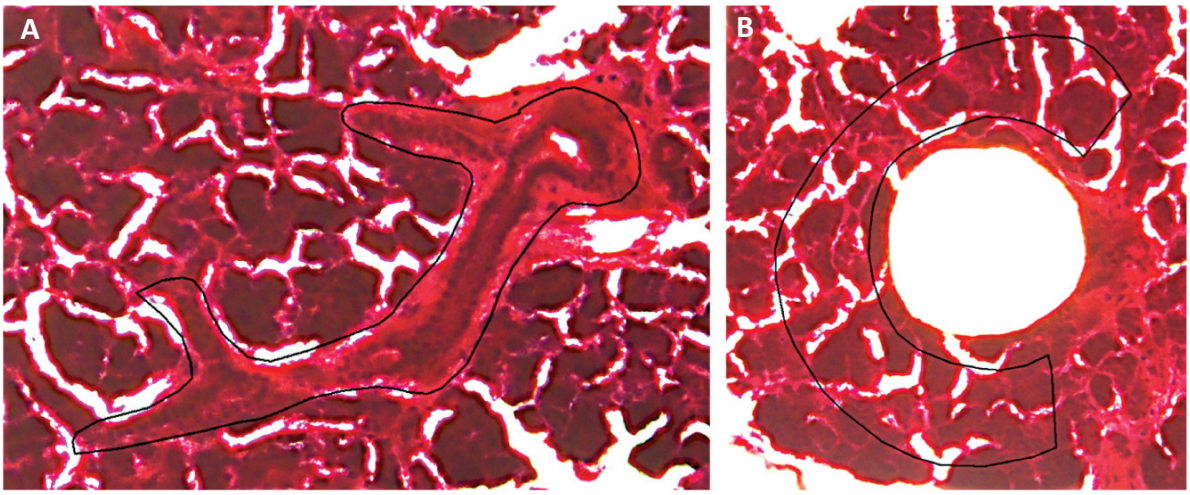


Figure 36 – Further LCM work is focusing on the pancreatic ducts and acini
LCM images from patient 290B. The black lines tracing the borders of the structure indicate the margins of the laser microdissection.

(A) A pancreatic duct demarcated during LCM. The area of tissue obtained was 28,466 μm^2.

(B) A crescentic acinar section surrounding islet PD37726d_lo0030, outlined in black. This has an area of 36,582 μm^2. The islet was previously excised and leaves a white circle in the centre of picture. There is a clear margin of tissue left between the islet and the acinar crescent, to ensure no remnants of the islets are cut with the acinar tissue.

As shown in Figure 36, while the ducts have been taken in the same way as the islets have, a different approach has been used for the acini. Crescents of acinar tissue surrounding islets are being collected with a clear margin of tissue being left between the islets and acinar crescents, to prevent cross-contamination of the cell types. With the whole-genome sequence data from the islets, ultra-deep targeted sequencing of the surrounding acinar crescents, and nearby ducts, can be performed using the mutations identified in the islet whole-genome data as a custom bait. The targeted genotyping of both of these exocrine tissues will then shed light on whether they share the same variants and therefore, ancestry.

To conclude, this body of work serves as a starting point for examining the somatic mutations in the pancreatic islets. Building on this foundation, future work is already under way. Single-cell derived splenocyte colonies are currently being cultured and once sequenced, the phylogenetic tree reconstruction will hopefully reveal an unrivalled insight into the phylogeny of the islets of Langerhans. Immunohistochemistry is a readily available resource to further clarify the lineages seen in each islet and this could frame these results in a more appropriate context. The prospect of transdifferentiation is exciting and if shared ancestry between the endocrine and exocrine pancreas is proven, a remarkable new frontier could open up in regenerative medicine, one that could lead to the development of novel, translational therapies for diabetic patients.

7. References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., Nik-Zainal, S., & Stratton, M. R. (2015). Clock-like mutational processes in human somatic cells. *Nat Genet, 47*(12), 1402-1407. doi:10.1038/ng.3441

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W. T., Boot, A., . . . Stratton, M. R. (2018). The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*.

Andrews, T. S., & Hemberg, M. (2018). Identifying cell populations with scRNASeq. *Mol Aspects Med, 59*, 114-122. doi:10.1016/j.mam.2017.07.002

Beck, P., & Daughaday, W. H. (1967). Human placental lactogen: studies of its acute metabolic effects and disposition in normal man. *J Clin Invest, 46*(1), 103-110. doi:10.1172/JCI105503

Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D. C., Tamuri, A. U., . . . Stratton, M. R. (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature, 513*(7518), 422-425. doi:10.1038/nature13448

Bjerknes, M. (1986). A test of the stochastic theory of stem cell differentiation. *Biophys J, 49*(6), 1223-1227. doi:10.1016/S0006-3495(86)83751-1

Bonner-Weir, S., Baxter, L. A., Schuppin, G. T., & Smith, F. E. (1993). A second pathway for regeneration of adult exocrine and endocrine pancreas. A possible recapitulation of embryonic development. *Diabetes, 42*(12), 1715-1720.

Bonner-Weir, S., Guo, L., Li, W. C., Ouziel-Yahalom, L., Lysy, P. A., Weir, G. C., & Sharma, A. (2012). Islet neogenesis: a possible pathway for beta-cell replenishment. *Rev Diabet Stud, 9*(4), 407-416. doi:10.1900/RDS.2012.9.407

Bonner-Weir, S., Inada, A., Yatoh, S., Li, W. C., Aye, T., Toschi, E., & Sharma, A. (2008). Transdifferentiation of pancreatic ductal cells to endocrine beta-cells. *Biochem Soc Trans, 36*(Pt 3), 353-356. doi:10.1042/BST0360353

Brennand, K., Huangfu, D., & Melton, D. (2007). All beta cells contribute equally to islet growth and maintenance. *PLoS Biol, 5*(7), e163. doi:10.1371/journal.pbio.0050163

Bruens, L., Ellenbroek, S. I. J., van Rheenen, J., & Snippert, H. J. (2017). In Vivo Imaging Reveals Existence of Crypt Fission and Fusion in Adult Mouse Intestine. *Gastroenterology, 153*(3), 674-677 e673. doi:10.1053/j.gastro.2017.05.019

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., . . . Holmes, I. H. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology, 17*(1), 66. doi:10.1186/s13059-016-0924-1

Butler, A. E., Cao-Minh, L., Galasso, R., Rizza, R. A., Corradin, A., Cobelli, C., & Butler, P. C. (2010). Adaptive changes in pancreatic beta cell fractional area and beta cell turnover in human pregnancy. *Diabetologia, 53*(10), 2167-2176. doi:10.1007/s00125-010-1809-6

Butler, A. E., Janson, J., Bonner-Weir, S., Ritzel, R., Rizza, R. A., & Butler, P. C. (2003). Beta-cell deficit and increased beta-cell apoptosis in humans with type 2 diabetes. *Diabetes, 52*(1), 102-110.

Cabrera, O., Berman, D. M., Kenyon, N. S., Ricordi, C., Berggren, P. O., & Caicedo, A. (2006). The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proc Natl Acad Sci U S A, 103*(7), 2334-2339. doi:10.1073/pnas.0510790103

Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O., & Stein, L. D. (2017). Pan-cancer analysis of whole genomes. *bioRxiv*.

Campbell-Thompson, M. L., Heiple, T., Montgomery, E., Zhang, L., & Schneider, L. (2012). Staining protocols for human pancreatic islets. *J Vis Exp*(63), e4068. doi:10.3791/4068

Castaing, M., Duvillie, B., Quemeneur, E., Basmaciogullari, A., & Scharfmann, R. (2005). Ex vivo analysis of acinar and endocrine cell development in the human embryonic pancreas. *Dev Dyn, 234*(2), 339-345. doi:10.1002/dvdy.20547

Castaing, M., Peault, B., Basmaciogullari, A., Casal, I., Czernichow, P., & Scharfmann, R. (2001). Blood glucose normalization upon transplantation of human embryonic pancreas into beta-cell-deficient SCID mice. *Diabetologia, 44*(11), 2066-2076. doi:10.1007/s001250100012

Chakravarthy, H., Gu, X., Enge, M., Dai, X., Wang, Y., Damond, N., . . . Kim, S. K. (2017). Converting Adult Pancreatic Islet alpha Cells into beta Cells by Targeting Both Dnmt1 and Arx. *Cell Metab, 25*(3), 622-634. doi:10.1016/j.cmet.2017.01.009

Chen, H., & Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics, 12*(1), 35. doi:10.1186/1471-2105-12-35

Cheng, H., & Bjerknes, M. (1985). Whole population cell kinetics and postnatal development of the mouse intestinal epithelium. *Anat Rec, 211*(4), 420-426. doi:10.1002/ar.1092110408

Clarke, R. M. (1972). The effect of growth and of fasting on the number of villi and crypts in the small intestine of the albino rat. *J Anat, 112*(Pt 1), 27-33.

Cnop, M., Hughes, S. J., Igoillo-Esteve, M., Hoppa, M. B., Sayyed, F., van de Laar, L., . . . Clark, A. (2010). The long lifespan and low turnover of human islet beta cells estimated by mathematical modelling of lipofuscin accumulation. *Diabetologia, 53*(2), 321-330. doi:10.1007/s00125-009-1562-x

Dahl, D. B. (2003). An improved merge-split sampler for conjugate Dirichlet process mixture models. *Univ. Wisconsin-Madison Tech. Rep., 1086*, 1–32

DeFronzo, R. A., & Tripathy, D. (2009). Skeletal muscle insulin resistance is the primary defect in type 2 diabetes. *Diabetes Care, 32 Suppl 2*, S157-163. doi:10.2337/dc09-S302

Dor, Y., Brown, J., Martinez, O. I., & Melton, D. A. (2004). Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation. *Nature, 429*(6987), 41-46. doi:10.1038/nature02520

El Ouaamari, A., Dirice, E., Gedeon, N., Hu, J., Zhou, J. Y., Shirakawa, J., . . . Kulkarni, R. N. (2016). SerpinB1 Promotes Pancreatic beta Cell Proliferation. *Cell Metab, 23*(1), 194-205. doi:10.1016/j.cmet.2015.12.001

Enge, M., Arda, H. E., Mignardi, M., Beausang, J., Bottino, R., Kim, S. K., & Quake, S. R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell, 171*(2), 321-330 e314. doi:10.1016/j.cell.2017.09.004

Escribano, O., Guillen, C., Nevado, C., Gomez-Hernandez, A., Kahn, C. R., & Benito, M. (2009). Beta-Cell hyperplasia induced by hepatic insulin resistance: role of a liver-pancreas endocrine axis through insulin receptor A isoform. *Diabetes, 58*(4), 820-828. doi:10.2337/db08-0551

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., . . . Campbell, P. J. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res, 45*(D1), D777-D783. doi:10.1093/nar/gkw1121

Fowler, M. J. (2008). Microvascular and Macrovascular Complications of Diabetes. *Clinical Diabetes, 26*(2), 77.

Gatfield, D., & Izaurralde, E. (2004). Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in Drosophila. *Nature, 429*(6991), 575-578. doi:10.1038/nature02559

Gregg, B. E., Moore, P. C., Demozay, D., Hall, B. A., Li, M., Husain, A., . . . Rhodes, C. J. (2012). Formation of a human beta-cell population within pancreatic islets is set early in life. *J Clin Endocrinol Metab, 97*(9), 3197-3206. doi:10.1210/jc.2012-1206

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L. B., Tubio, J. M. C., Papaemmanuil, E., . . . Bova, G. S. (2015). The evolutionary history of lethal metastatic prostate cancer. *Nature, 520*(7547), 353-357. doi:10.1038/nature14347

Guz, Y., Nasir, I., & Teitelman, G. (2001). Regeneration of pancreatic beta cells from intra-islet precursor cells in an experimental model of diabetes. *Endocrinology, 142*(11), 4956-4968. doi:10.1210/endo.142.11.8501

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell, 144*(5), 646-674. doi:10.1016/j.cell.2011.02.013

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97-109. doi:10.1093/biomet/57.1.97

Hines, K. E. (2015). A primer on Bayesian inference for biophysical systems. *Biophys J, 108*(9), 2103-2113. doi:10.1016/j.bpj.2015.03.042

Hoang, M. L., Kinde, I., Tomasetti, C., McMahon, K. W., Rosenquist, T. A., Grollman, A. P., . . . Papadopoulos, N. (2016). Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A, 113*(35), 9846-9851. doi:10.1073/pnas.1607794113

Horiguchi, S., & Kamisawa, T. (2010). Major duodenal papilla and its normal anatomy. *Dig Surg, 27*(2), 90-93. doi:10.1159/000288841

Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., . . . Kumar, R. (2013). TERT promoter mutations in familial and sporadic melanoma. *Science, 339*(6122), 959-961. doi:10.1126/science.1230062

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., & Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science, 339*(6122), 957-959. doi:10.1126/science.1229259

Ionescu-Tirgoviste, C., Gagniuc, P. A., Gubceac, E., Mardare, L., Popescu, I., Dima, S., & Militaru, M. (2015). A 3D map of the islet routes throughout the healthy human pancreas. *Sci Rep, 5*, 14634. doi:10.1038/srep14634

Jacovetti, C., Matkovich, S. J., Rodriguez-Trejo, A., Guay, C., & Regazzi, R. (2015). Postnatal beta-cell maturation is associated with islet-specific microRNA changes induced by nutrient shifts at weaning. *Nat Commun, 6*, 8084. doi:10.1038/ncomms9084

Jager, M., Blokzijl, F., Sasselli, V., Boymans, S., Janssen, R., Besselink, N., . . . Cuppen, E. (2018). Measuring mutation accumulation in single human adult stem cells by whole-genome sequencing of organoid cultures. *Nat Protoc, 13*(1), 59-78. doi:10.1038/nprot.2017.111

Jennings, R. E., Berry, A. A., Kirkwood-Wilson, R., Roberts, N. A., Hearn, T., Salisbury, R. J., . . . Hanley, N. A. (2013). Development of the human pancreas from foregut to endocrine commitment. *Diabetes, 62*(10), 3514-3522. doi:10.2337/db12-1479

Jeon, J., Correa-Medina, M., Ricordi, C., Edlund, H., & Diez, J. A. (2009). Endocrine cell clustering during human pancreas development. *J Histochem Cytochem, 57*(9), 811-824. doi:10.1369/jhc.2009.953307

Jo, J., Kilimnik, G., Kim, A., Guo, C., Periwal, V., & Hara, M. (2011). Formation of pancreatic islets involves coordinated expansion of small islets and fission of large interconnected islet-like structures. *Biophys J, 101*(3), 565-574. doi:10.1016/j.bpj.2011.06.042

Jones, D., Raine, K. M., Davies, H., Tarpey, P. S., Butler, A. P., Teague, J. W., . . . Campbell, P. J. (2016). cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics, 56*, 15 10 11-15 10 18. doi:10.1002/cpbi.20

Ju, Y. S., Martincorena, I., Gerstung, M., Petljak, M., Alexandrov, L. B., Rahbari, R., . . . Stratton, M. R. (2017). Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature, 543*(7647), 714-718. doi:10.1038/nature21703

Kim, H. S., & Lee, M. K. (2016). beta-Cell regeneration through the transdifferentiation of pancreatic cells: Pancreatic progenitor cells in the pancreas. *J Diabetes Investig, 7*(3), 286-296. doi:10.1111/jdi.12475

Kodama, S., Kuhtreiber, W., Fujimura, S., Dale, E. A., & Faustman, D. L. (2003). Islet regeneration during the reversal of autoimmune diabetes in NOD mice. *Science, 302*(5648), 1223-1227. doi:10.1126/science.1088949

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Lin, F., Chen, Z. E., & Wang, H. L. (2015). Utility of immunohistochemistry in the pancreatobiliary tract. *Arch Pathol Lab Med, 139*(1), 24-38. doi:10.5858/arpa.2014-0072-RA

Lipsett, M., & Finegood, D. T. (2002). beta-cell neogenesis during prolonged hyperglycemia in rats. *Diabetes, 51*(6), 1834-1841.

Lyttle, B. M., Li, J., Krishnamurthy, M., Fellows, F., Wheeler, M. B., Goodyer, C. G., & Wang, R. (2008). Transcription factor expression in the developing human fetal endocrine pancreas. *Diabetologia, 51*(7), 1169-1180. doi:10.1007/s00125-008-1006-z

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., . . . Voet, T. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods, 12*(6), 519-522. doi:10.1038/nmeth.3370

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., . . . Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell, 171*(5), 1029-1041 e1021. doi:10.1016/j.cell.2017.09.042

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., . . . Campbell, P. J. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science, 348*(6237), 880-886. doi:10.1126/science.aaa6806

Menge, B. A., Tannapfel, A., Belyaev, O., Drescher, R., Muller, C., Uhl, W., . . . Meier, J. J. (2008). Partial pancreatectomy in adult humans does not provoke beta-cell regeneration. *Diabetes, 57*(1), 142-149. doi:10.2337/db07-1294

Messier, B., & Leblond, C. P. (1960). Cell proliferation and migration as revealed by radioautography after injection of thymidine-H3 into male rats and mice. *Am J Anat, 106*, 247-285. doi:10.1002/aja.1001060305

Mezza, T., Muscogiuri, G., Sorice, G. P., Clemente, G., Hu, J., Pontecorvi, A., . . . Kulkarni, R. N. (2014). Insulin resistance alters islet morphology in nondiabetic humans. *Diabetes, 63*(3), 994-1007. doi:10.2337/db13-1013

Muraro, M. J., Dharmadhikari, G., Grun, D., Groen, N., Dielen, T., Jansen, E., . . . van Oudenaarden, A. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst, 3*(4), 385-394 e383. doi:10.1016/j.cels.2016.09.002

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., . . . Breast Cancer Working Group of the International Cancer Genome, C. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell, 149*(5), 979-993. doi:10.1016/j.cell.2012.04.024

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., . . . Breast Cancer Working Group of the International Cancer Genome, C. (2012). The life history of 21 breast cancers. *Cell, 149*(5), 994-1007. doi:10.1016/j.cell.2012.04.023

Papastamoulis, P. (2016). label.switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs. *J. Stat. Softw., 69*, 24. doi:10.18637/jss.v069.c01

Papastamoulis, P., & Iliopoulos, G. (2010). An Artificial Allocations Based Solution to the Label Switching Problem in Bayesian Analysis of Mixtures of Distributions. *Journal of Computational and Graphical Statistics, 19*(2), 313-331.

Perl, S., Kushner, J. A., Buchholz, B. A., Meeker, A. K., Stein, G. M., Hsieh, M., . . . Tisdale, J. F. (2010). Significant human beta-cell turnover is limited to the first three decades of life as determined by in vivo thymidine analog incorporation and radiocarbon dating. *J Clin Endocrinol Metab, 95*(10), E234-239. doi:10.1210/jc.2010-0932

Piper, K., Brickwood, S., Turnpenny, L. W., Cameron, I. T., Ball, S. G., Wilson, D. I., & Hanley, N. A. (2004). Beta cell differentiation during early human pancreas development. *J Endocrinol, 181*(1), 11-23.

Pusztaszeri, M. P., Seelentag, W., & Bosman, F. T. (2006). Immunohistochemical expression of endothelial markers CD31, CD34, von Willebrand factor, and Fli-1 in normal human tissues. *J Histochem Cytochem, 54*(4), 385-395. doi:10.1369/jhc.4A6514.2005

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.0): R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Raine, K. M., Van Loo, P., Wedge, D. C., Jones, D., Menzies, A., Butler, A. P., . . . Campbell, P. J. (2016). ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr Protoc Bioinformatics, 56*, 15 19 11-15 19 17. doi:10.1002/cpbi.17

Rall, L. B., Pictet, R. L., Williams, R. H., & Rutter, W. J. (1973). Early differentiation of glucagon-producing cells in embryonic pancreas: a possible developmental role for glucagon. *Proc Natl Acad Sci U S A, 70*(12), 3478-3482.

Rieck, S., & Kaestner, K. H. (2010). Expansion of beta-cell mass in response to pregnancy. *Trends Endocrinol Metab, 21*(3), 151-158. doi:10.1016/j.tem.2009.11.001

Riedel, M. J., Asadi, A., Wang, R., Ao, Z., Warnock, G. L., & Kieffer, T. J. (2012). Immunohistochemical characterisation of cells co-producing insulin and glucagon in the developing human pancreas. *Diabetologia, 55*(2), 372-381. doi:10.1007/s00125-011-2344-9

Roberts, N. D. (2018). *Patterns of somatic genome rearrangement in human cancer (Doctoral Thesis).* (Doctoral Thesis), University of Cambridge,

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., & Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol, 17*, 31. doi:10.1186/s13059-016-0893-4

RStudio Team. (2016). RStudio: Integrated Development for R (Version 1.1.453). Boston, MA, USA: RStudio, Inc. Retrieved from http://www.rstudio.com/

Sarkar, S. A., Kobberup, S., Wong, R., Lopez, A. D., Quayum, N., Still, T., . . . Hutton, J. C. (2008). Global gene expression profiling and histochemical analysis of the developing human fetal pancreas. *Diabetologia, 51*(2), 285-297. doi:10.1007/s00125-007-0880-0

Seymour, P. A., Bennett, W. R., & Slack, J. M. (2004). Fission of pancreatic islets during postnatal growth of the mouse. *J Anat, 204*(2), 103-116. doi:10.1111/j.1469-7580.2004.00265.x

Snippert, H. J., van der Flier, L. G., Sato, T., van Es, J. H., van den Born, M., Kroon-Veenboer, C., . . . Clevers, H. (2010). Intestinal crypt homeostasis results from

neutral competition between symmetrically dividing Lgr5 stem cells. *Cell, 143*(1), 134-144. doi:10.1016/j.cell.2010.09.016

Socorro, M., Criscimanna, A., Riva, P., Tandon, M., Prasadan, K., Guo, P., . . . Esni, F. (2017). Identification of Newly Committed Pancreatic Cells in the Adult Mouse Pancreas. *Sci Rep, 7*(1), 17539. doi:10.1038/s41598-017-17884-z

Sorenson, R. L., & Brelje, T. C. (1997). Adaptation of islets of Langerhans to pregnancy: beta-cell growth, enhanced insulin secretion and the role of lactogenic hormones. *Horm Metab Res, 29*(6), 301-307. doi:10.1055/s-2007-979040

Spijker, H. S., Song, H., Ellenbroek, J. H., Roefs, M. M., Engelse, M. A., Bos, E., . . . de Koning, E. J. (2015). Loss of beta-Cell Identity Occurs in Type 2 Diabetes and Is Associated With Islet Amyloid Deposits. *Diabetes, 64*(8), 2928-2938. doi:10.2337/db14-1752

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature, 458*(7239), 719-724. doi:10.1038/nature07943

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association, 101*(476), 1566-1581. doi:10.1198/016214506000000302

Tischler, G., & Leonard, S. (2014). biobambam: tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine, 9*, 13-13. doi:10.1186/1751-0473-9-13

Van Assche, F. A., Aerts, L., & De Prins, F. (1978). A morphological study of the endocrine pancreas in human pregnancy. *Br J Obstet Gynaecol, 85*(11), 818-820.

van der Meulen, T., Mawla, A. M., DiGruccio, M. R., Adams, M. W., Nies, V., Dolleman, S., . . . Huising, M. O. (2017). Virgin Beta Cells Persist throughout Life at a Neogenic Niche within Pancreatic Islets. *Cell Metab, 25*(4), 911-926 e916. doi:10.1016/j.cmet.2017.03.017

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., . . . Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A, 107*(39), 16910-16915. doi:10.1073/pnas.1009843107

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R. C., Liaw, W. H. A., Lumley, T., . . . Venables, B. (2015). gplots: Various R Programming Tools for Plotting

Data (Version 3.0.1). Retrieved from https://CRAN.R-project.org/package=gplots

Wen, X. Y., Hegele, R. A., Wang, J., Wang, D. Y., Cheung, J., Wilson, M., . . . Stewart, A. K. (2003). Identification of a novel lipase gene mutated in lpd mice with hypertriglyceridemia and associated with dyslipidemia in humans. *Hum Mol Genet, 12*(10), 1131-1143.

World Health Organization. (2016). *Global report on diabetes.* Retrieved from http://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=5C2AFA5E158FB3E62F43519100F63B75?sequence=1

Yee, T. W. (2015). *Vector generalized linear and additive models: with an implementation in R*. New York, NY, USA: Springer.

Yu, G., Smith David, K., Zhu, H., Guan, Y., & Lam Tommy, T.-Y. (2016). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution, 8*(1), 28-36. doi:10.1111/2041-210X.12628

Zheng, Y., Ley, S. H., & Hu, F. B. (2018). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol, 14*(2), 88-98. doi:10.1038/nrendo.2017.151

Zou, L., & Elledge, S. J. (2003). Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. *Science, 300*(5625), 1542-1548. doi:10.1126/science.1083430

Zulewski, H., Abraham, E. J., Gerlach, M. J., Daniel, P. B., Moritz, W., Muller, B., . . . Habener, J. F. (2001). Multipotential nestin-positive stem cells isolated from adult pancreatic islets differentiate ex vivo into pancreatic endocrine, exocrine, and hepatic phenotypes. *Diabetes, 50*(3), 521-533.

8. Appendices

8.1 Variant clustering with the n-dimensional hierarchical Dirichlet process

To cluster the VAFs across multiple samples from the same patient, an n-dimensional hierarchical Dirichlet process was employed. This algorithm was written and designed by Peter Campbell. Much like the Bayesian Dirichlet process used previously, this method requires the number of variant reads per sample as well as the sequencing depth (Nik-Zainal, Van Loo, et al., 2012). There are no prior definitions of how many mutations are in each cluster, and each mutation can be allocated to any cluster. An upper limit of 100 clusters was set; however the number of clusters generated here was far less than that. Diploid cells are assumed. Each cluster takes on a location within an n-dimensional VAF hypercube but the exact location of this is unknown until the completion of the process. Both the distribution of clone sizes and number of variants per clone are modelled as a Dirichlet process, in a hierarchical Bayesian model with the variant reads and total sequencing depth.

The total read depth for mutation *i* in sample *j* is defined as: $n_{i,j}, i = 1, \dots, N, j = 1, \dots, M$ where *N* is the number of somatic mutations across all *M* samples. The reference allele, $y_{i,j}$, is similarly defined using the number of reads supporting the reference sequence. The distribution of the reference allele approximates a binomial distribution of the mutation total read depth and the expected proportion of reads supporting the reference allele ($\pi_{i,j}$). As such, $y_{i,j} \sim \mathrm{Bin}(n_{i,j}, \pi_{i,j})$ where $\pi_{i,j}$ is a Dirichlet process:

$$\pi_i \sim DP(\alpha P_0) \in [0,1]^M$$

P is defined as:

$$P = \sum_{h=1}^{\infty} \omega_h \delta_{\pi_h}$$

In this case, $\pi_h \sim P_0$, where δ_π is the point mass at π and ω_h is the weight of the h^{th} mutation cluster. As this is a stick-breaking representation of the Dirichlet process, ω_h is defined as:

$$\omega_h = V_h \Pi_{l<h}(1 - V_l)$$

The beta distribution is then used to estimate V_h as $V_h \sim Beta(1 - \alpha)$. As priors, $P_0 \sim U(0,1)^M$ and $\alpha \sim \Gamma(0.01, 0.01)$. The posterior distribution of the Dirichlet process is

then modelled using the Gibbs sampler. This involves the sequential sampling of each parameter in the joint distribution and the extraction of a univariate conditional distribution, based on the previous sampling of all other parameters (Hines, 2015). In this case, each mutation is assigned a cluster and stick-breaking weights are adjusted based on the conditional conjugate beta posterior distributions. Draws from the posterior distribution of $(\pi_h|-)$ are then used to update the cluster positions in the *M*-dimensional VAF hypercube.

In large polyclonal samples, these clusters tend to be on the edges of the VAF hypercube and as such, a large region of low probability becomes apparent in the posterior distribution. This aspect of the posterior distribution can then go unsampled due to the low probability and the Gibbs sampling can then be limited. To counter this, a merge-split step is performed after each iteration of the Gibbs sampler using the Metropolis-Hastings proposal for conjugate distributions (Dahl, 2003). Briefly, this involves the random sampling of two mutations and if they are in different clusters, merging the two variants is considered based on the beta-binomial distribution of mutations already allocated. Should the two random variants be in the same clusters, a split step is then considered in a similar fashion (Dahl, 2003). Each merge-split option produces a Metropolis-Hastings ratio and the split or merge is accepted with this probability (Hastings, 1970). The posterior distribution for α can then be refined using this clustering.

The Gibbs sampler was run for 15,000 iterations and the first 10,000 were discarded. The R package label.switching (v1.7, https://CRAN.R-project.org/package=label.switching) (Papastamoulis, 2016) was integrated into the process in order to resolve the label switching problem associated with Markov Chain Monte Carlo outputs, through the use of an Equivalence Classes Representatives (ECR) algorithm (Papastamoulis & Iliopoulos, 2010).

8.2 Copy number analysis revealed no detectable gains or losses across samples

The mean ploidy was 1.97 with ASCAT and 2 with Battenberg (Nik-Zainal, Van Loo, et al., 2012; Raine et al., 2016; Van Loo et al., 2010). This fits with a normal sample containing no significant copy number changes. The ASCAT and Battenberg plot for sample PD37726d_lo0055 can be seen in Figure S1.

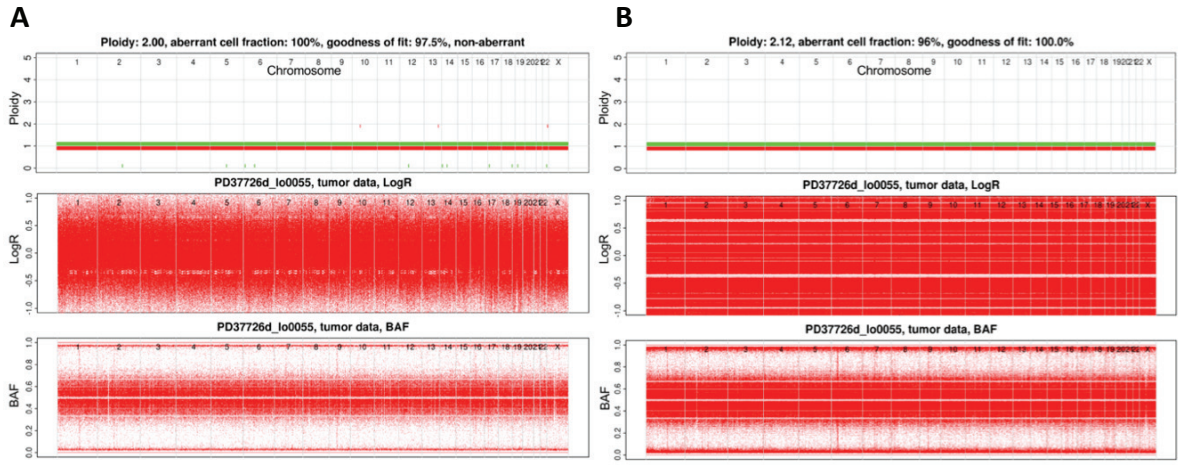


Figure S1 - The copy number analysis of sample PD37726d_lo0055

The x-axes in all six plots is the genomic position, split up by chromosome number. LogR is the log of the coverage at each variant site, compared to the reference sample. The BAF is the B-allele frequency with the B-allele being the mutant variant.

(A) The plots produced by ASCAT shows a clear ploidy of 2, with very few deviations. The top plot summarises this with the few individual red and green marks above and below the value y=1. The LogR and BAF plots show no significant changes in allele frequency or deviations in coverage.

(B) The plot produced by Battenberg supports the ASCAT result, with a ploidy of 2.12 and a goodness of it equal to 100%. The top plot summarises this with the few individual red and green marks above and below the value y=1. The LogR and BAF plots show no significant changes in allele frequency or deviations in coverage.

8.3 Additional support for the mutation calling filters

To validate these three filters, the variants that were filtered out in the unmatched data were analysed. These variants include those with a mean VAF≥0.4, a low over-dispersion parameter (ρ-value) or poor coverage at the variant site. This does not include any variants filtered out by manual inspection of the variants. Per sample, the range of mutations removed was 92,603 to 94,435 and the mean number of removed variants per sample was 94,204 (Figure S2).

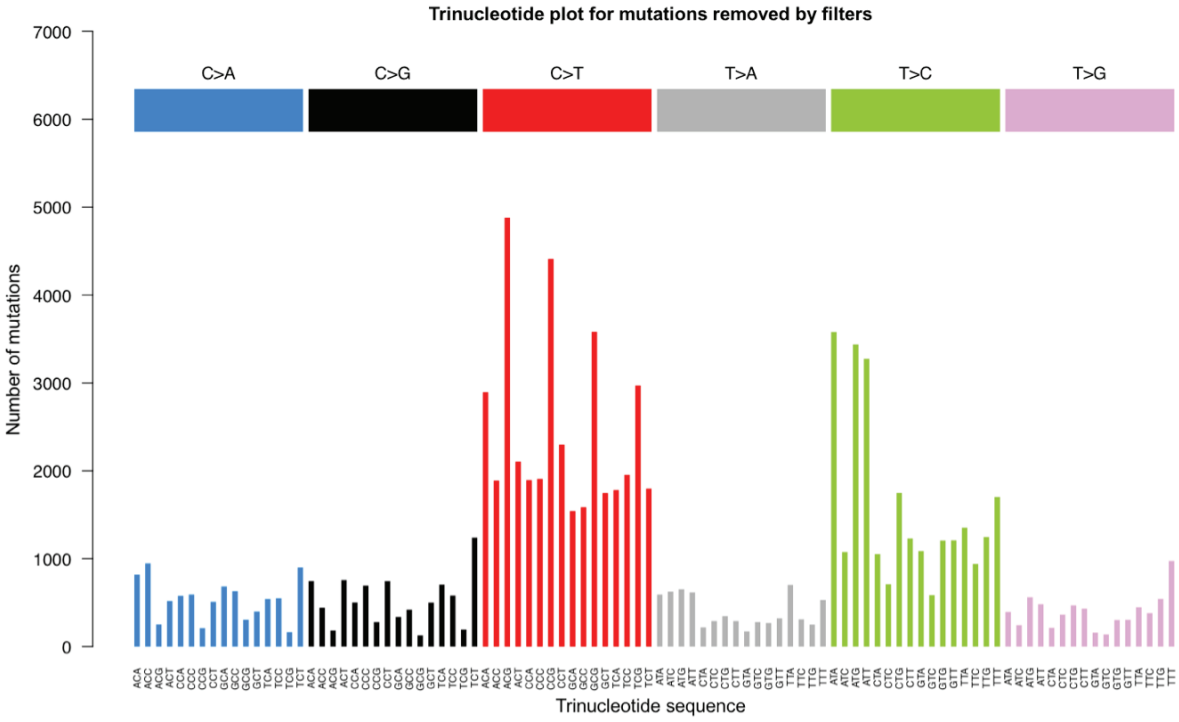


Figure S2 – The entire set of variants filtered out from the unmatched data
The 96-trinucleotide bar plot shows the unique variants removed from the unmatched data. There is a clear dominance of C>T mutations, followed by T>C mutations.

Many of the mutations included in this filtered out cohort would be expected to be germline SNVs. These variants take on high VAFs and are present throughout all samples. Artefacts however would be expected to have a lower VAF and be less common amongst the samples. The efficient removal of artefacts, through the use of the filters described, appears to be supported when subsetting these mutations to those with a VAF less than 0.1 (Figure S3).

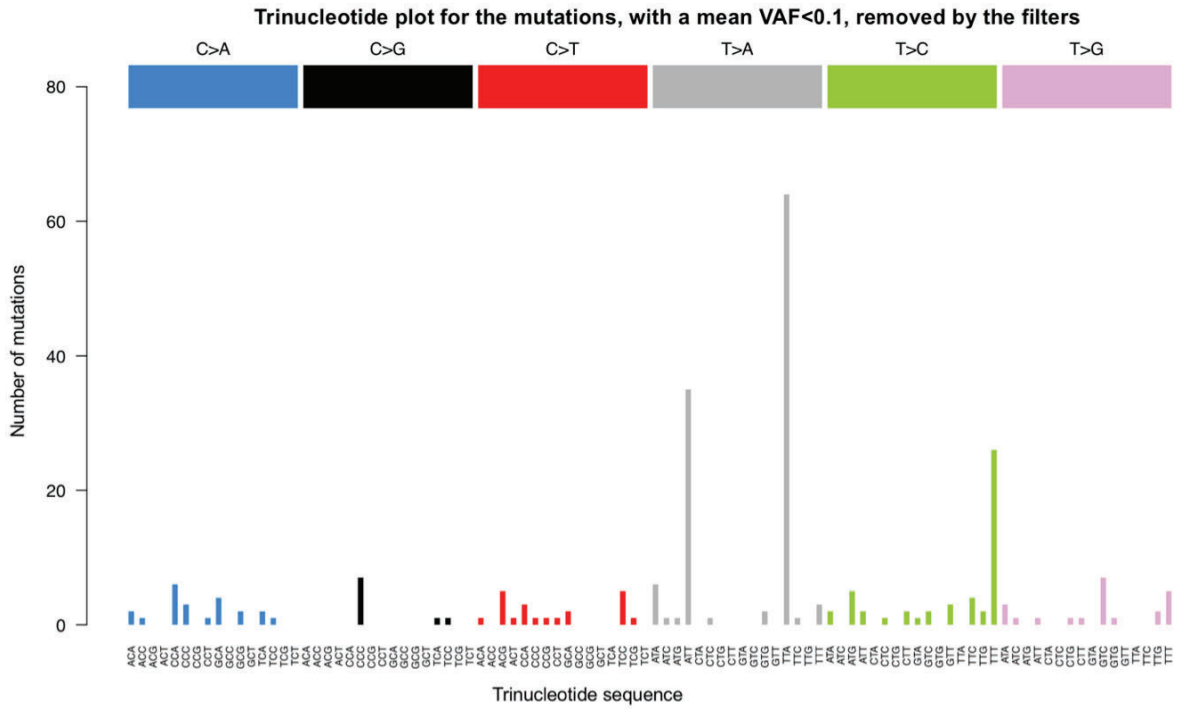


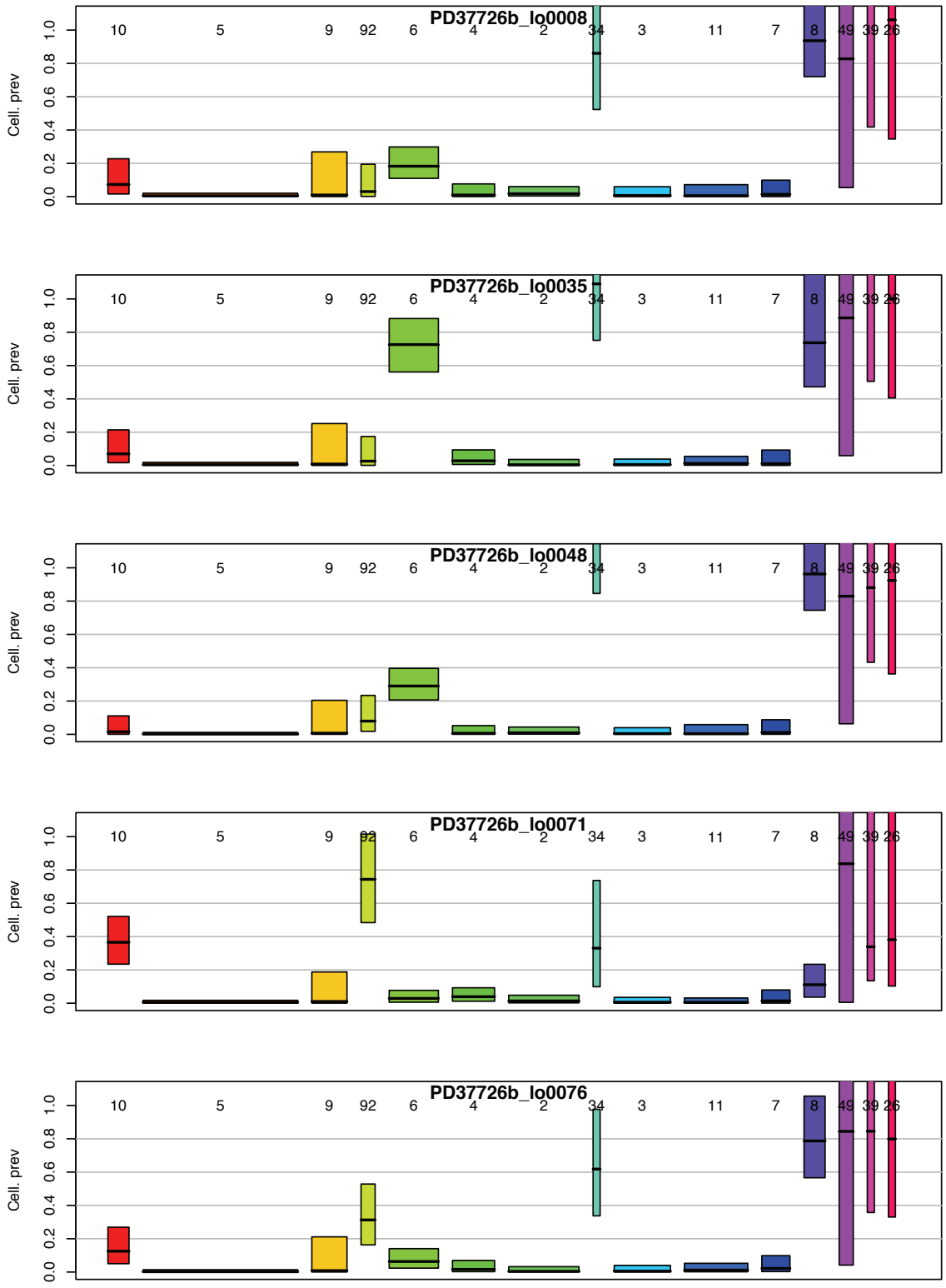
Figure S3 – The low VAF variants that were filtered out

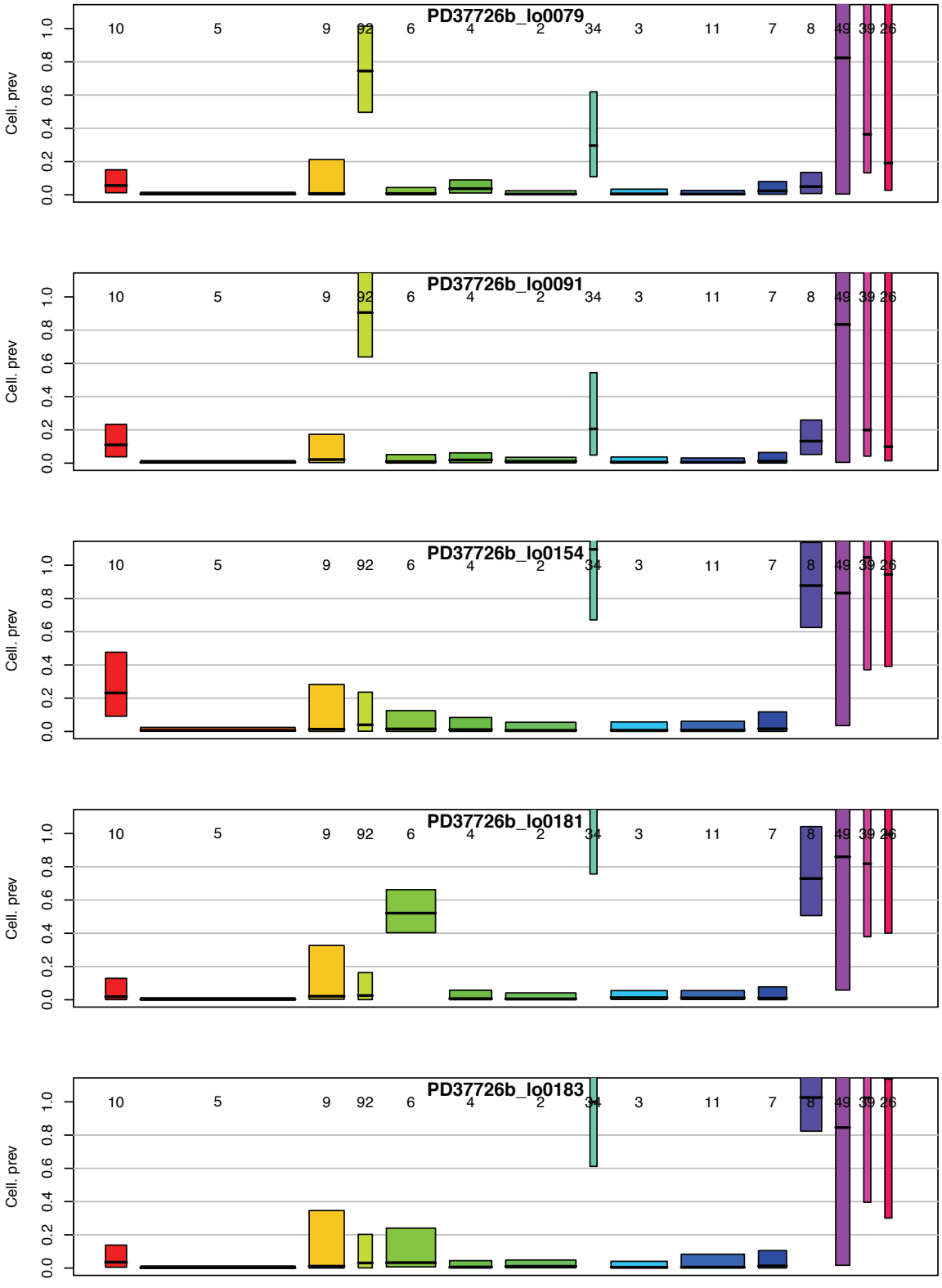
From the 94,548 variants filtered out of the unmatched data, 497 had a mean VAF<0.1. The majority of these are T>A mutations and resemble known sequencing artefacts.

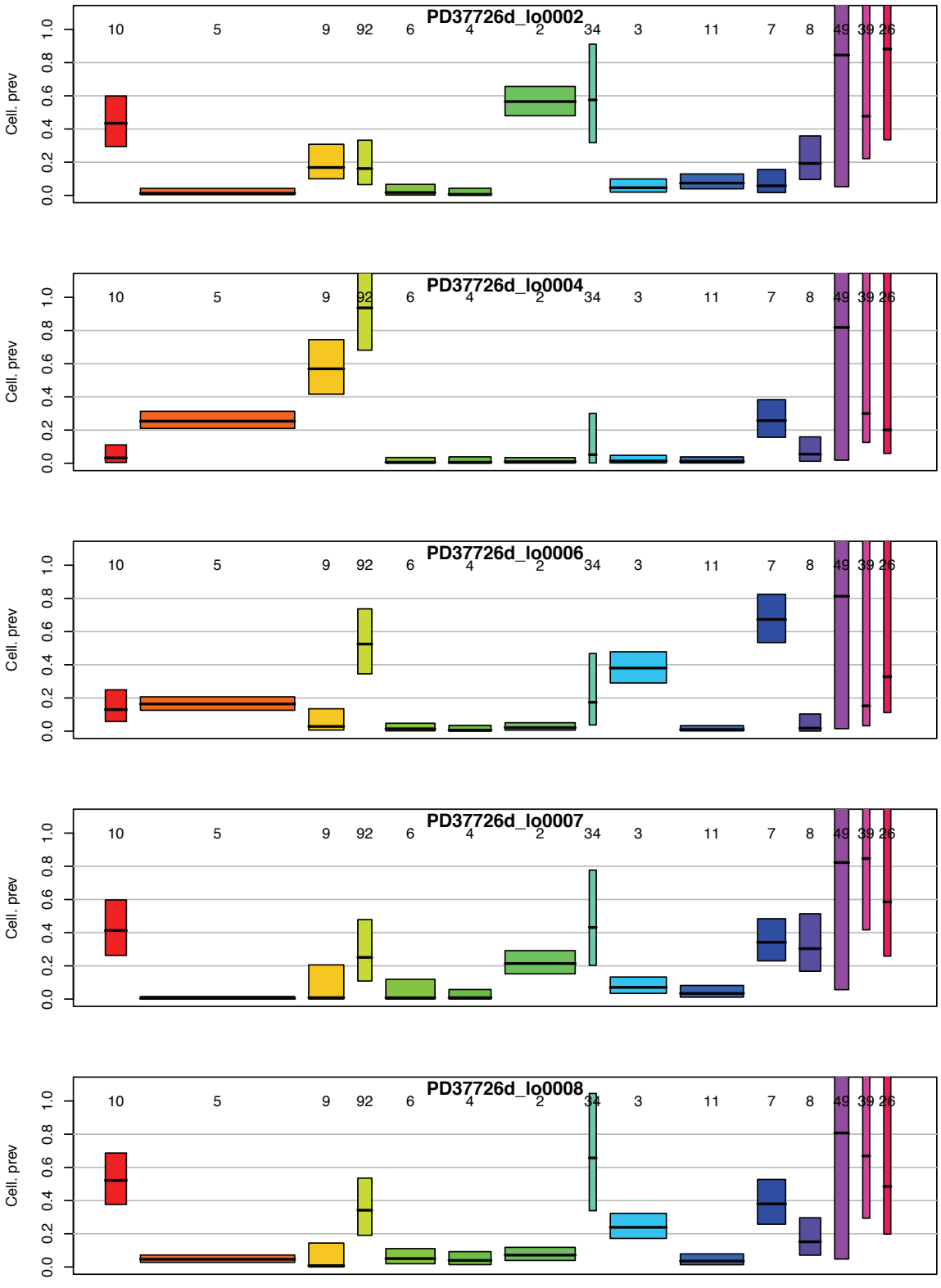
The 497 variants with a VAF less than 0.1 are made up mostly of T>A mutations, a transversion likely generated by the fragmentase enzyme mix (New England Biolabs), during whole-genome library preparation. Comparing Figure S3 to Figures 11 and 12, there is a clear removal of these T>A mutations through the use of the beta-binomial distribution and a depth filter. It appears therefore, that these two filters have proved useful in reducing sequencing artefacts.

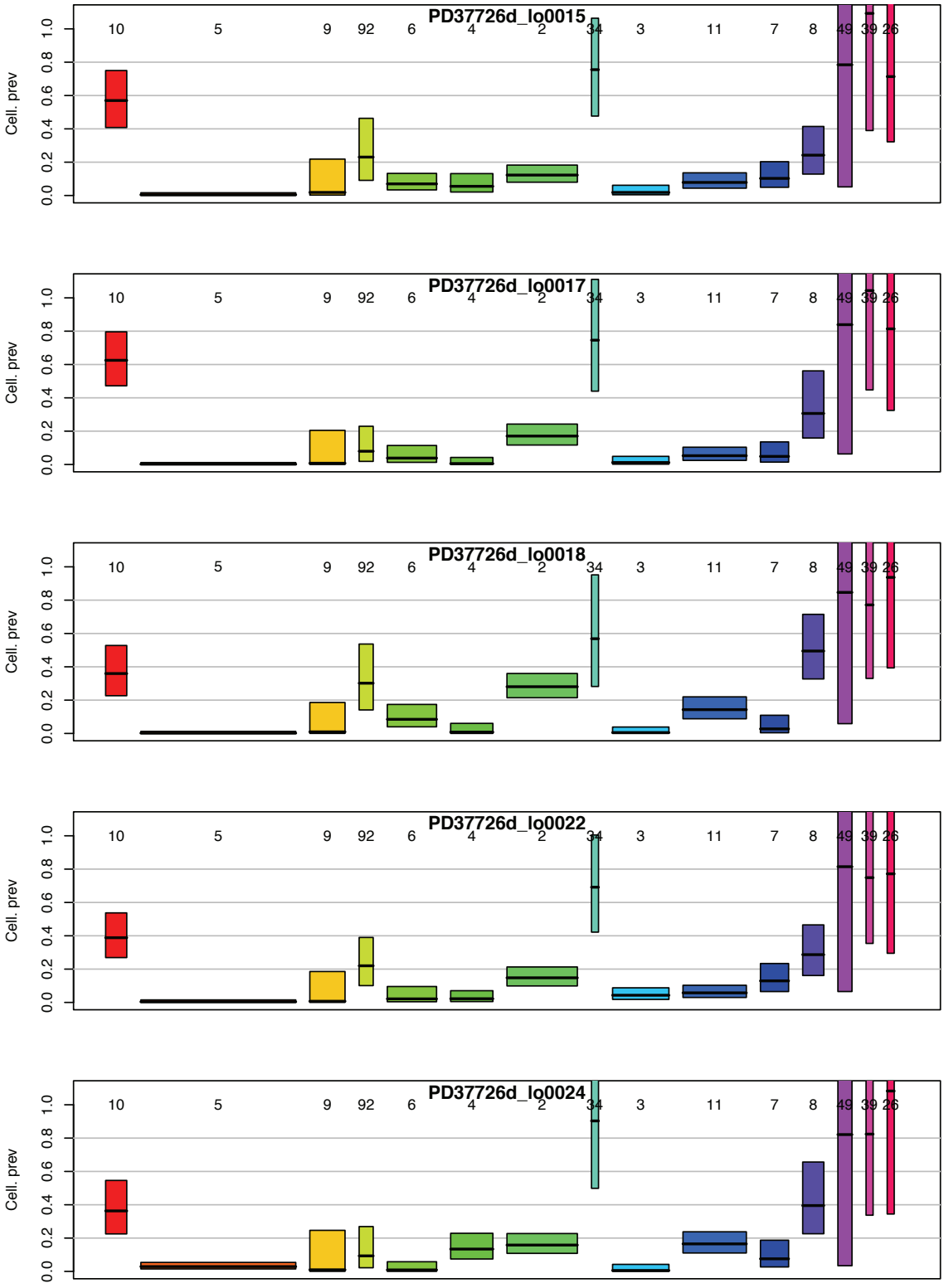
8.4 The pigeonhole principle was applied to each of the 42 samples to reconstruct the phylogenetic tree

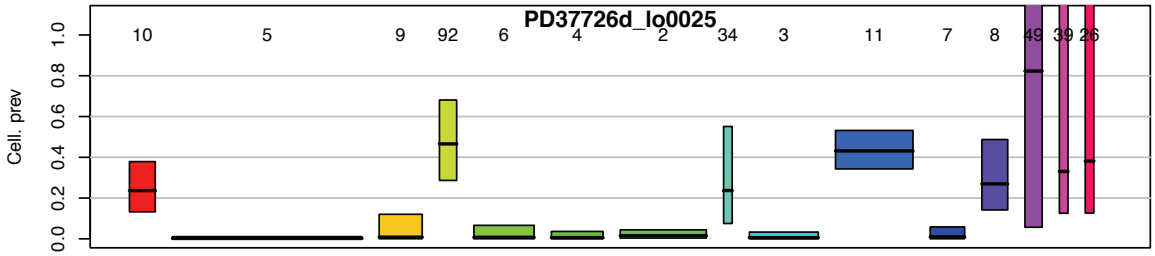
The boxes that were generated from the n-HDP clustering algorithm for all 42 samples are shown in Figure S4. These figures were generated in collaboration with Federico Abascal.

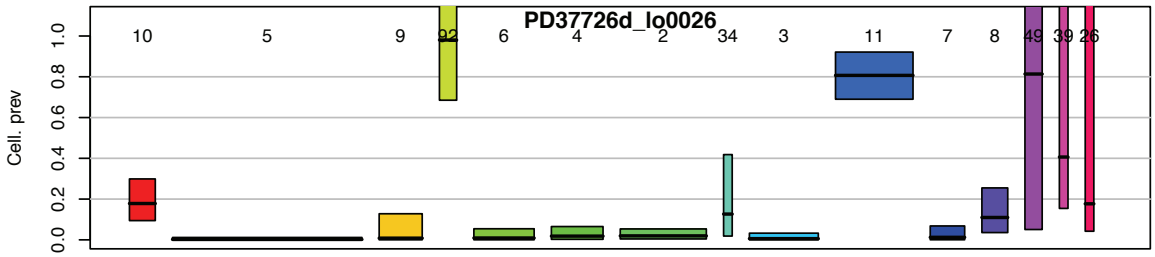

PD37726b_lo0008
PD37726b_lo0035
PD37726b_lo0048
PD37726b_lo0071
PD37726b_lo0076
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

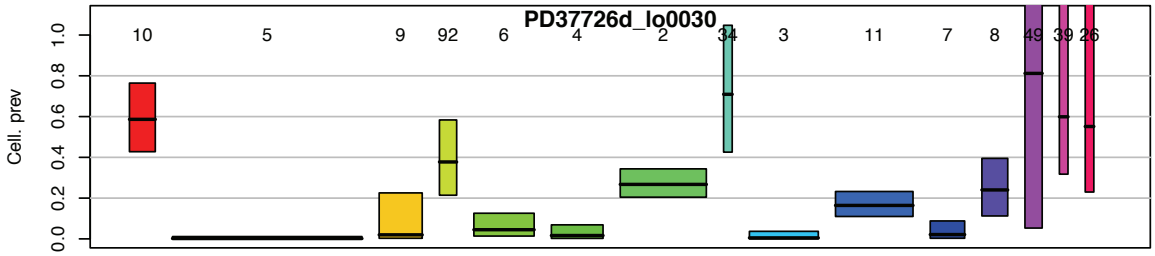

PD37726b_lo0079
PD37726b_lo0091
PD37726b_lo0154
PD37726b_lo0181
PD37726b_lo0183
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

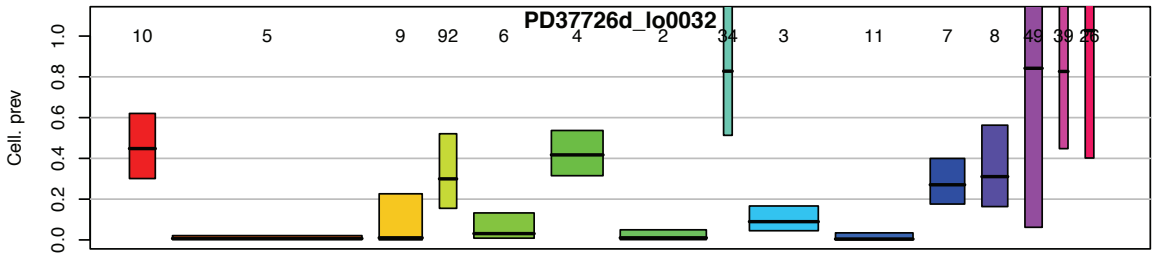

PD37726d_lo0002
PD37726d_lo0004
PD37726d_lo0006
PD37726d_lo0007
PD37726d_lo0008
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

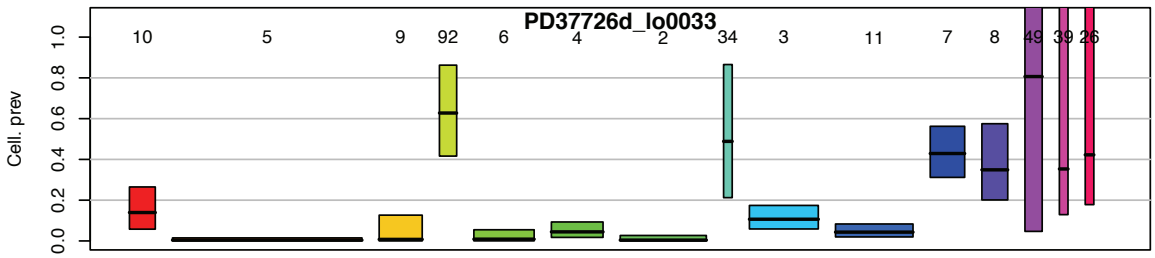

PD37726d_lo0015
PD37726d_lo0017
PD37726d_lo0018
PD37726d_lo0022
PD37726d_lo0024
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

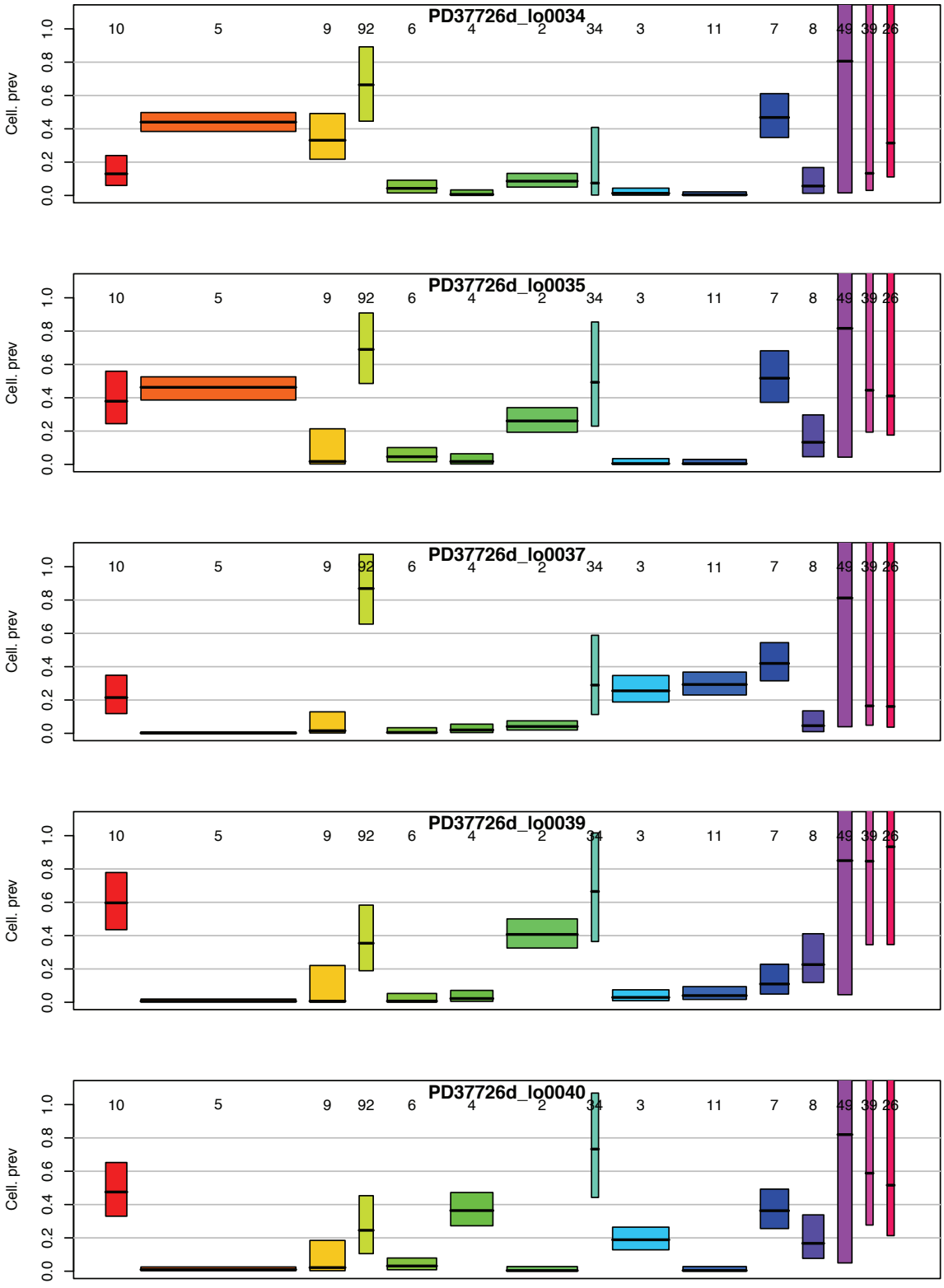

PD37726d_lo0025
Cell. prev
0.0 0.2 0.4 0.6 0.8 1.0
10 5 9 92 6 4 2 34 3 11 7 8 49 39 26

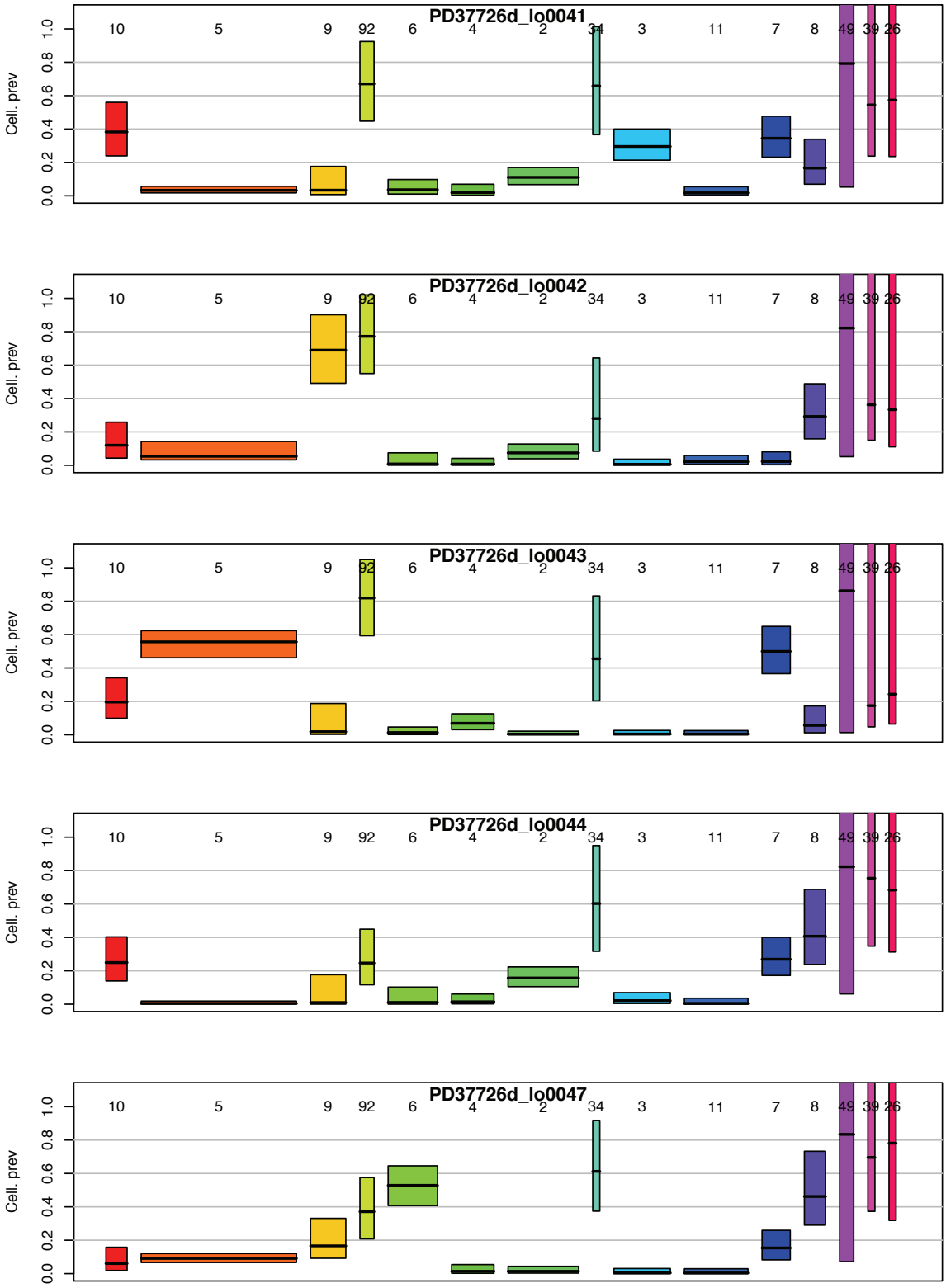

PD37726d_lo0026
Cell. prev
0.0 0.2 0.4 0.6 0.8 1.0
10 5 9 92 6 4 2 34 3 11 7 8 49 39 26

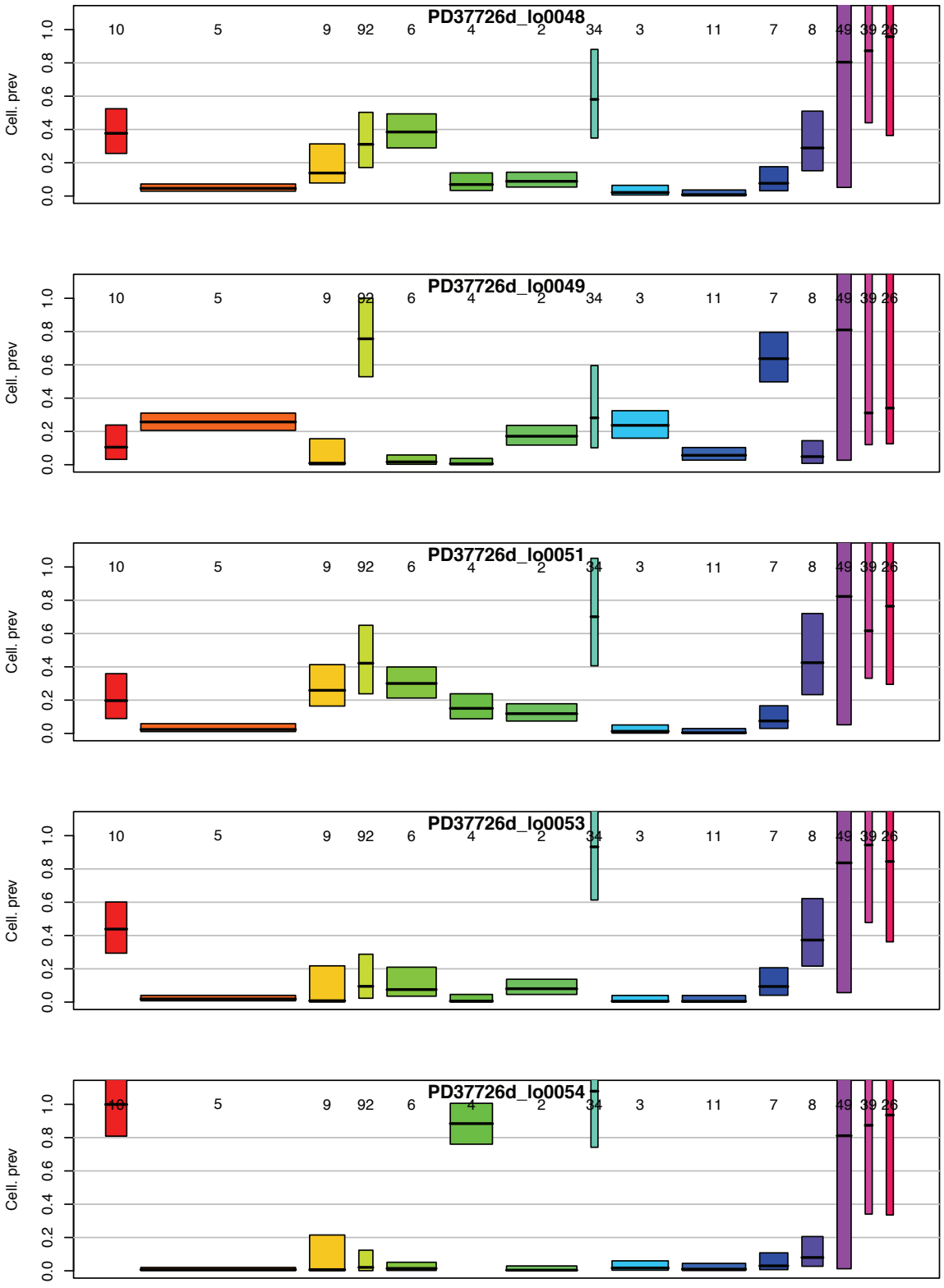

PD37726d_lo0030
Cell. prev
0.0 0.2 0.4 0.6 0.8 1.0
10 5 9 92 6 4 2 34 3 11 7 8 49 39 26


PD37726d_lo0032
Cell. prev
0.0 0.2 0.4 0.6 0.8 1.0
10 5 9 92 6 4 2 34 3 11 7 8 49 39 26


PD37726d_lo0033
Cell. prev
0.0 0.2 0.4 0.6 0.8 1.0
10 5 9 92 6 4 2 34 3 11 7 8 49 39 26


PD37726d_lo0034
PD37726d_lo0035
PD37726d_lo0037
PD37726d_lo0039
PD37726d_lo0040
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26


PD37726d_lo0041
PD37726d_lo0042
PD37726d_lo0043
PD37726d_lo0044
PD37726d_lo0047
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26


PD37726d_lo0048
PD37726d_lo0049
PD37726d_lo0051
PD37726d_lo0053
PD37726d_lo0054
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

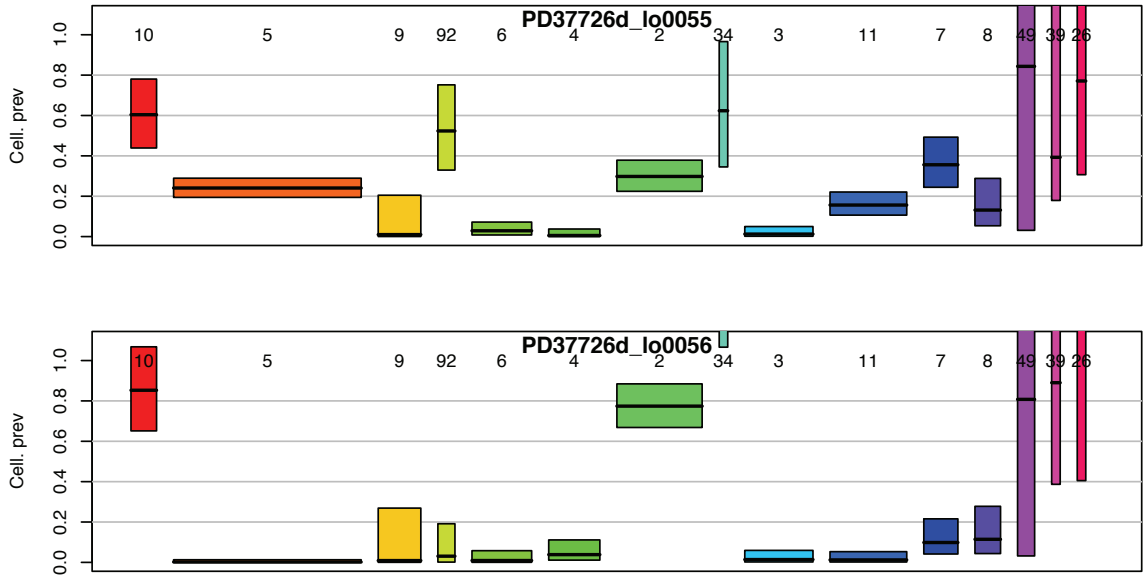


Figure S4 – All 42 boxes used in the phylogenetic tree reconstruction

These boxes represent all 32 islets and ten bladder urothelium samples that underwent clustering with n-HDP. Per sample, clusters (x-axis) and the cell fraction per cluster (y-axis) are shown. The cell fraction is equal to double the VAF. The sample name is at the top with the prefix “PD37726b” representing a bladder urothelium sample and “PD37726d” representing an islet sample. Box width is proportional to the number of mutations while the length is the 95% credible interval. Cluster 49 was discarded as it appeared to be present in all cells.

Each of these boxplots was used to reconstruct the phylogenetic tree, placing the 32 islets and the ten bladder urothelial samples onto the tree. This revealed a shared MRCA and a missing split accounting for three bladder samples (Figures 30 and 31).