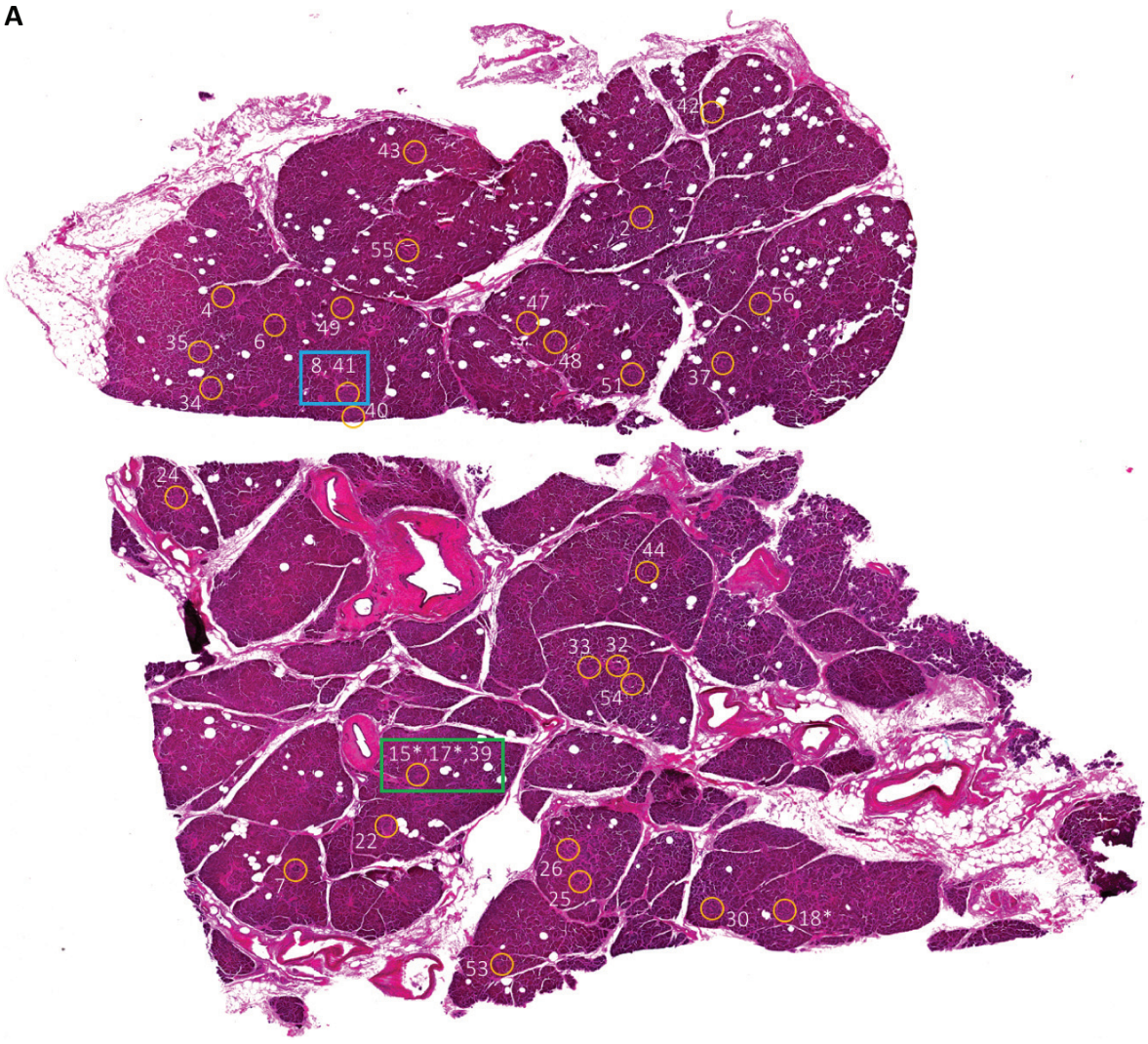
4. Results

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



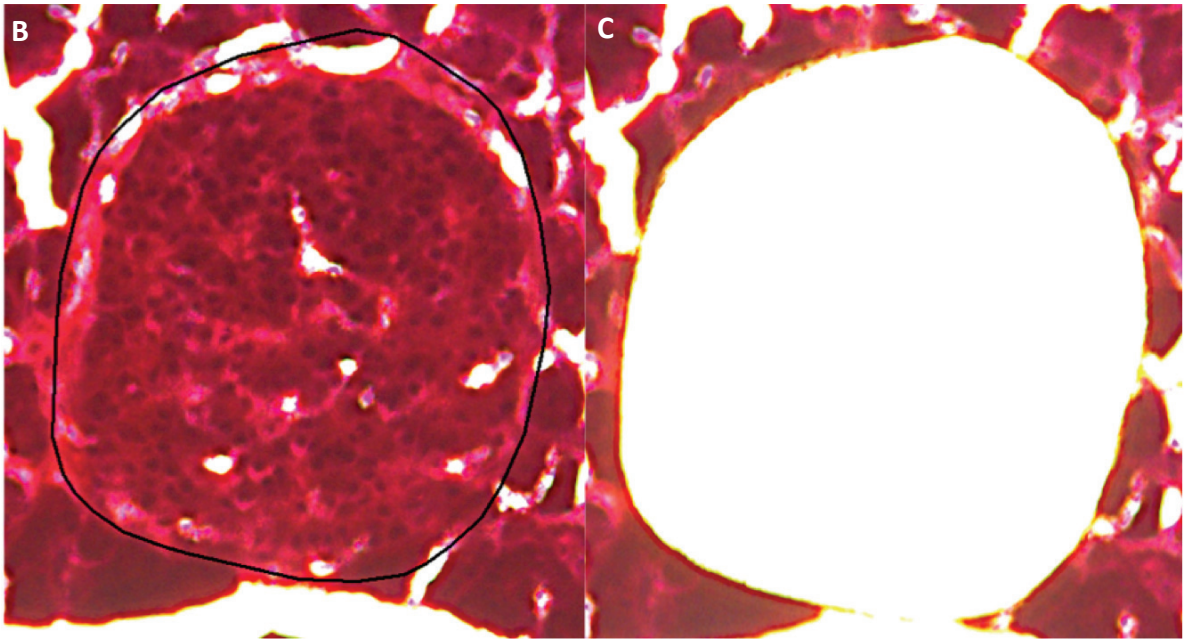


Figure 9 - Images from the LCM of the pancreatic islets

(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.

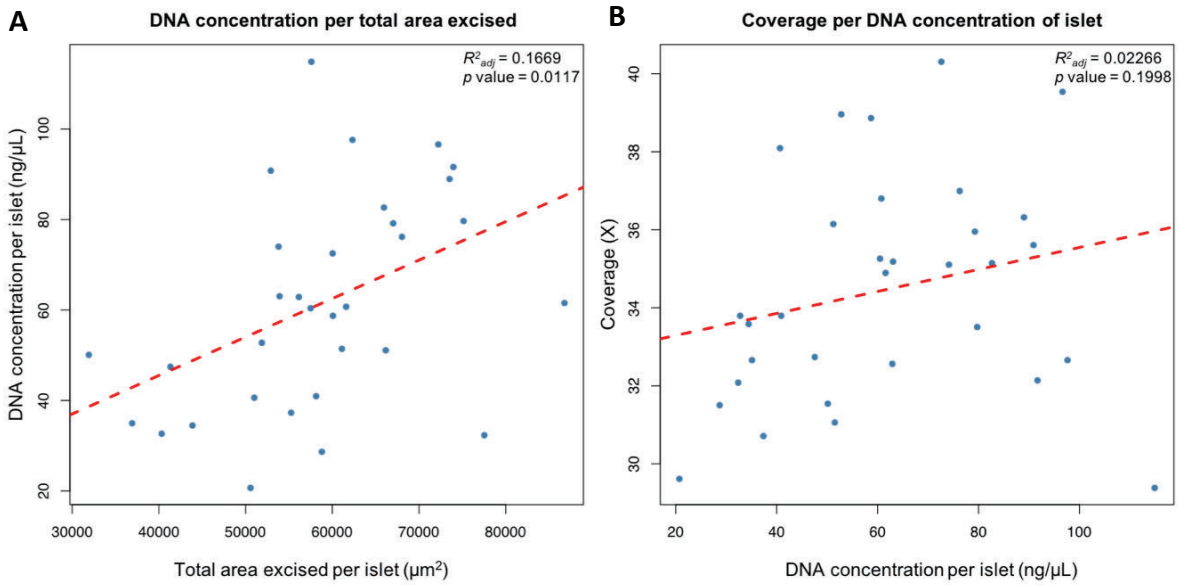


Figure 10 – The data metrics from the LCM workflow

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.

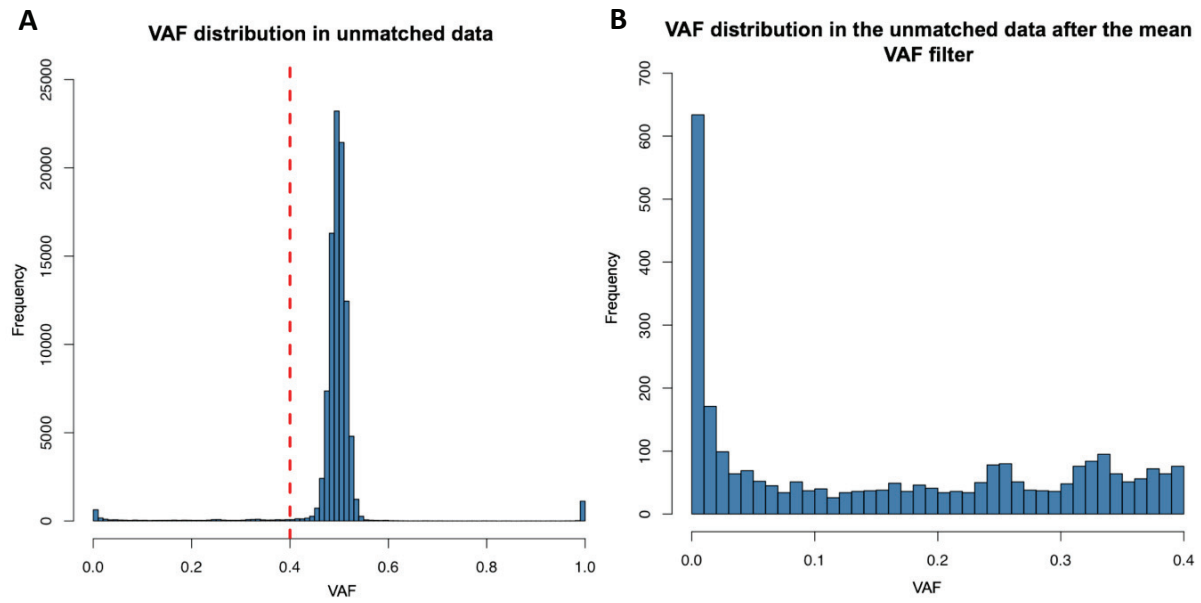
4.2 Successful identification of somatic mutations in individual islets

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.

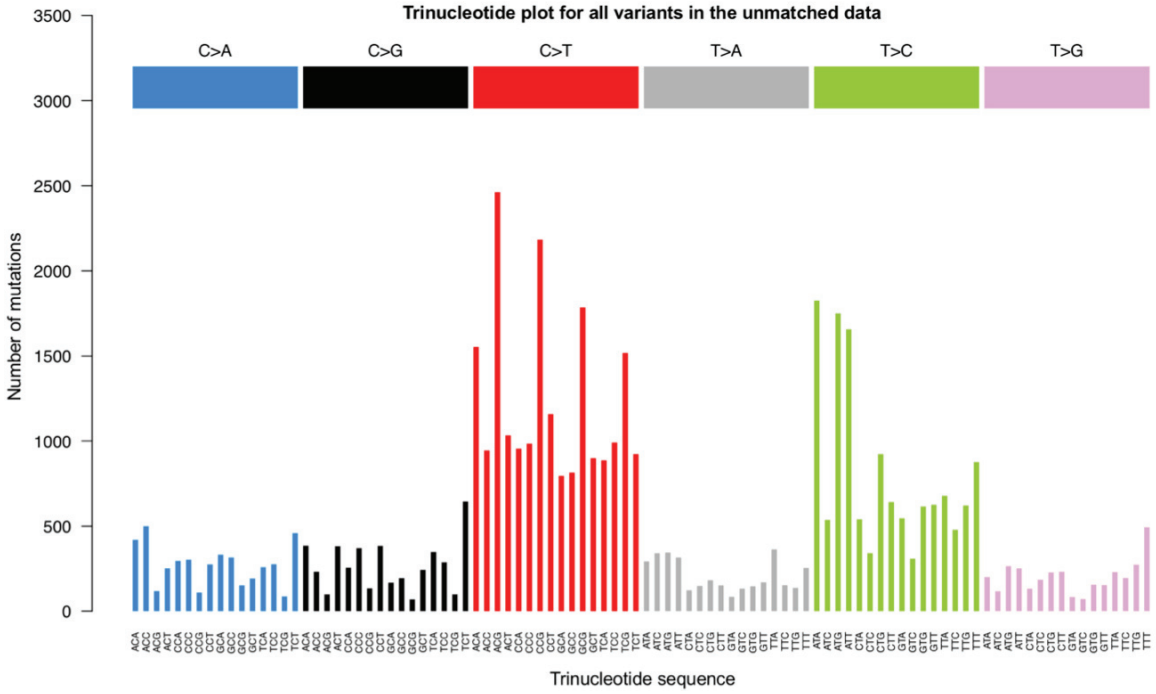
To approach this task, CaVEMan was run in two different ways: a standard run using a matched bladder urothelium (sample PD37726b_lo0071), and an unmatched run, using an unrelated WGS sample as a reference (section 3.7). The latter analysis results in the identification of both somatic and germline mutations,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

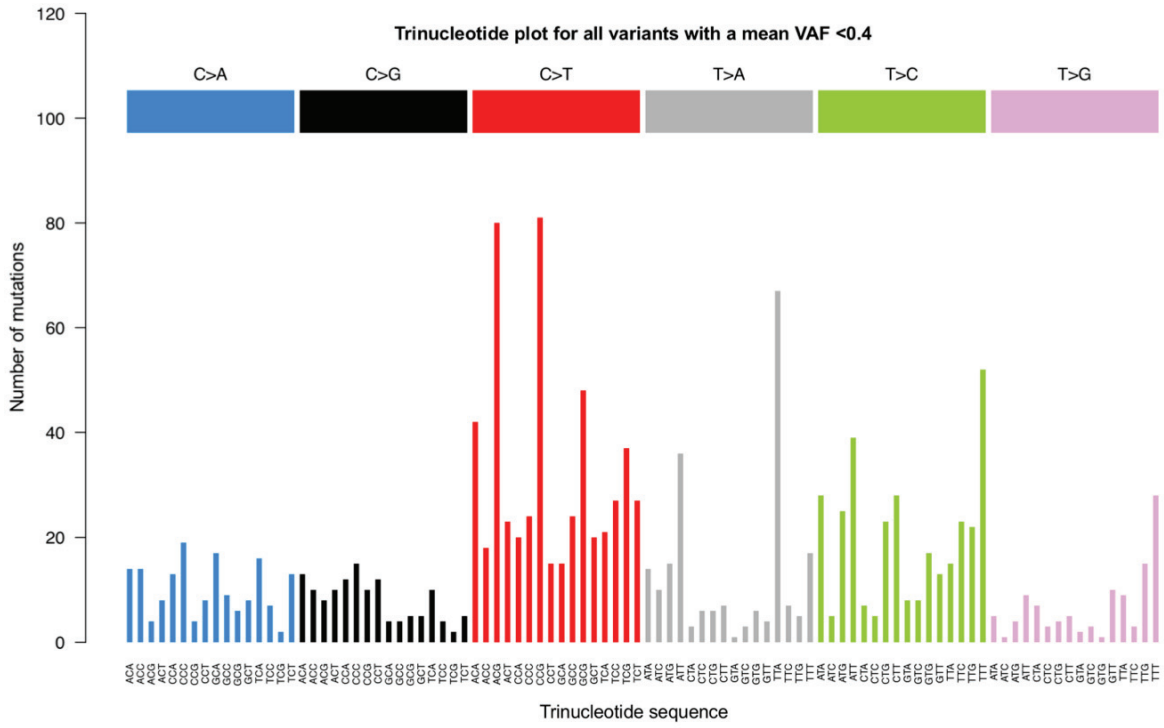

Trinucleotide plot for all variants with a mean VAF <0.4
C>A
C>G
C>T
T>A
T>C
T>G
Number of mutations
0
20
40
60
80
100
120
Trinucleotide sequence

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.

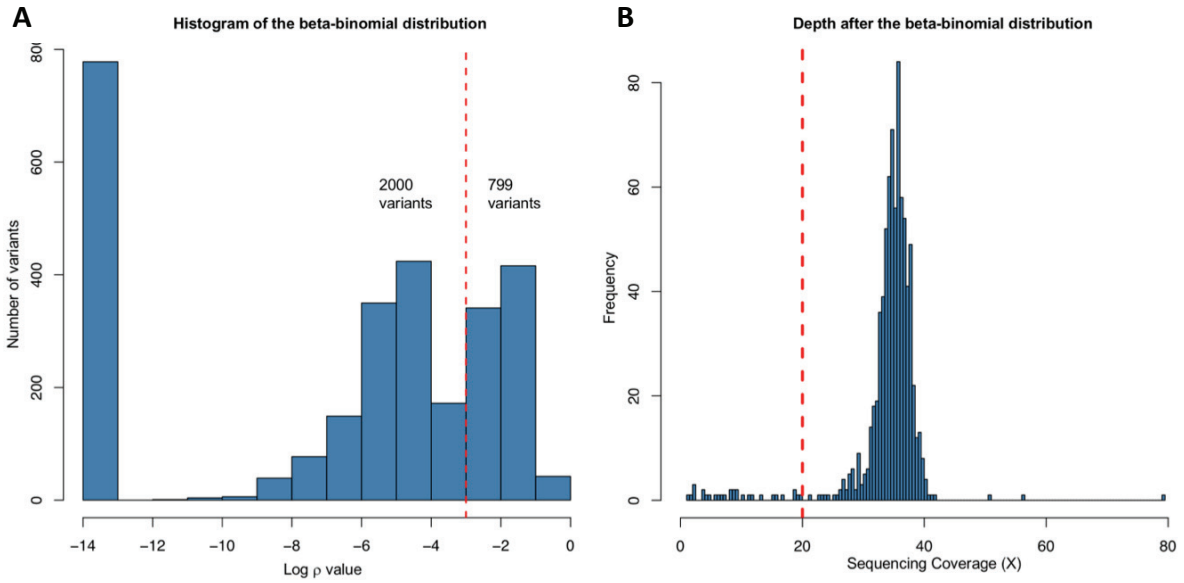
For consistency, the matched run was processed analogously and identified 1,318 variants initially, before this reduced to 1,284 after the mean VAF filter.

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).



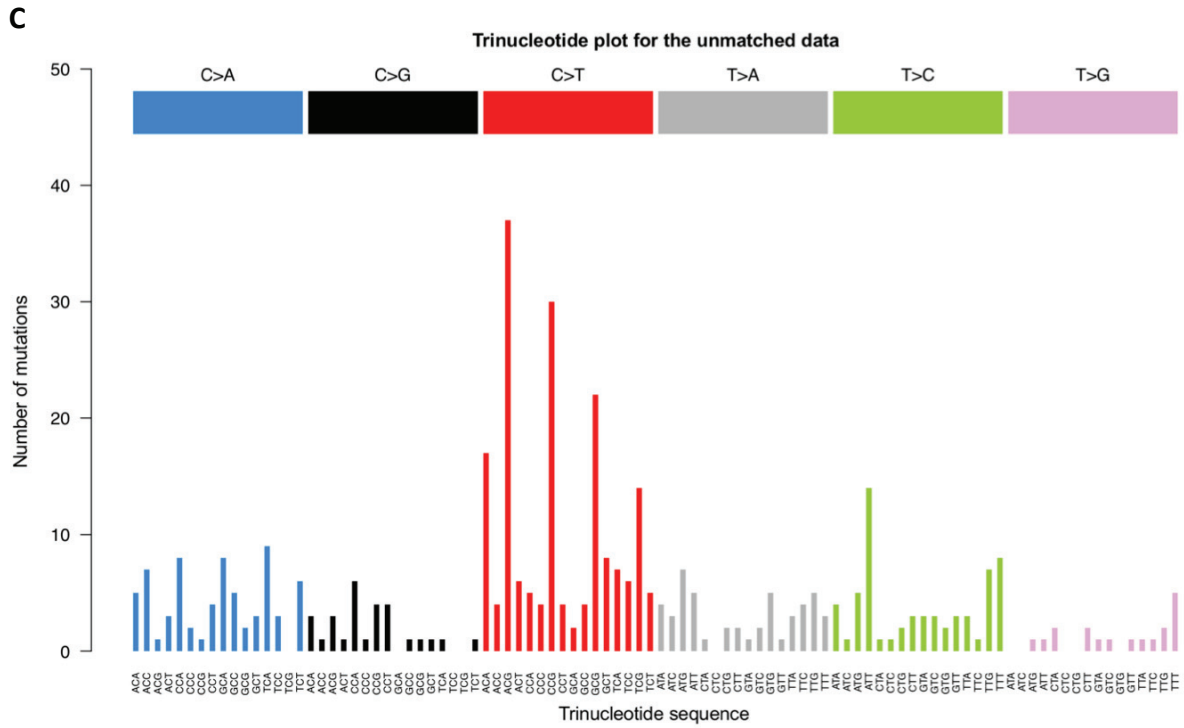


Figure 12 – The beta-binomial distribution identified over-dispersed variants
(A) The beta-binomial distribution applied to the 2,799 variants that passed all previous filters in the unmatched data. The vertical red line represents log ρ of -3, and 799 variants are shown exceeding the over-dispersion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

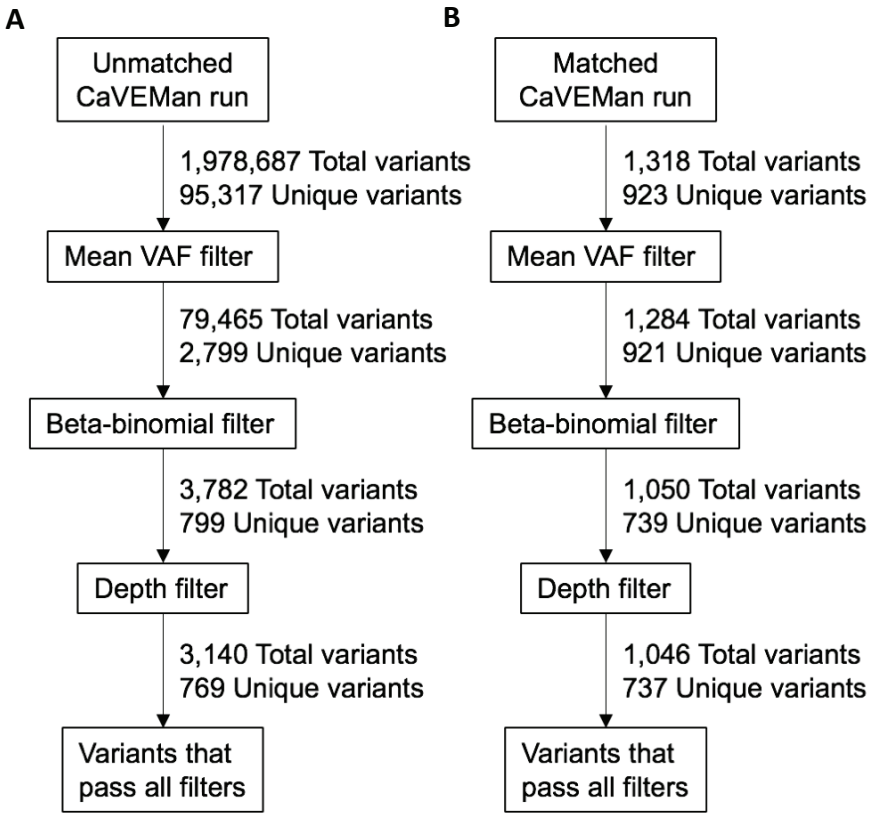


Figure 13 – The two workflows for the matched and unmatched data

(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.

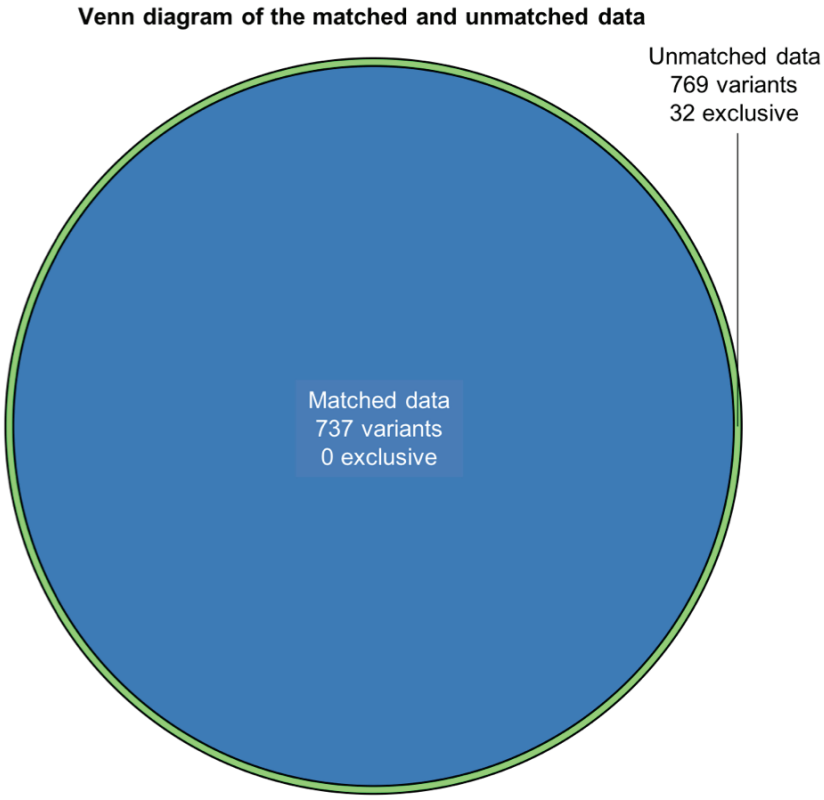


Figure 14 - A Venn diagram demonstrating the overlap between the matched and unmatched data

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.

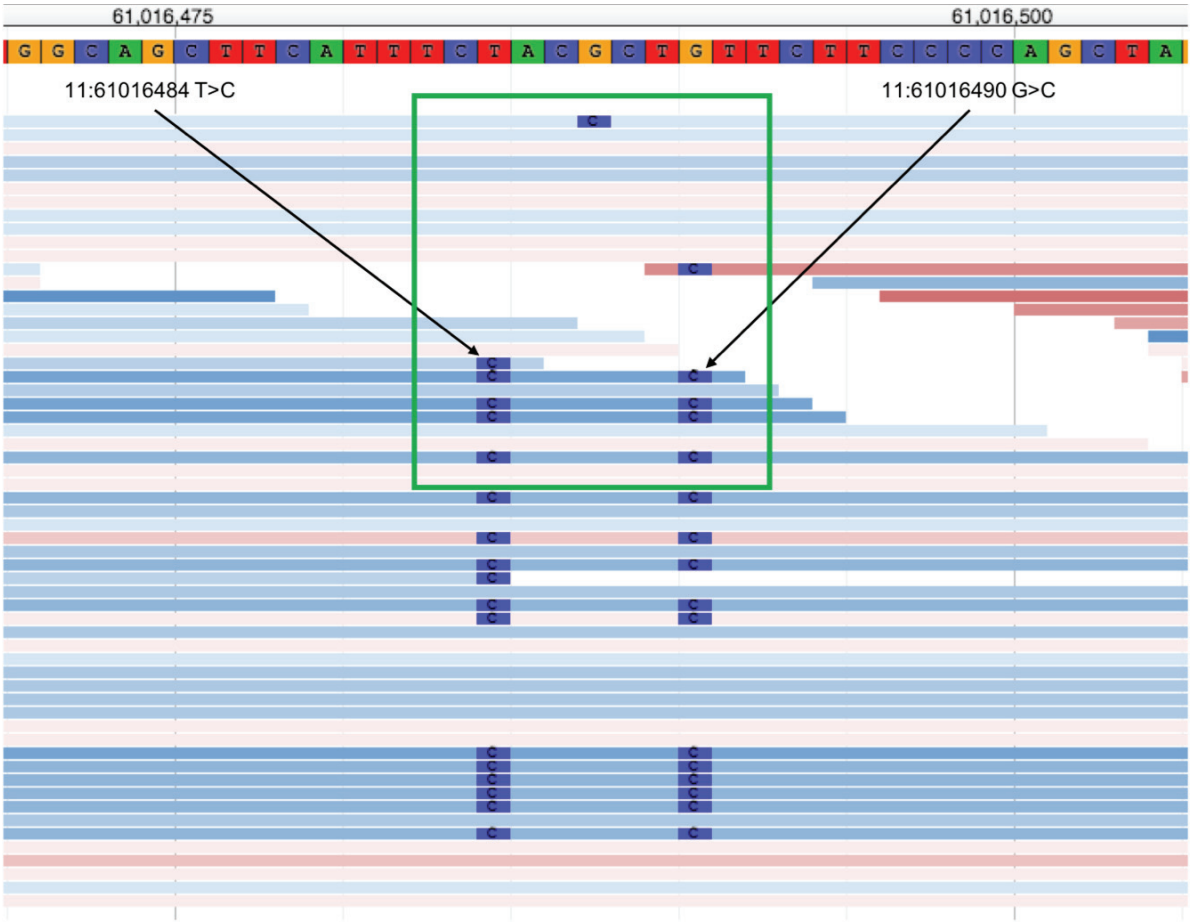


Figure 15 – Manual inspection of the two variants that were excluded from the unmatched data

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).

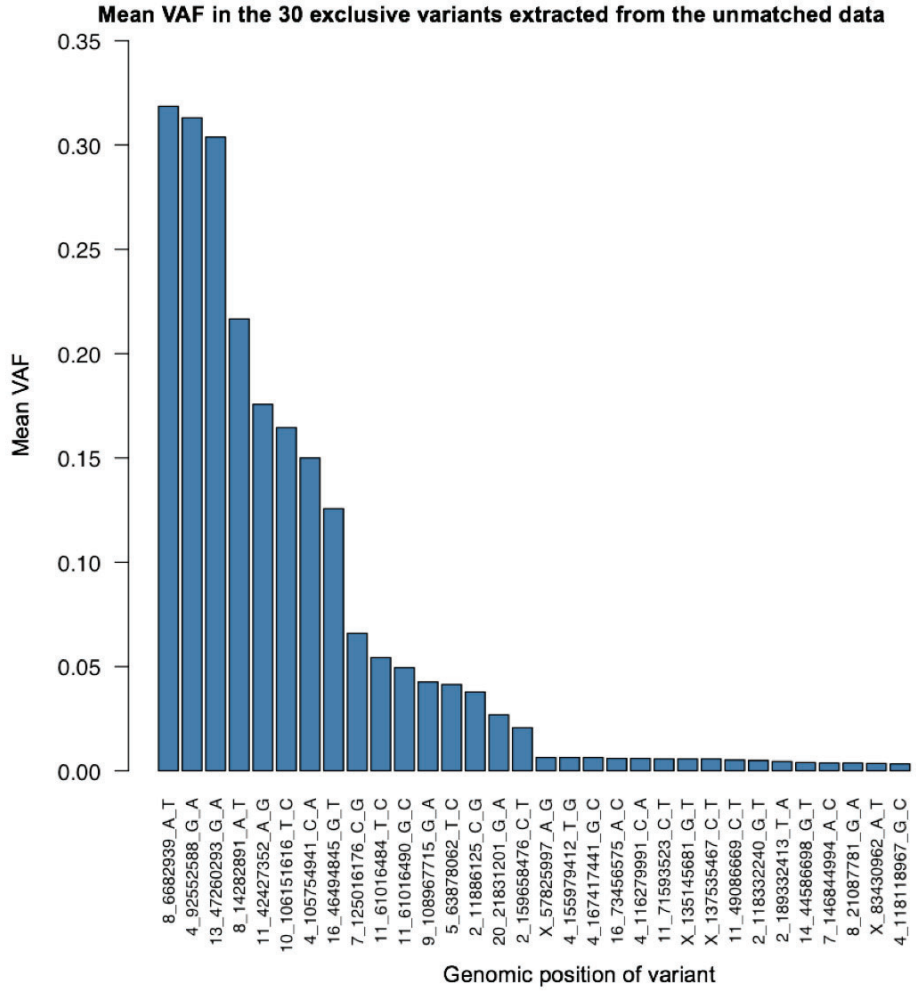


Figure 16 – The 30 exclusive variants showed a range of mean VAFs

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).

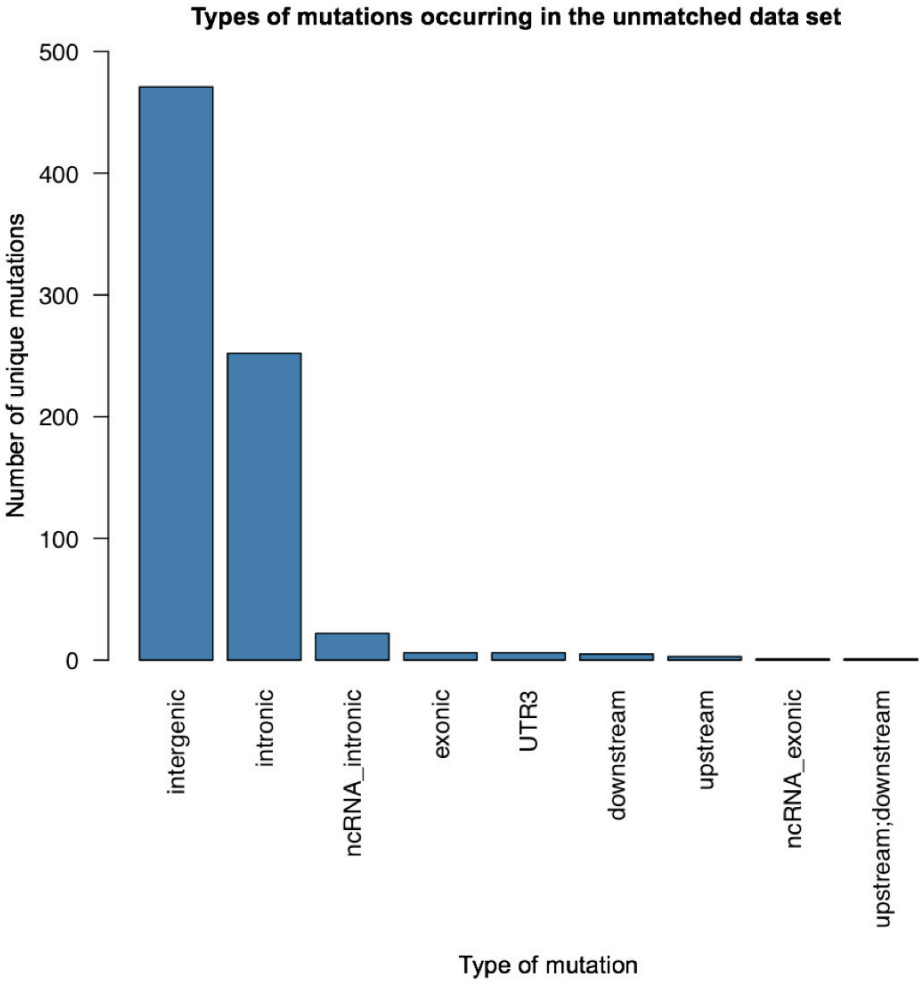


Figure 17 - The frequency of mutations in different regions of the genome
The vast majority of variants are found in intergenic and intronic locations, with very few in the exonic regions. ncRNA is non-coding RNA and UTR3 is the untranslated region in the 3' end of the gene.

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.

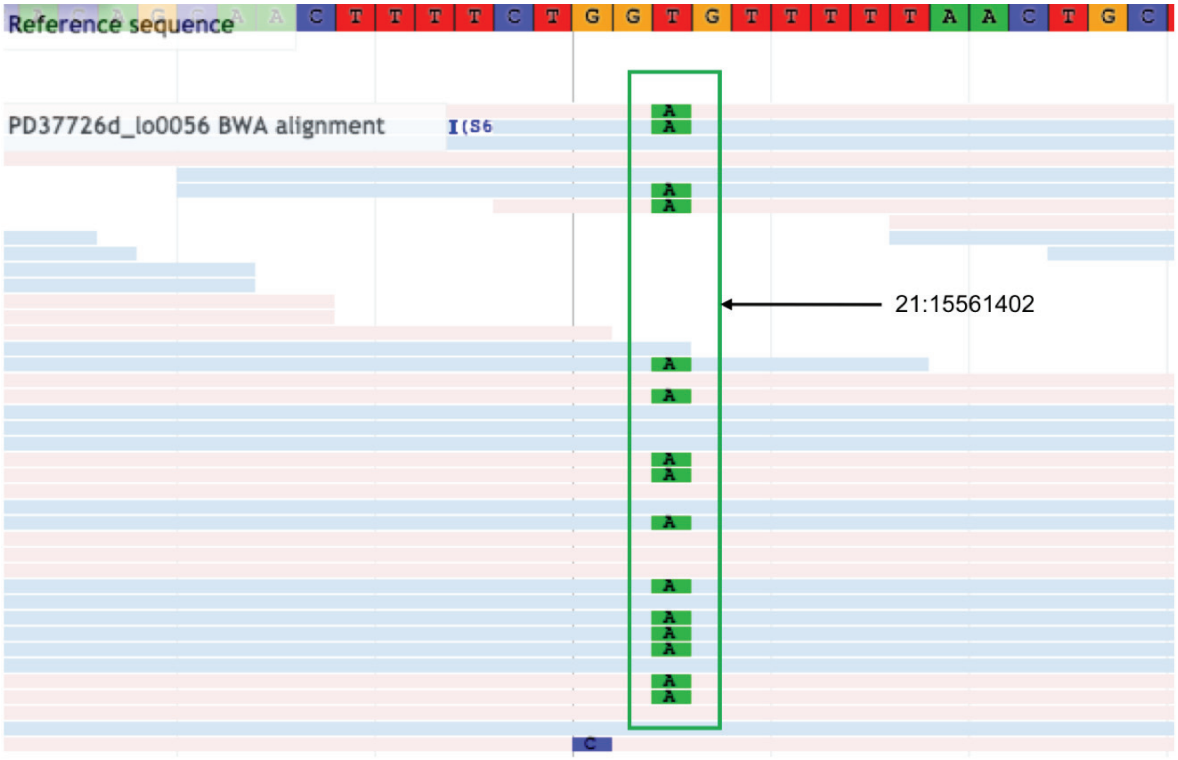


Figure 18 – The *LIPI* variant in sample PD37726d_lo0056

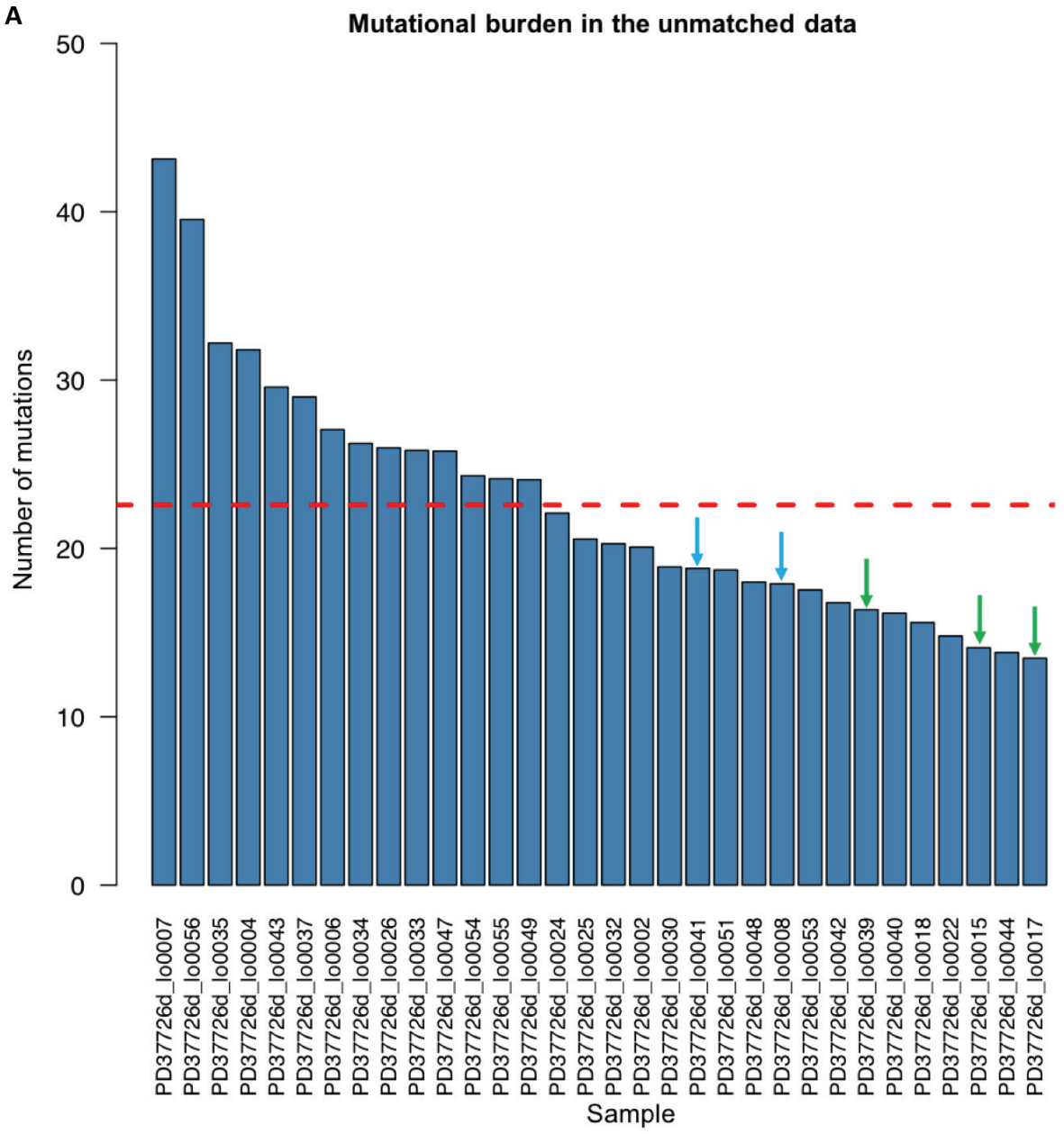
JBrowse screenshot showing the sequencing reads from sample PD37726d_lo0056 (Buels et al., 2016). Each horizontal bar is an individual read with the red representing a forward strand and blue being a reverse. The lighter the shade of these colours, the better the read quality.

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.



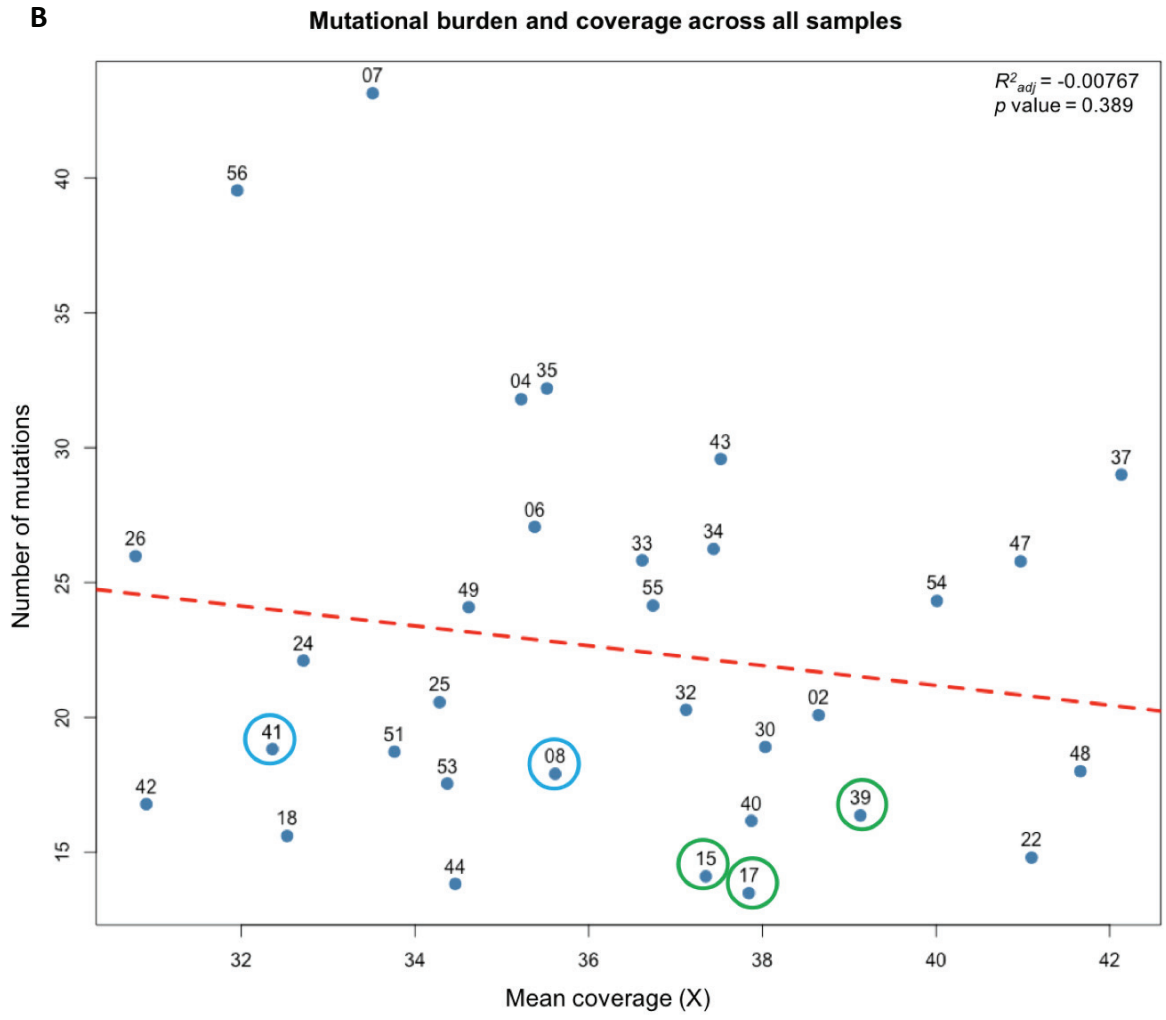


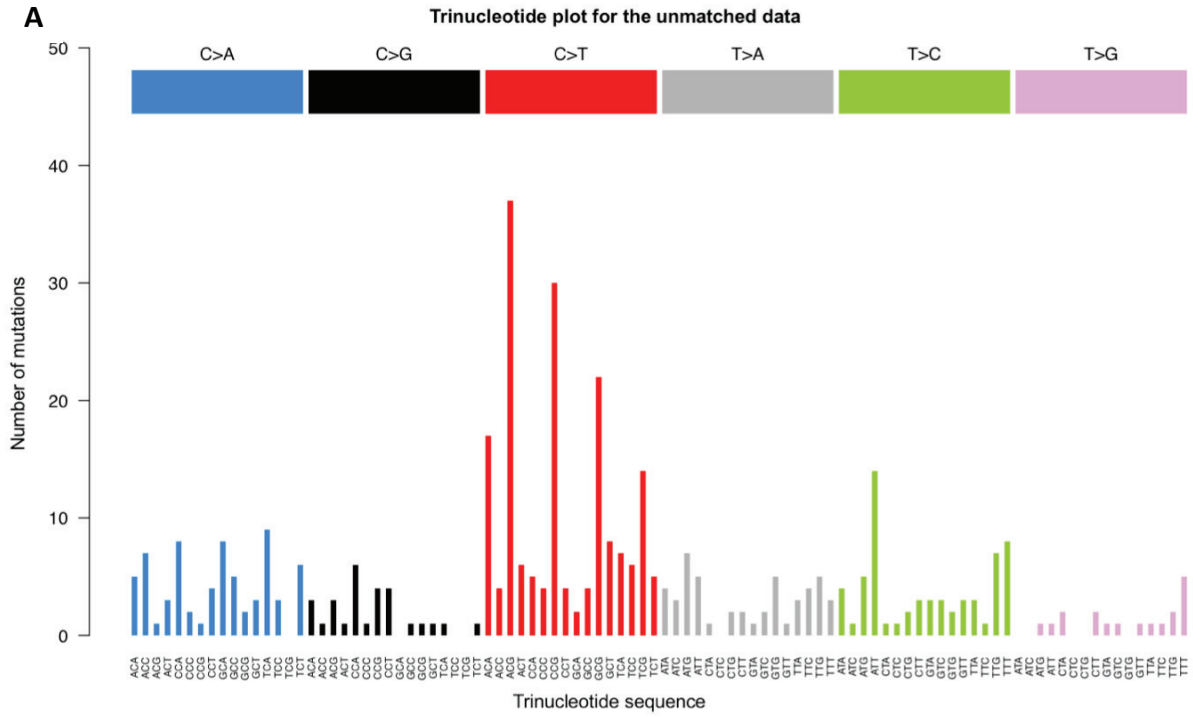
Figure 19 – The observed mutational burden in the pancreatic islets
Duplicates are highlighted in blue and triplicates in green.

(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).



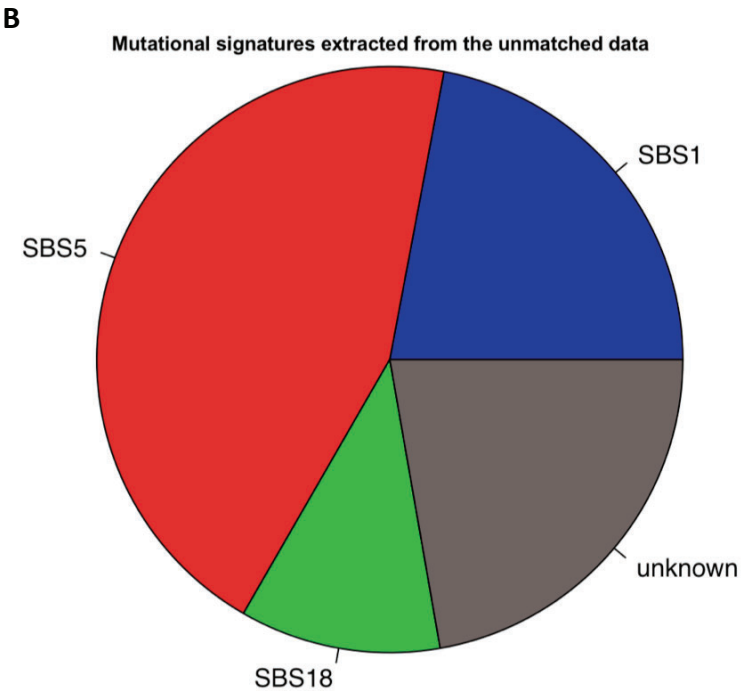


Figure 20 - The mutational signatures in the pancreatic islets
(A) The 96-trinucleotide bar plot for the 767 mutations in the unmatched data. C>T mutations are the most common base substitutions in the islets.

(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.

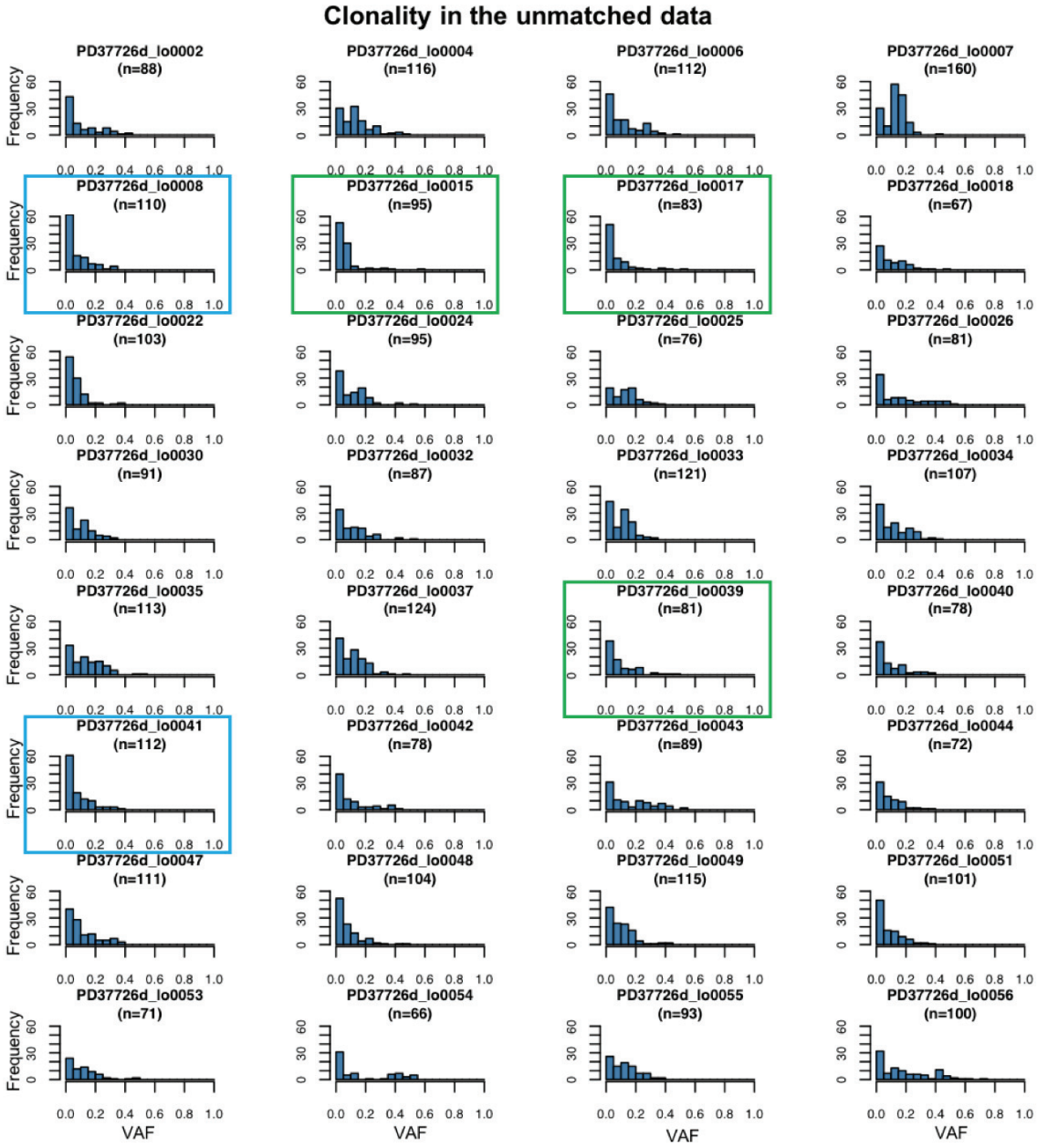


Figure 22 – The clonality of the 32 pancreatic islet samples

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).

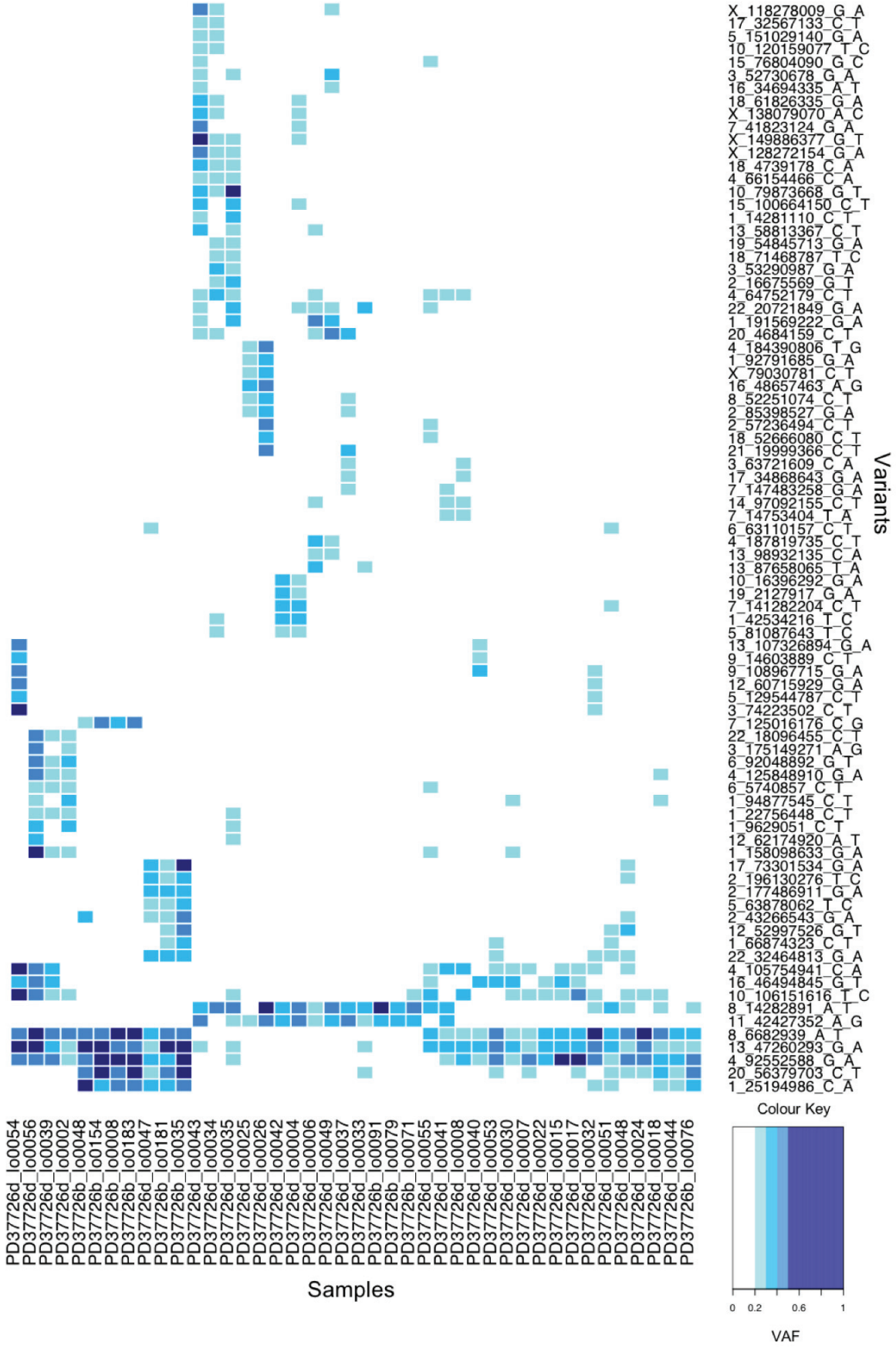
Table 4 – The different number of variants available for phylogenetic tree reconstruction, when including two additional criteria.

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

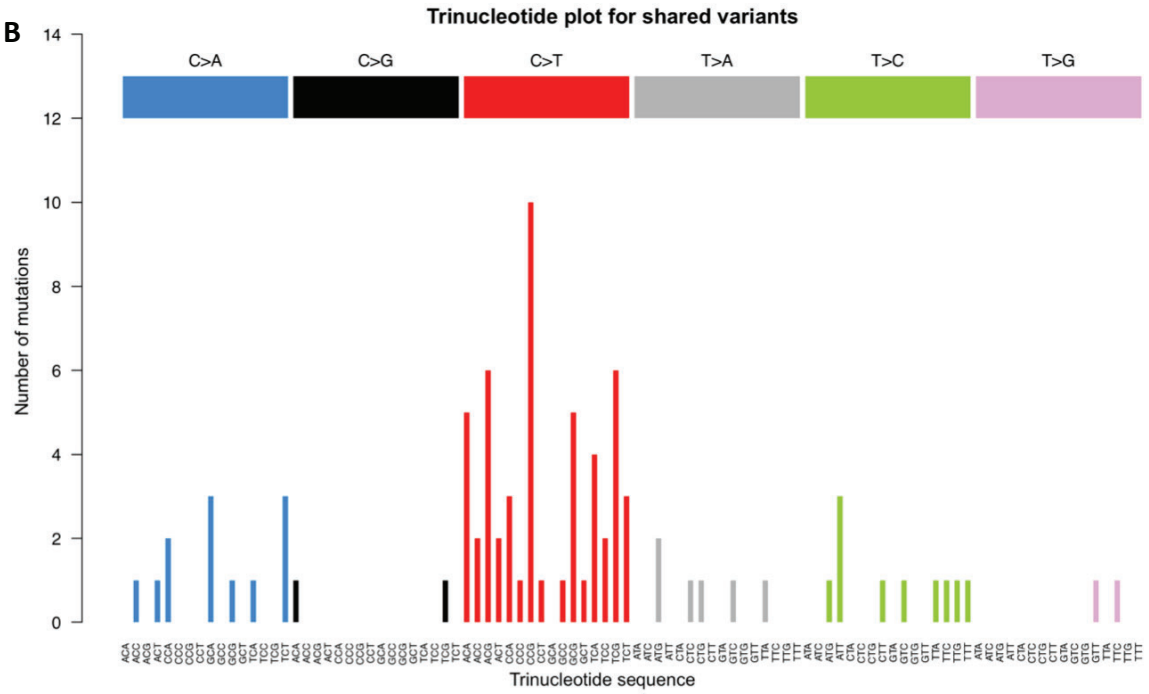


Figure 23 – The 84 variants used for phylogenetic tree reconstruction

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).

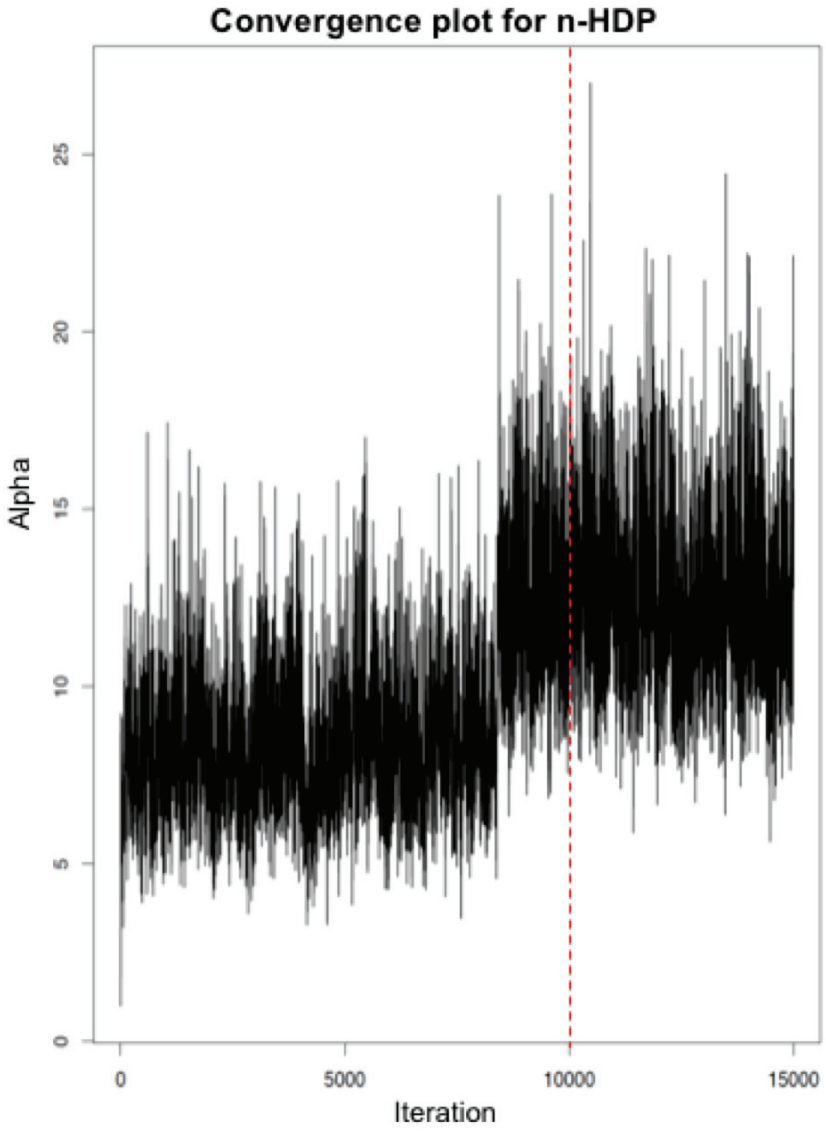


Figure 24 - The convergence plot generated by n-HDP

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.

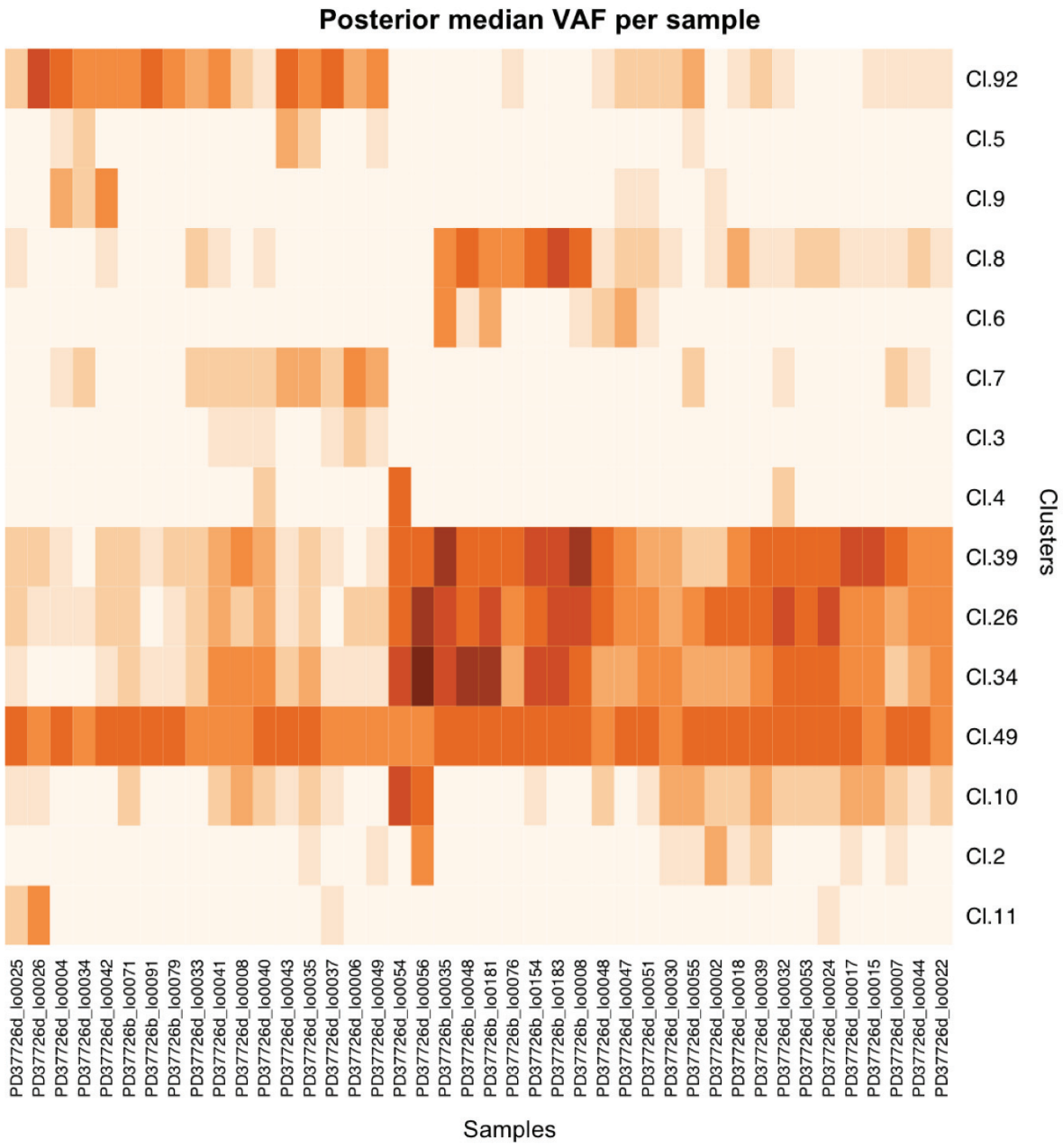


Figure 25 – The n-HDP clustering output

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.

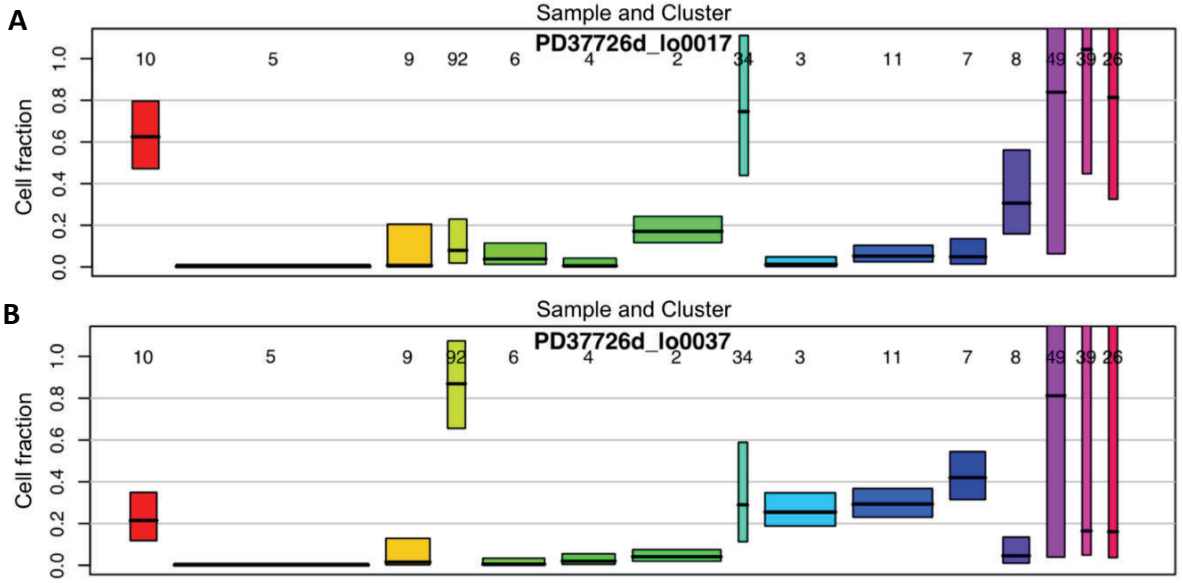


Figure 26 – The pigeonhole principle identified the phylogeny of clusters
Boxes each showing the cell fraction occupied by each cluster (y-axis). Cell fraction is equal to the VAF doubled. The numbers represent the cluster number assigned by n-HDP (x-axis). The sample name is at the top of each box.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

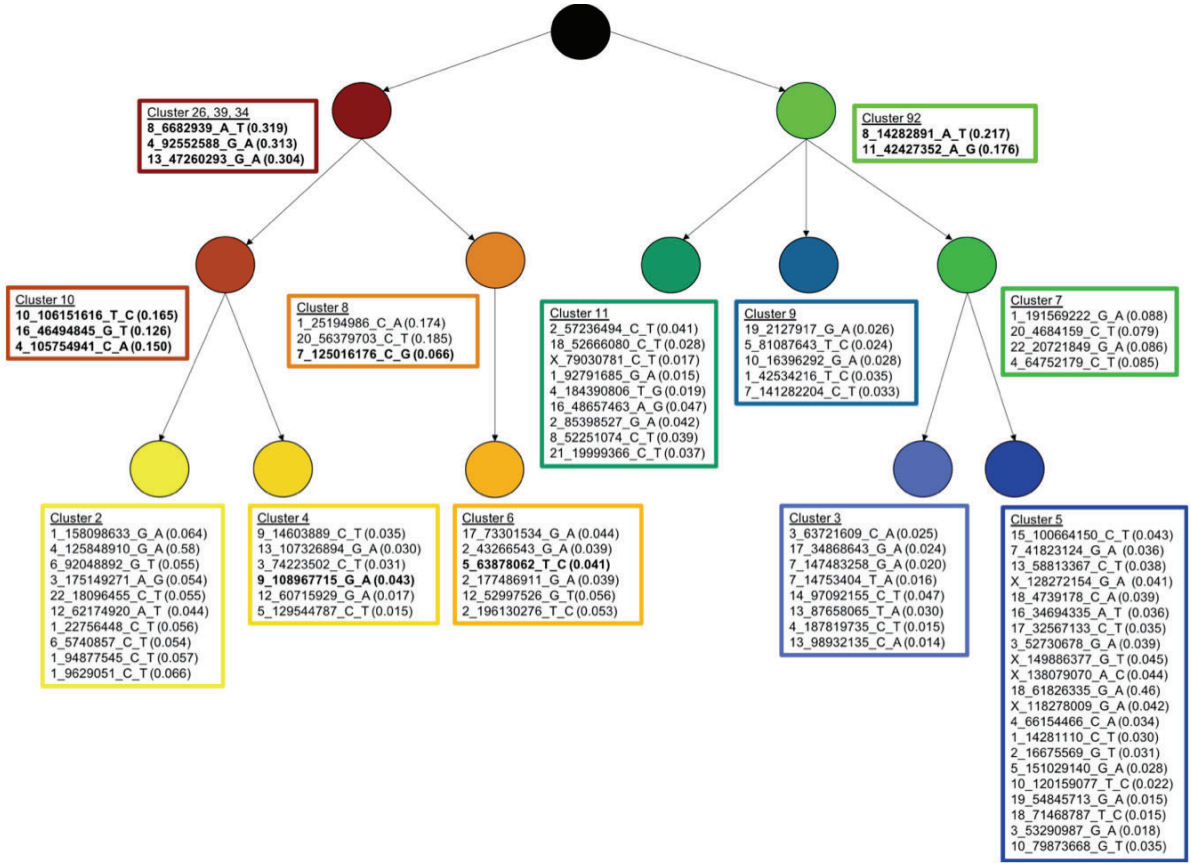


Figure 27 – The phylogenetic tree reconstructed with the n-HDP clustering
Each branch has an associated set of unique mutations with the mean VAF in brackets. The variants are assigned to an individual cluster, as per the n-HDP clustering. The variants in **bold** represent those recovered from the 30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

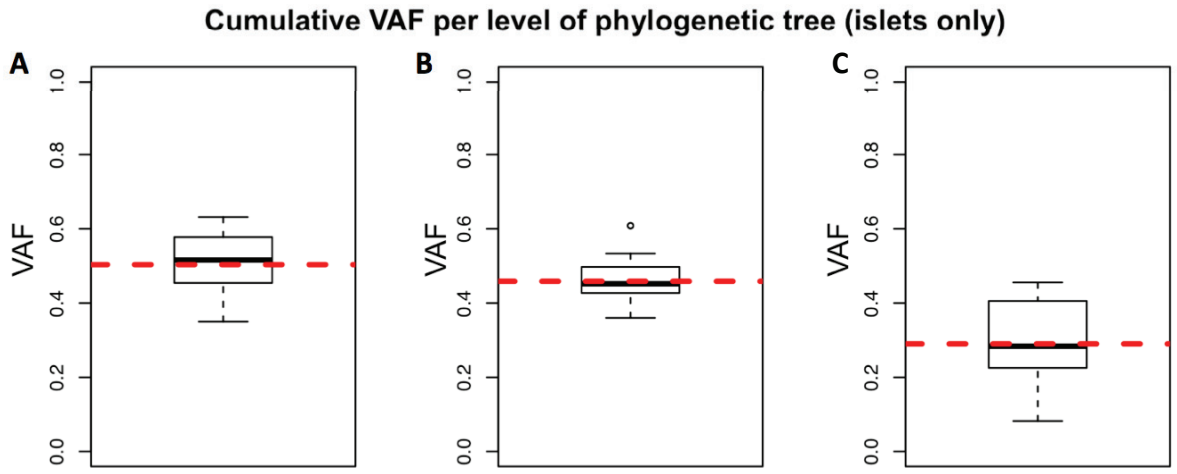


Figure 28 – The cumulative VAF for the islet samples

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).

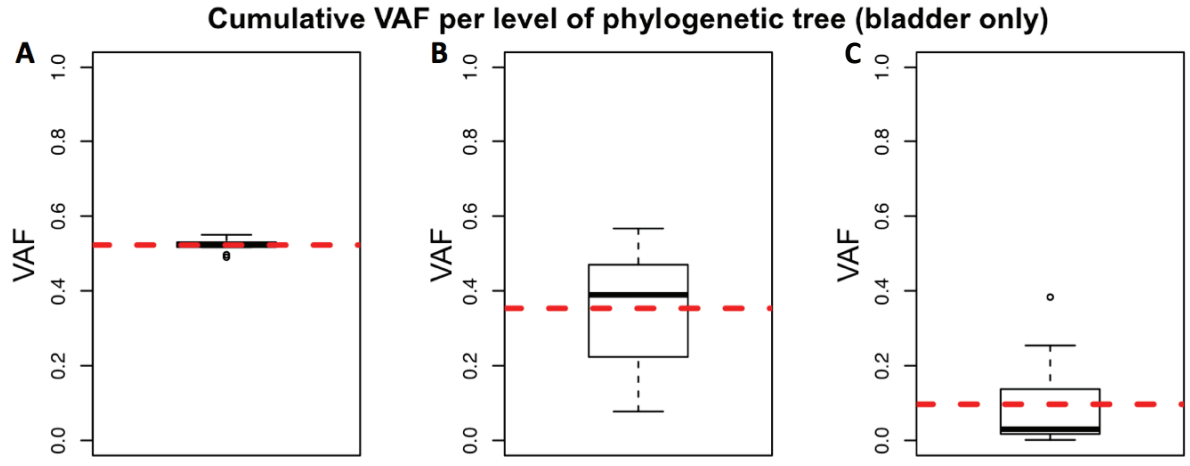


Figure 29 – The cumulative VAF for the bladder urothelium samples

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).

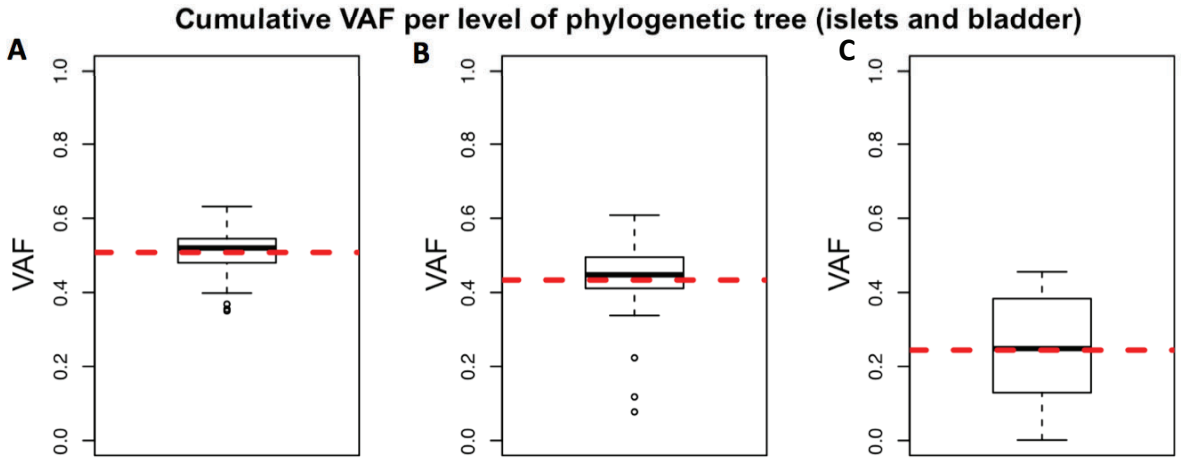


Figure 30 - The cumulative VAF for both the islets and bladder samples

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).

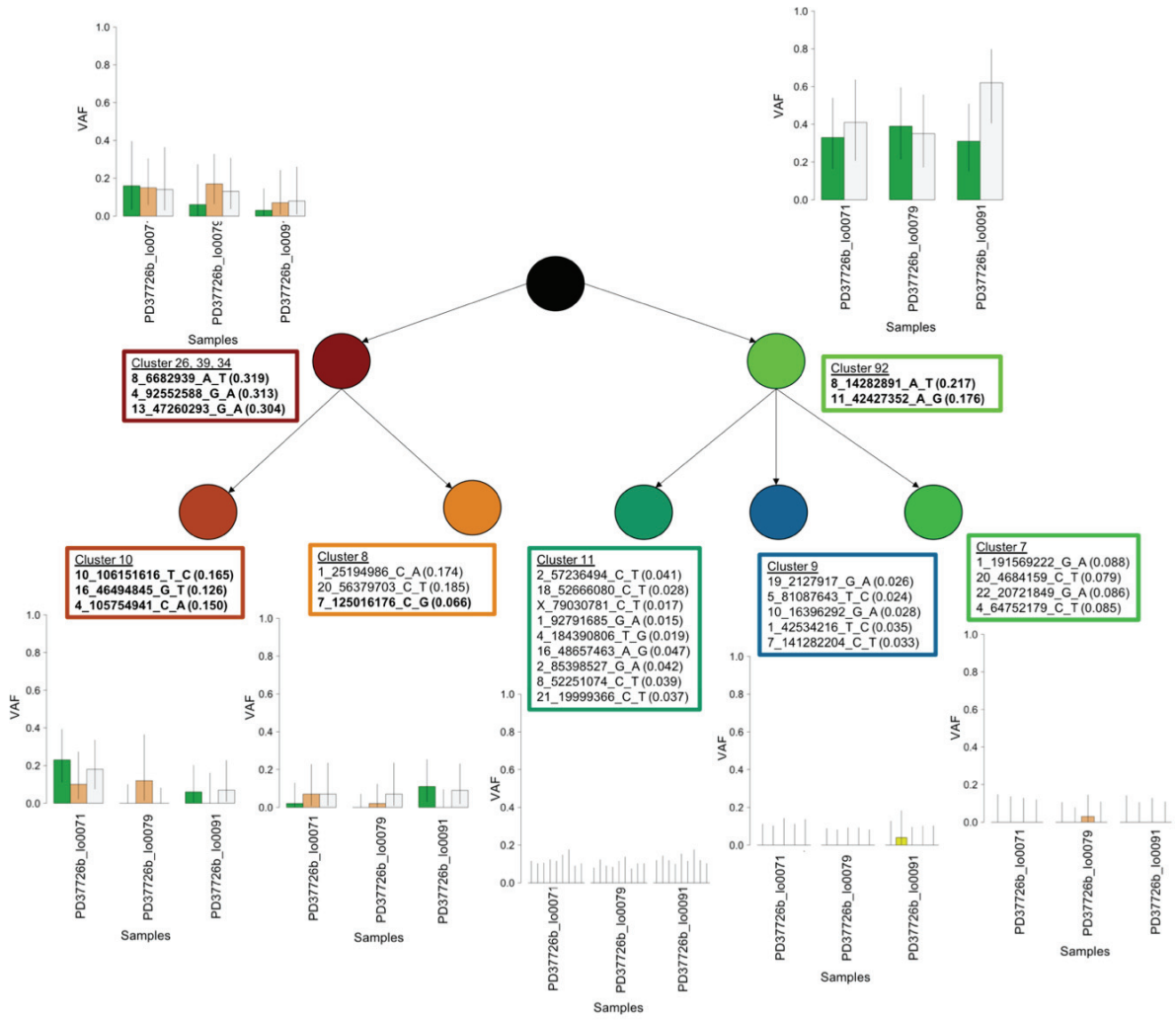


Figure 31 - The early phylogeny of the bladder triplicate

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).

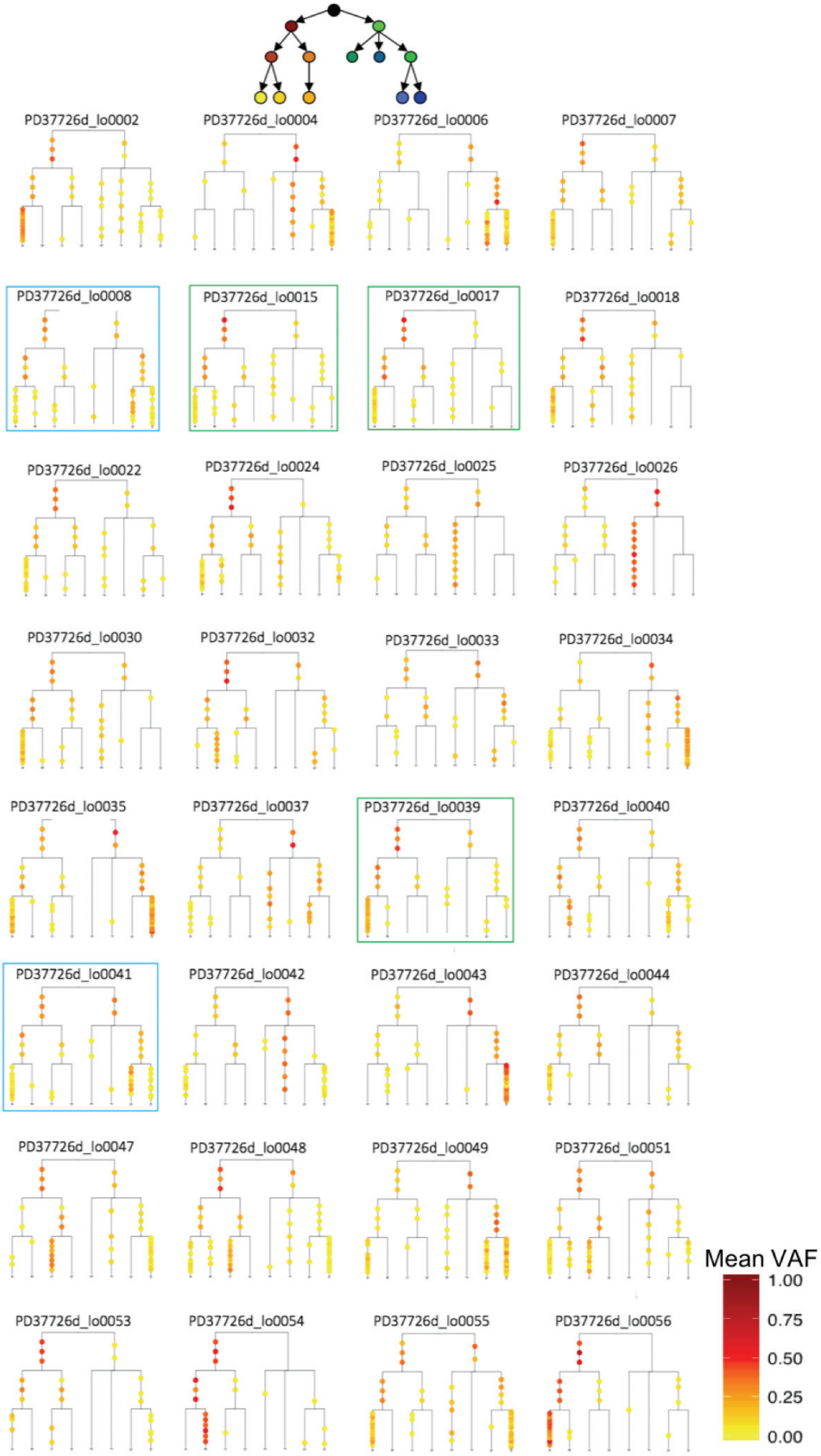

PD37726d_lo0002
PD37726d_lo0004
PD37726d_lo0006
PD37726d_lo0007
PD37726d_lo0008
PD37726d_lo0015
PD37726d_lo0017
PD37726d_lo0018
PD37726d_lo0022
PD37726d_lo0024
PD37726d_lo0025
PD37726d_lo0026
PD37726d_lo0030
PD37726d_lo0032
PD37726d_lo0033
PD37726d_lo0034
PD37726d_lo0035
PD37726d_lo0037
PD37726d_lo0039
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.

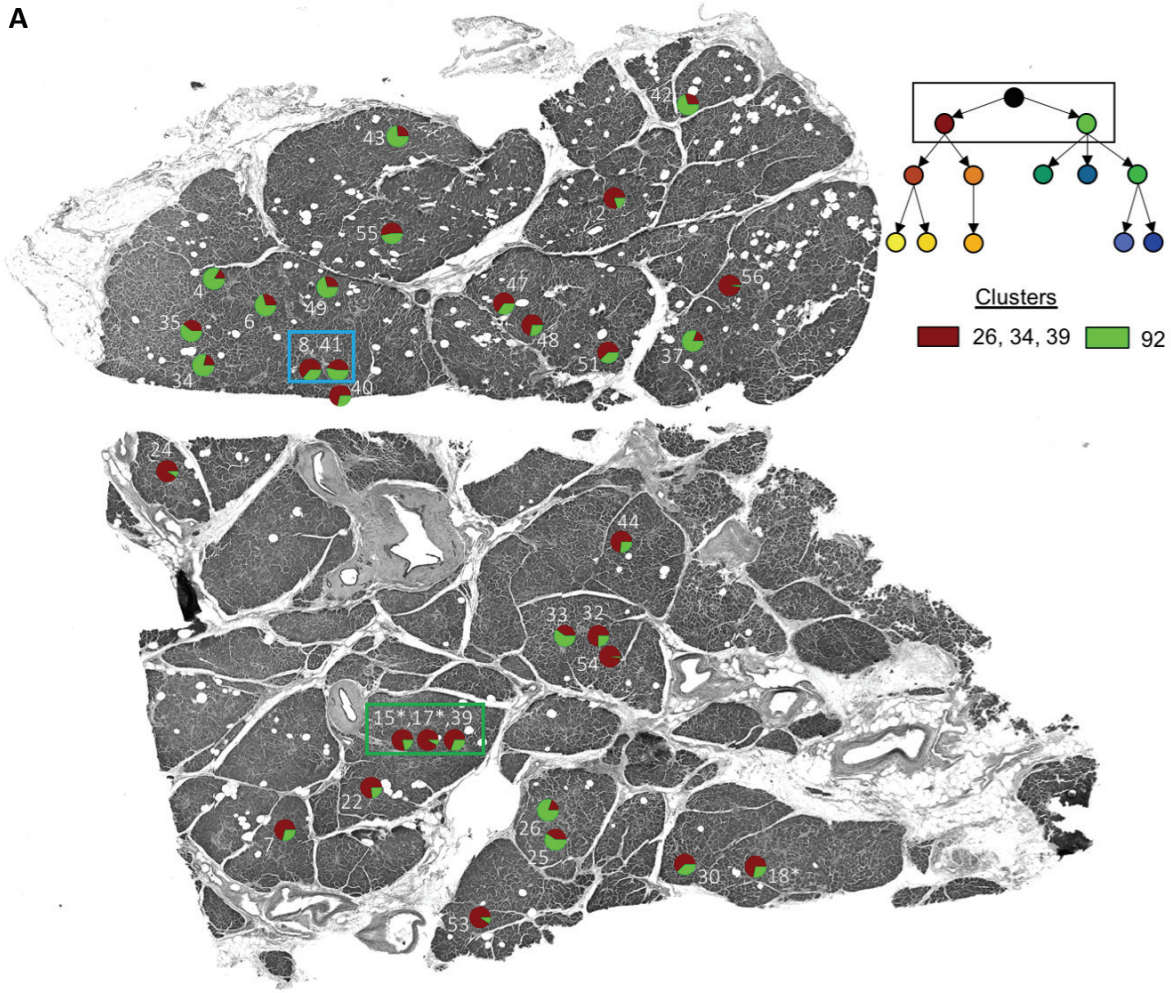
Figure 32 summarises how much the different embryonic lineages contribute to each islet. For example, sample PD37726d_lo0056 nearly entirely derives from the left most embryonic lineage of the tree. The VAFs close to 0.5 along the left-most branches show that the islet is almost entirely composed of 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).



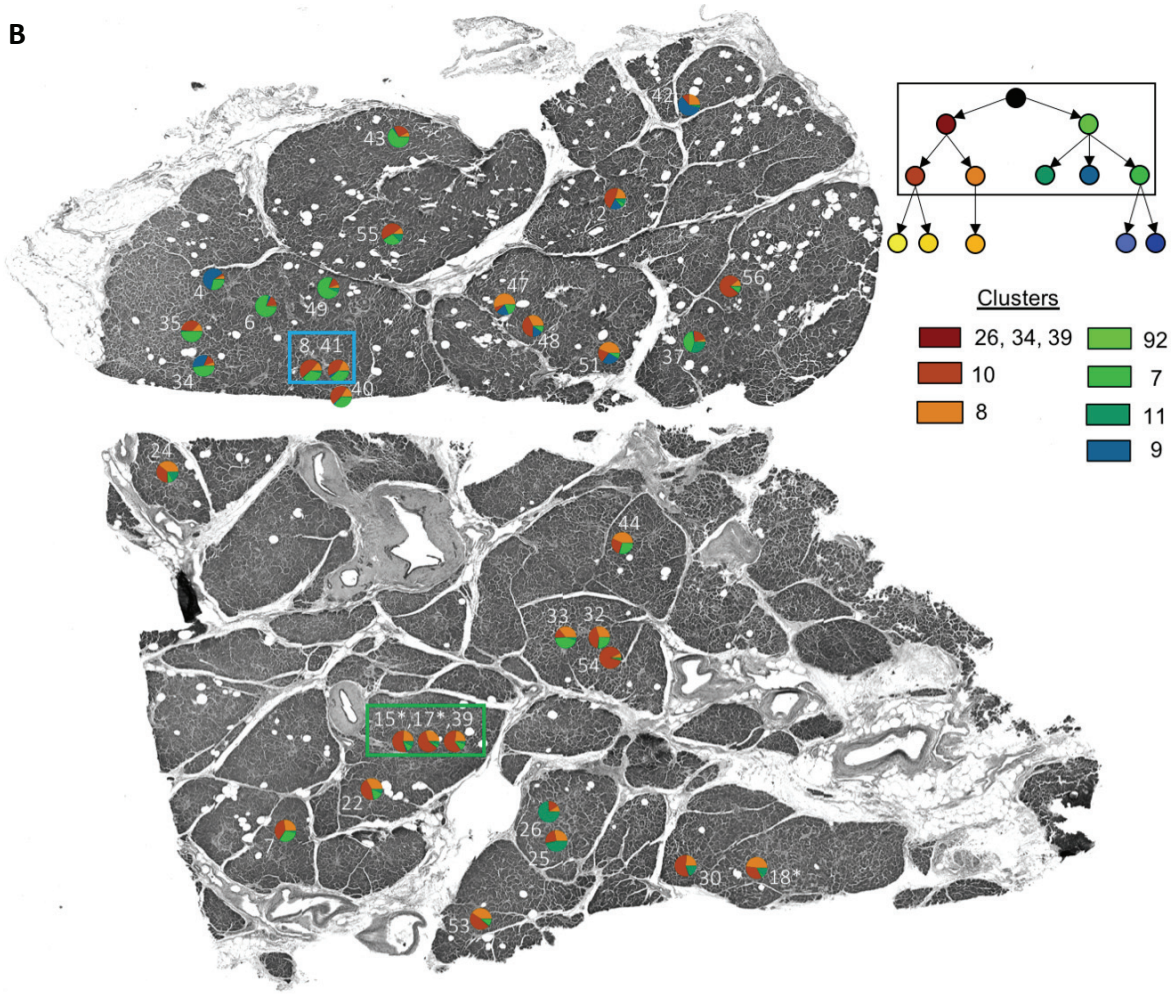


Figure 33 – The integrated spatial and phylogenetic information for each islet
Greyscale pancreas overview section used in LCM. Each pie chart represents the spatial location of an individual islet and the number labelling these pie charts is the suffix of each sample ("PD37726d_lo00*"). Islets labelled 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).