SUPPLEMENTARY INFORMATION FOR

Somatic mutant clone selection in human skin varies with body site

Joanna C Fowler[1], Charlotte King[1], Christopher Bryant[1], Michael Hall[1,3], Roshan Sood[1], Swee Hoe Ong[1], Eleanor Earp[1], David Fernandez-Antoran[1], Jonas Koeppel[1], Stefan Dentro[1,2], David Shorthouse[3], Amer Durrani[4], Kate Fife[4], Edward Rytina[4], Doreen Milne[4], Amit Roshan[4,5], Krishnaa Mahububani[6], Kourosh Saeb-Parsy[6], Benjamin A Hall[3], Moritz Gerstung[2,7], Philip H Jones[1,3*]

[1] Wellcome Sanger Institute, Hinxton CB10 1SA, UK

[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, UK

[3] MRC Cancer Unit, University of Cambridge, Hutchison-MRC Research Centre, Cambridge Biomedical Campus, Cambridge CB2 0XZ, UK

[4] Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus Cambridge, CB2 0QQ, UK.

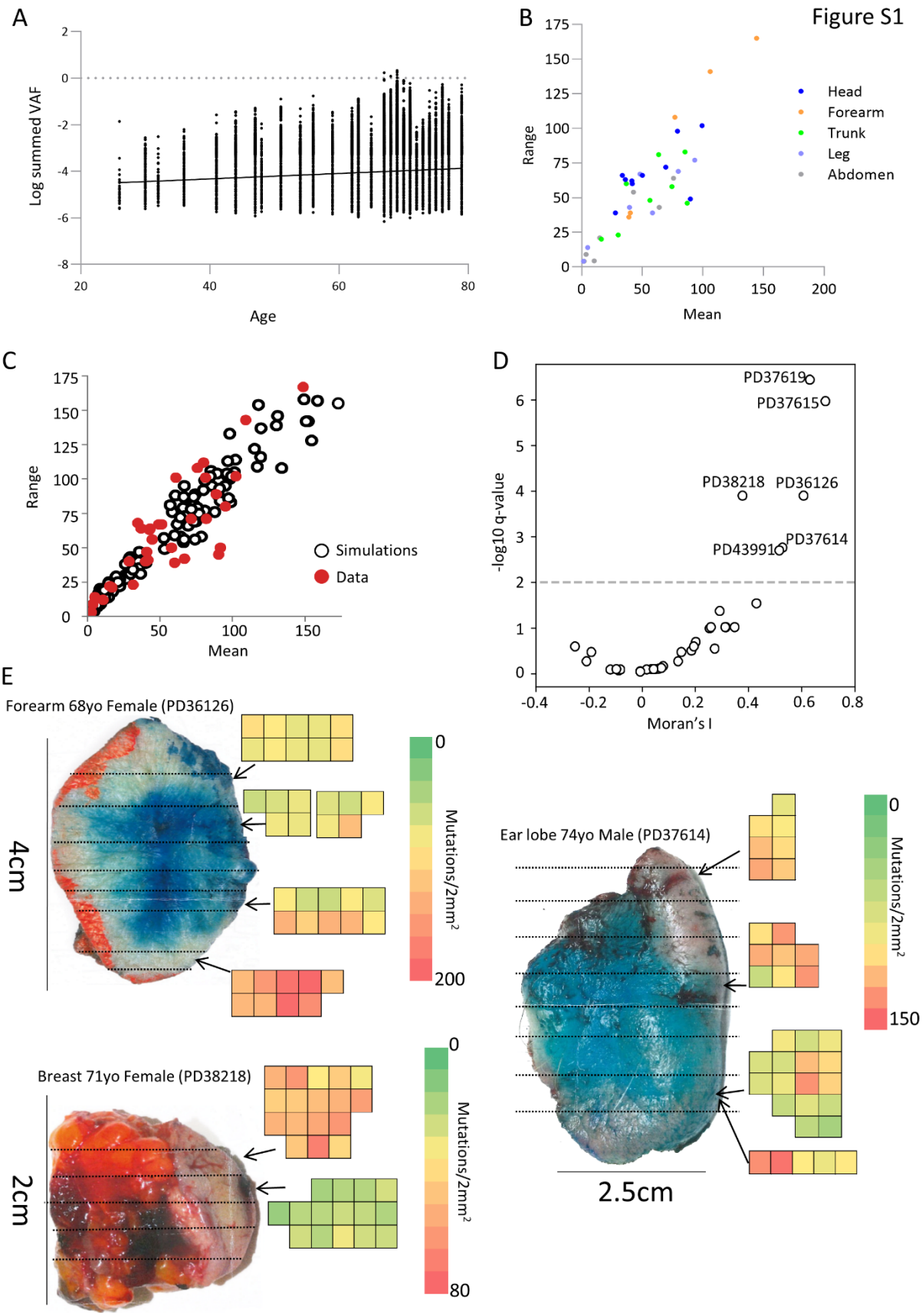
[5] Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

[6] Department of Surgery and Cambridge NIHR Biomedical Research Centre, Biomedical Campus, Cambridge CB2 0QQ

[7] European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, DE

*correspondence to pj3@sanger.ac.uk

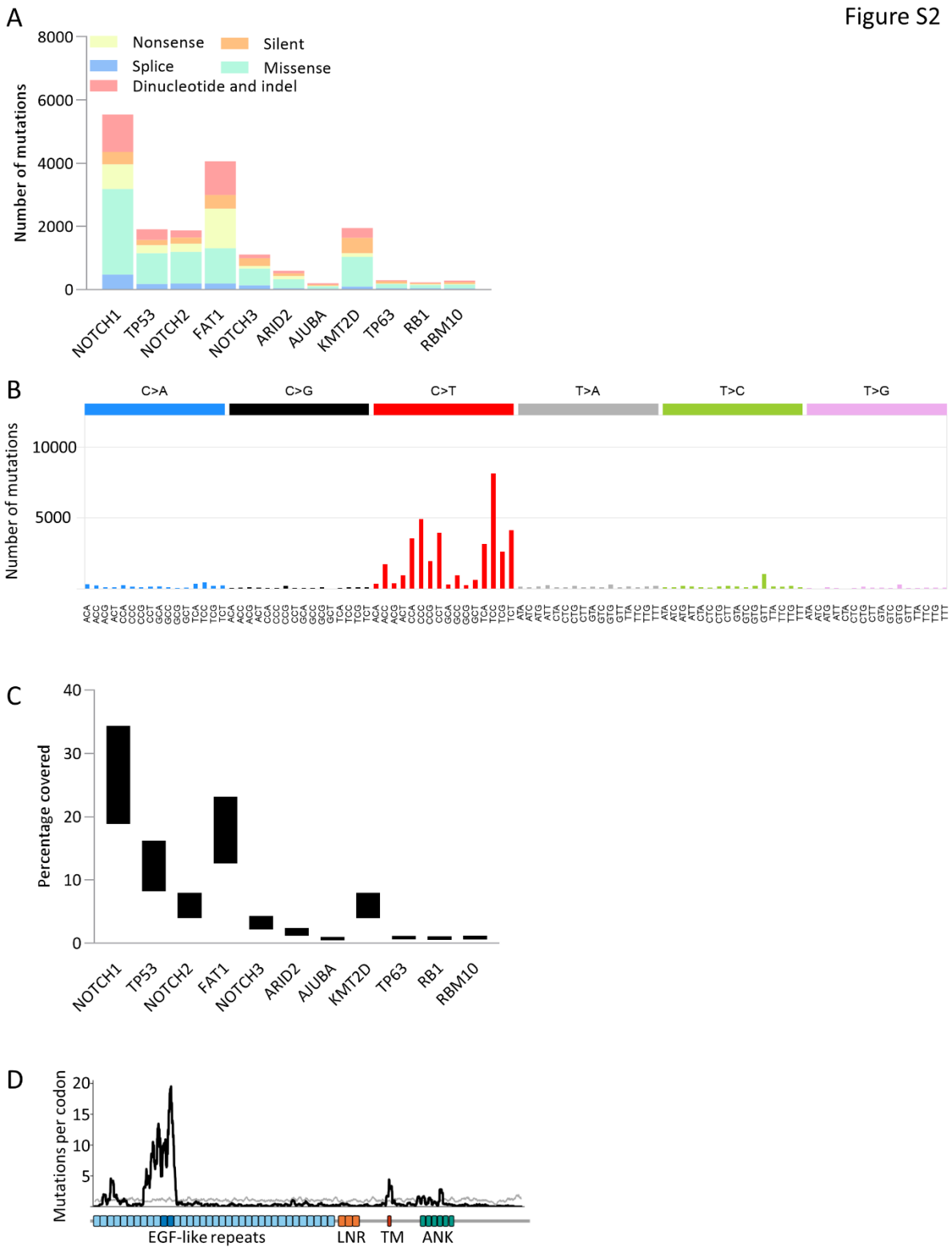
Figure S1


A
Log summed VAF
Age
B
Range
Mean
Head
Forearm
Trunk
Leg
Abdomen
C
Range
Mean
Simulations
Data
D
-log10 q-value
Moran's I
PD37619
PD37615
PD38218
PD36126
PD37614
PD43991
E
Forearm 68yo Female (PD36126)
4cm
Mutations/2mm²
0
200
Breast 71yo Female (PD38218)
2cm
80
Ear lobe 74yo Male (PD37614)
2.5cm
150

Supplemental figure 1 - Assessing mutational variability in human skin

a. Log of summed VAF. A simple linear regression finds that there is an increase in clone size (VAF) with age ($p < 2.2 \times 10^{16}$, slope = 0.01161, intercept = -4.79356), however, this should be interpreted with caution because, even after log transformation, the data does not meet the normal distribution assumption. Note for this analysis PD38219 was removed as an outlier because of their predominantly non-UV signature
b. Mean mutation number per sample vs the range (difference between the highest and lowest mutation count in single biopsies) for each patient. Different colours represent different body sites.
c. Mean mutation number per sample vs the range for each patient (red) and simulations (open black markers).
d. Patients whose epithelia shows a greater than expected difference in the number of mutations/ $2mm^2$ gridded compared to their adjacent neighbours. Moran's I and -log10 q-value for each donor are shown. Values of Moran's I closer to 1 show increased positive autocorrelation (more clustering of high values/clustering of low values).
e. Further examples of epidermis which show a gradient in the number of mutations across the tissue. Heat maps show the number of mutations in each $2mm^2$ grid.

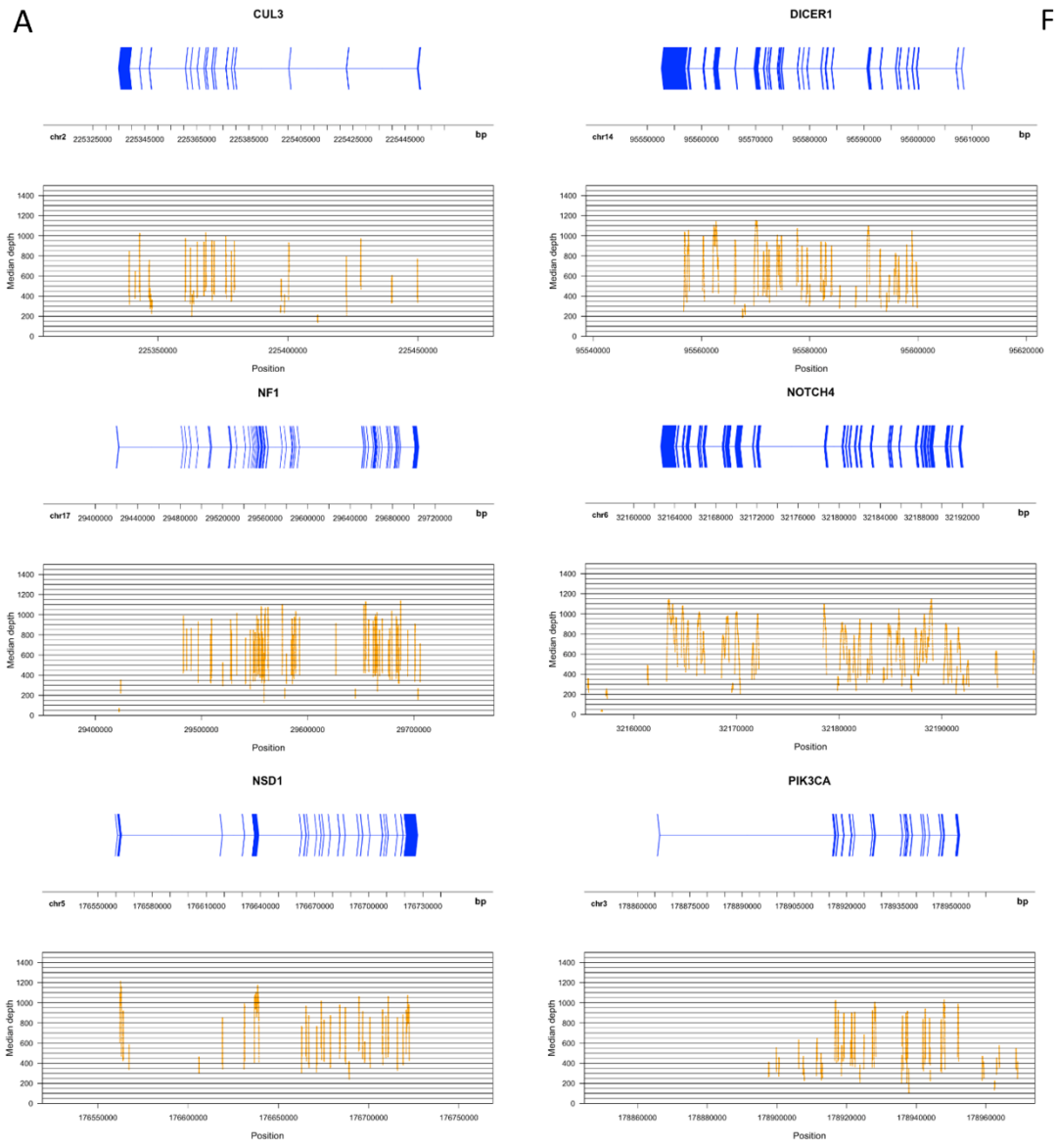
Figure S2

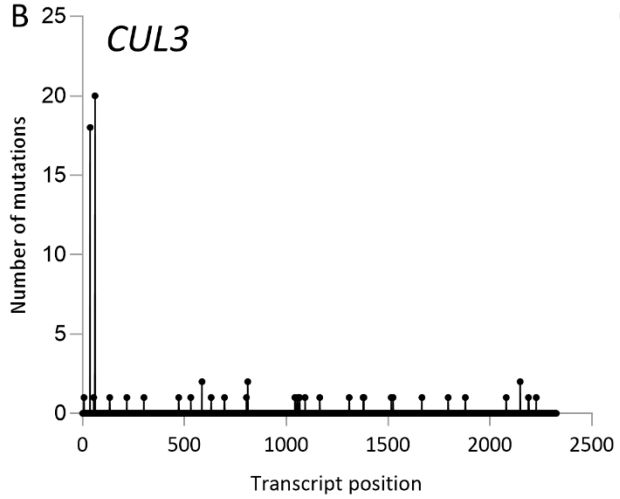

A
Nonsense
Silent
Splice
Missense
Dinucleotide and indel
Number of mutations
8000
6000
4000
2000
0
NOTCH1
TP53
NOTCH2
FAT1
NOTCH3
ARID2
AJUBA
KMT2D
TP63
RB1
RBM10
B
C>A
C>G
C>T
T>A
T>C
T>G
Number of mutations
10000
5000
C
Percentage covered
40
30
20
10
0
NOTCH1
TP53
NOTCH2
FAT1
NOTCH3
ARID2
AJUBA
KMT2D
TP63
RB1
RBM10
D
Mutations per codon
20
15
10
5
EGF-like repeats
LNR
TM
ANK

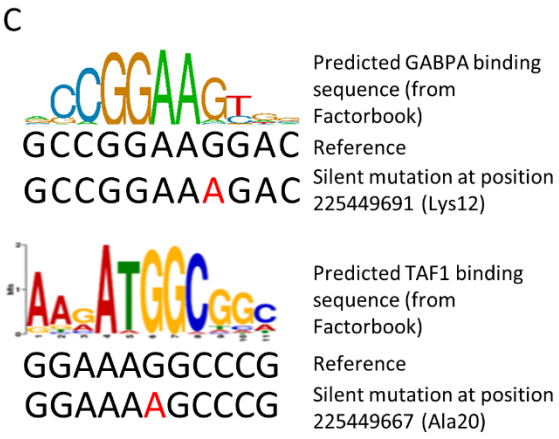
Supplemental figure 2 – Positive selection across different body sites

a. Number of different types of mutations for positively selected genes (globalq<0.01, dN/dS >2)
b. Mutational spectrum of single nucleotide variants from all 2mm^2 grid samples from targeted sequencing. The bar plot shows the number of mutations in each of the 96 possible trinucleotide contexts.
c. Percentage of mutant epithelia for all positively selected genes (global q<0.01, dN/dS >2)
d. Sliding window plot of missense mutations per codon in *NOTCH1*. Observed counts shown by the black line. Expected counts assuming that missense mutations were distributed across the gene according to the mutational spectrum shown in grey. Domains of NOTCH1 protein shown underneath the x-axis. EGF-repeats shown in blue, with ligand-binding site EGF repeats 11-12 shown in dark blue.

Figure S3

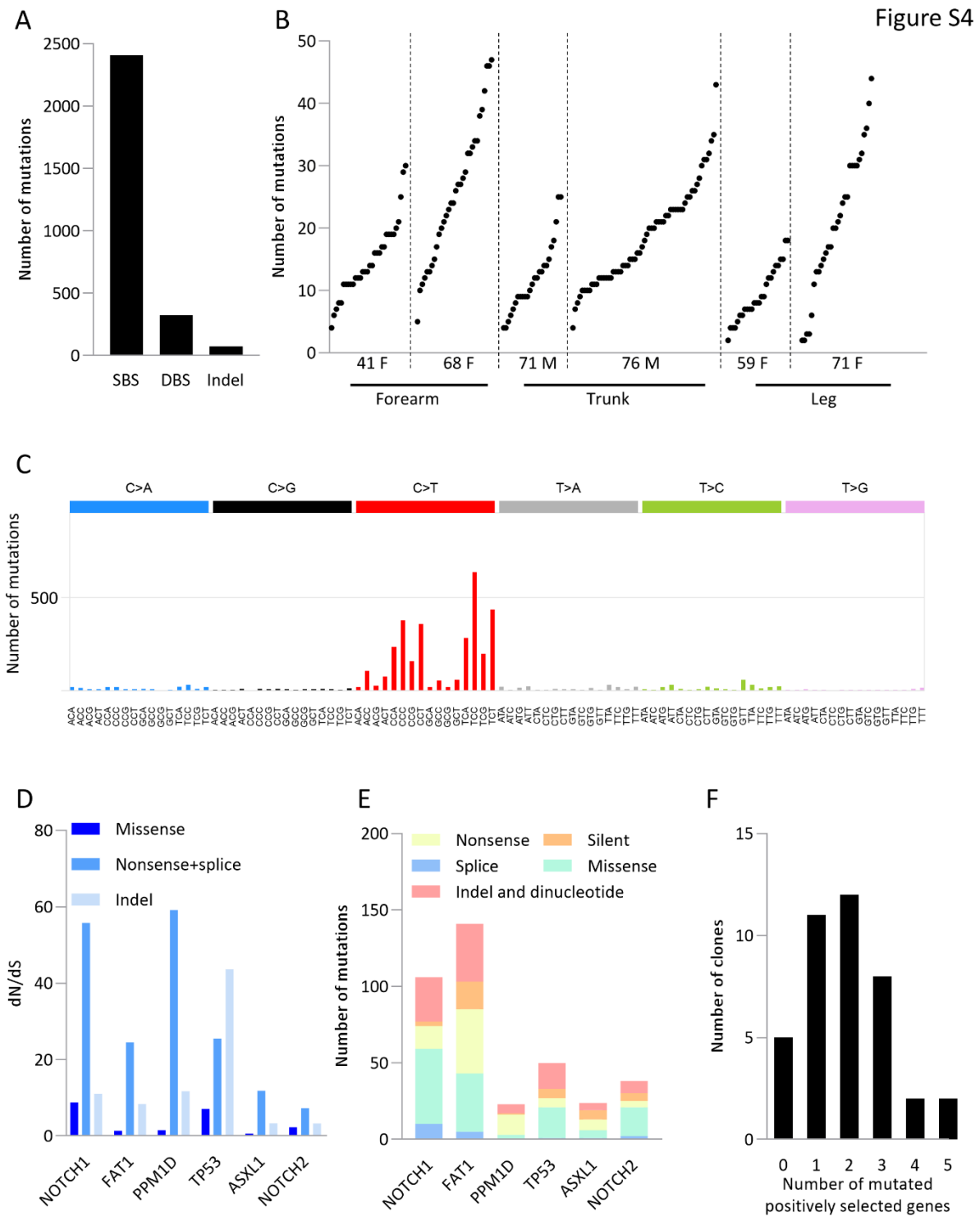

A
CUL3
DICER1
NF1
NOTCH4
NSD1
PIK3CA
Median depth
Position
bp


B
CUL3
Number of mutations
Transcript position


C
GCCGGAAGGAC
GCCGGAAAGAC
Predicted GABPA binding sequence (from Factorbook)
Reference
Silent mutation at position 225449691 (Lys12)
GGAAAGGCCCG
GGAAAAGCCCG
Predicted TAF1 binding sequence (from Factorbook)
Reference
Silent mutation at position 225449667 (Ala20)

Supplemental figure 3 – Negative selection is not due to lower sequencing coverage

a. Sequence coverage per gene of all genes showing negative dN/dS selection. In each case gene structure is shown in the top panel and the median sequence depth per base is shown below. Sequence baits are designed to cover only the exonic regions explaining the lack of sequence depth in the introns.
b. Distribution of number of *CUL3* silent mutations along the transcript length of the gene.
c. Consequence of recurrent *CUL3* silent mutations on predicted transcription factor binding sites.



Supplemental figure 4 – Sequence analysis of data from 0.25mm diameter punches is highly similar to that from 2mm² grid samples

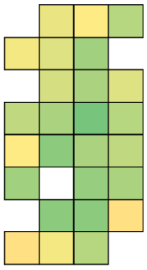
a. Total number of mutations from targeted sequencing of 232 0.25mm diameter punches from 6 donor from different body sites. Number of single base (SBS), double base (DBS) and insertion/ deletions (indel) are shown. Proportions are comparable to that seen in 2mm² gridded sample sequencing.

b. Number of mutations per 0.25mm diameter punch sample. Each point represents a sample. Patients are ordered by age and divided by body site. Age and gender of the patient is shown below.
c. Mutational spectrum of single nucleotide variants from all 0.25mm diameter punch samples from targeted sequencing. The bar plot shows the number of mutations in each of the 96 possible trinucleotide contexts.
d. dN/dS ratios for missense, nonsense/ splice substitutions and insertions/deletions (indel) for all 0.25mm diameter punch samples. Only genes with global $q<0.01$ are shown.
e. Number of different types of mutations for positively selected genes. Proportions are similar to that seen from targeting sequencing from $2mm^2$ gridded samples.
f. Number of positively selected driver genes carrying non-synonymous mutations from whole genome sequencing of clonal 0.25mm diameter punch samples.

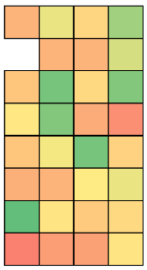
Figure S5

A

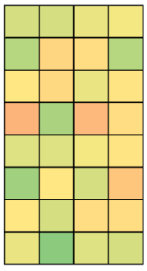
59yo Female leg (PD38331)



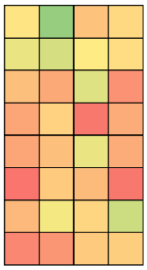
71yo Female leg (PD38334)



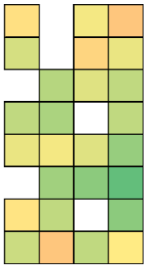
41yo Female forearm (PD37577)



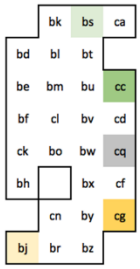
68yo Female forearm (PD36126)

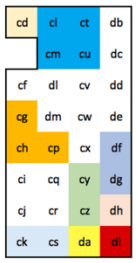


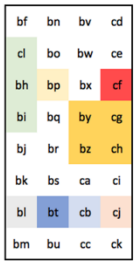
71yo Male trunk (PD38215)

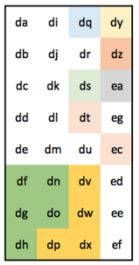


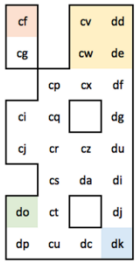
B



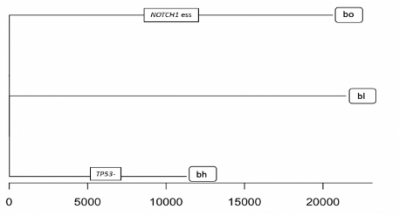




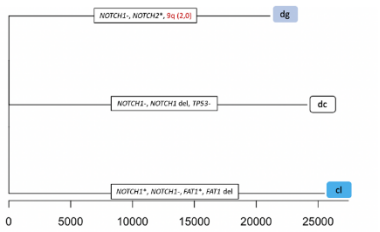




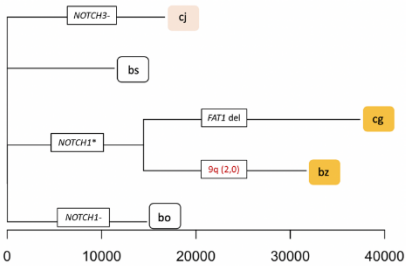
C

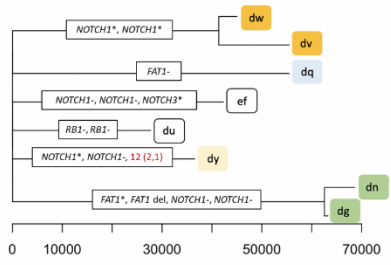


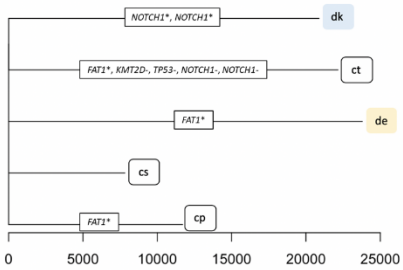
Number of substitutions



Number of substitutions







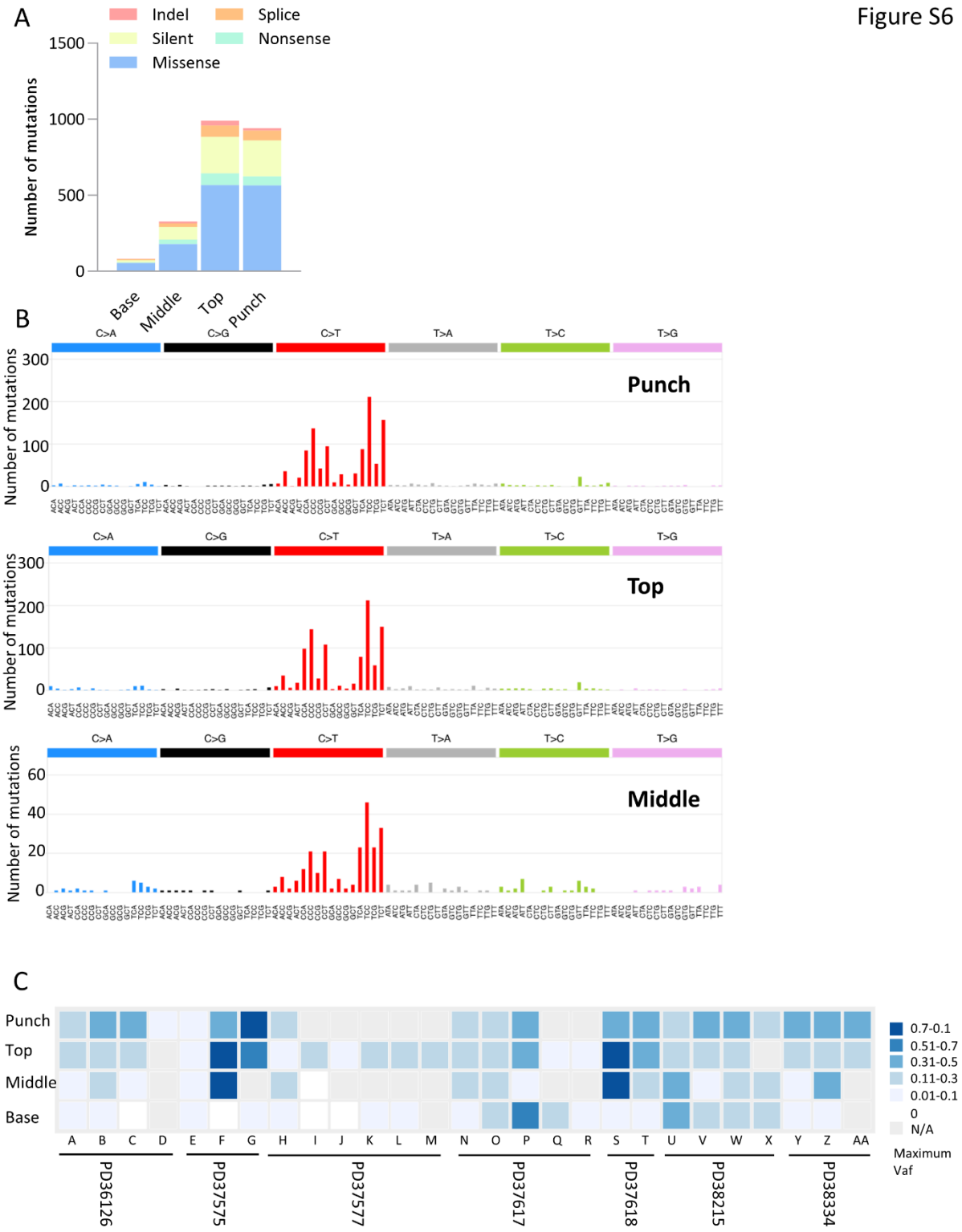
0 Mutations/0.25mm punch 50

* Nonsense	Ess Essential splice
- Missense	CNA

Supplemental figure 5 – Heatmaps and mutational variability of 0.25mm diameter punch samples

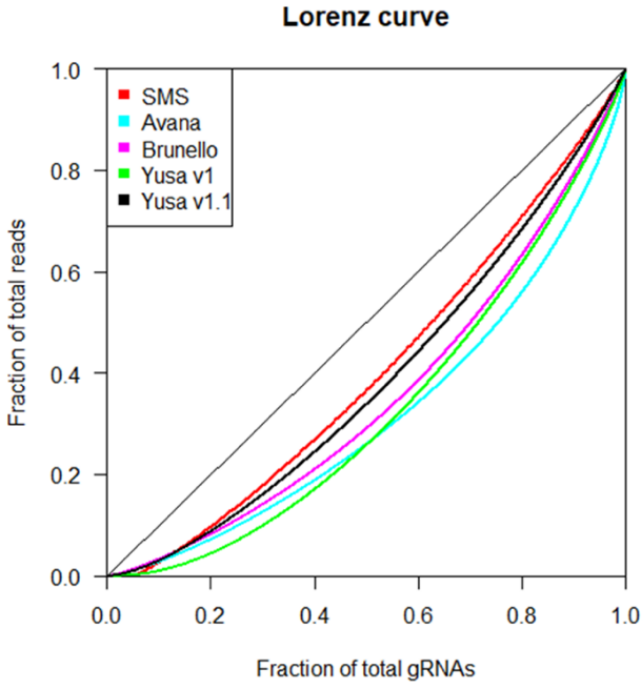
a. Heat map of number of mutations per punch across other five donors. Each square represents an individual 0.25mm diameter sample.
b. Epidermal clone map area for each donor. Mutations with a VAF of at least 0.2 are shown, with each clone represented in a different colour. Polyclonal samples are shown in white.
c. Maximum parsimony tree of clonal substitutions detected in 24 whole-genome 0.25mm diameter punch samples from five additional donors. Branch lengths are equivalent to the number of clonal single and double base substitutions and are annotated with clonal non-synonymous mutations detected in the 13 genes found to be under positive selection. Within each branch, driver mutations are arbitrarily ordered. Copy number alterations are shown in red.

Figure S6


A
Indel
Splice
Silent
Nonsense
Missense
Number of mutations
1500
1000
500
0
Base
Middle
Top
Punch
B
C>A
C>G
C>T
T>A
T>C
T>G
Number of mutations
300
200
100
0
Punch
Top
60
40
20
Middle
C
Punch
Top
Middle
Base
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z AA
PD36126
PD37575
PD37577
PD37617
PD37618
PD38215
PD38334
0.7-0.1
0.51-0.7
0.31-0.5
0.11-0.3
0.01-0.1
0
N/A
Maximum Vaf

Supplemental figure 6 – Mutational spectrum of epidermal 0.25mm diameter punch, top or middle regions of the hair follicle

a. Number of different types of mutations for each part of the follicle.
b. Mutational spectrum for different parts of the follicle, excluding the base where there were too few mutations (<200) inhibiting mutational signature decomposition. The bar plot shows the number of mutations in each of the 96 possible trinucleotide contexts.
c. Maximum variant allele frequency (VAF) of any mutation in each section of the follicle. Each column is a follicle with the patient noted below. Each row is a different part of the follicle.



Supplemental figure 7: Gini index of Crispr/CAS9 screen libraries

The 'ineq' R package was used to generate Lorenz curves and to calculate a Gini index for the generated library and widely used genome-wide libraries. The Gini index is a measurement of statistical dispersion and reflects the evenness of read count distribution. A lower Gini index is indicative of a more even distribution (**Figure S7**).

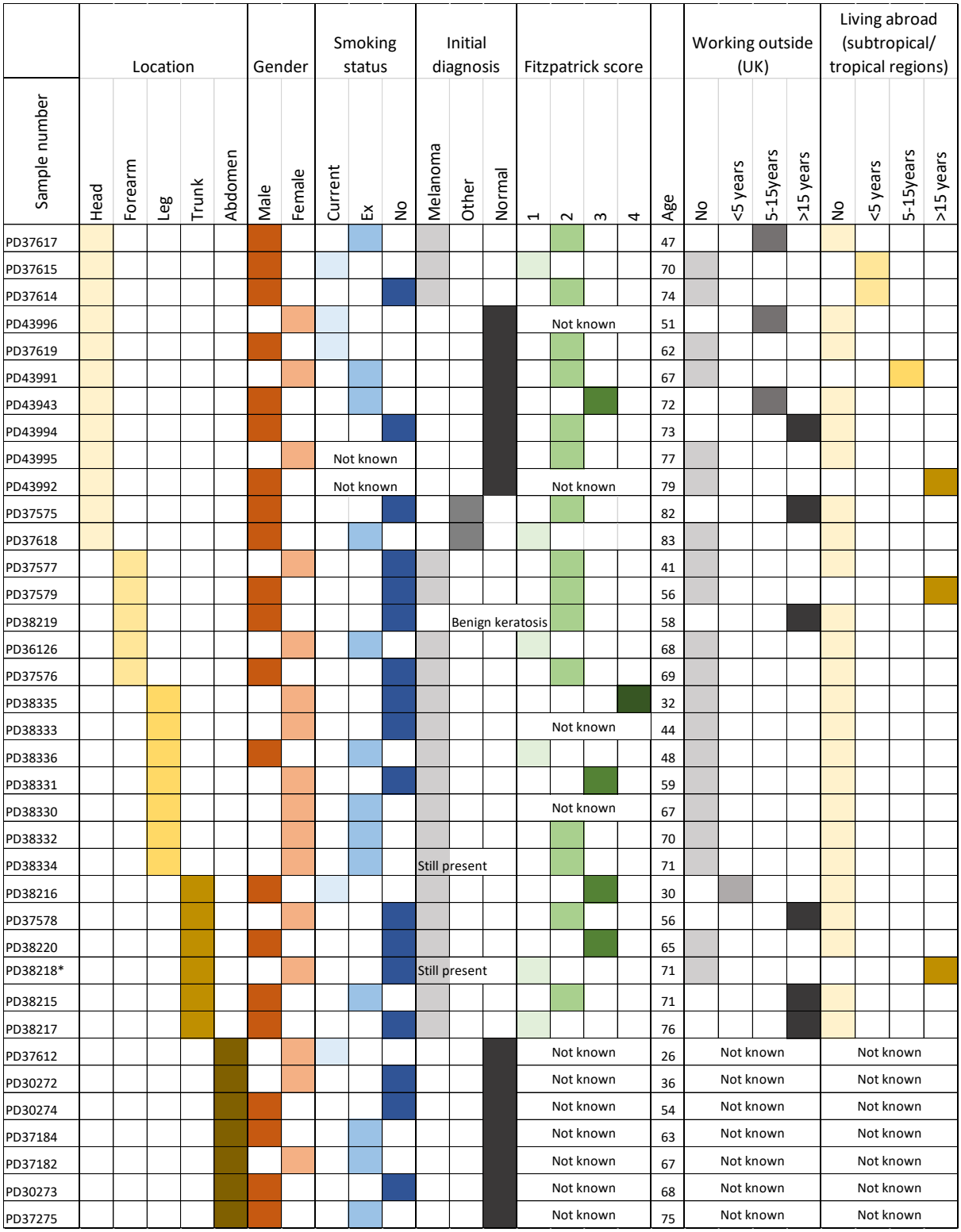


Table S1 – Patient demographics

* Patient PD38218 contributed 3 independent trunk samples

	Mean	Median	Stdev
Targeted			
2mm2 grids			
All	668.7	667.0	111.0
Abdomen	682.7	660.7	146.1
Forearm	595.3	595.4	141.7
Head_melanoma	681.5	677.9	62.5
Head non-melanoma	613.5	603.4	78.6
Leg	727.4	720.4	77.1
Trunk	691.2	695.5	74.8
Punches			
All	44.9	42.3	20.3
Hair follicles			
All	372.3	410.7	158.7
Whole genome sequencing			
2mm2 grids			
All	49.5	45.7	8.1
Punches			
All	33.1	29.8	7.8

Table S2 - Mean sequence coverage of all experiments
Punches are 0.25mm diameter samples

List of excel Tables S3-S8

Table S3 – Lists of genes on targeted sequencing bait sets

Table S4 - List of all mutations detected in all 2mm^2 gridded samples

Table S5 – dN/dS ratios of all genes with q value <0.01 from targeted sequencing of 2mm^2 gridded samples

Table S6 – Fold change in guide RNAs in targeted CRISPR-Cas9 knockout screen

Table S7 – List of all mutations detected in all 0.25mm diameter samples

Table S8 – List of all mutations detected in hair follicle samples