Code S4: Selection analyses

Sigurgeir Olafsson

2/17/2022

Introduction

This document describes the selection analyses carried out as part of the manuscript "Effects of psoriasis and phototreatment on the somatic mutation landscape of the skin" by Sigurgeir Ólafsson et al.

For selection analyses, we used the dNdScv software. Please see https://doi.org/10.1016/j.cell.2017.09.042 (https://doi.org/10.1016/j.cell.2017.09.042) and https://github.com/im3sanger/dndscv (https://github.com/im3sanger/dndscv)

The analyses use mutation calls provided in Supplementary Table 4 of the manuscript. The raw sequencing data has been made publicly available, please see the manuscript for details.

```
.libPaths("/lustre/scratch126/humgen/projects/psoriasis/R_packages_farm5_R4.1.0_install/")

library("seqinr")
library("Biostrings")
library("MASS")
library("GenomicRanges")
library("dndscv")
```

```
all_mutations <- read.table("/nfs/users/nfs_s/so11/phd/psoriasis/bsub_jupyter_lab/psoriasis/manuscript_data_and_f
igures/Supplementary_Table4_all_mutations.txt", h=T)
head(all_mutations)
```

```
##   PatientID         MicrobiopsyID    ClusterID             MutationID  Chr
## 1 patient01    P01L_LL_10,P01L_17 patient01_36  P01:chr1:10018498:A:T chr1
## 2 patient01               P01L_16  patient01_4 P01:chr1:100268476:A:C chr1
## 3 patient01            P01L_LL_13 patient01_28 P01:chr1:100348186:G:A chr1
## 4 patient01               P01L_20 patient01_38 P01:chr1:100743353:G:A chr1
## 5 patient01             P01H_LL_8 patient01_20   P01:chr1:1007997:C:T chr1
## 6 patient01 P01L_LL_11,P01L_LL_14 patient01_33 P01:chr1:100941476:C:T chr1
##    Pos_hg38 Ref Alt type
## 1  10018498   A   T  SBS
## 2 100268476   A   C  SBS
## 3 100348186   G   A  SBS
## 4 100743353   G   A  SBS
## 5   1007997   C   T  SBS
## 6 100941476   C   T  SBS
```

```
table(all_mutations$type)
```

```
##
##    DBS  Indel    SBS
##  54537   3670 582754
```

Exome wide test for selection on the gene-level

I will first conduct an unbiased test for positive selection on the level of individual genes. dNdScv does not annotate double-base mutations (DBSs) but lumps them together with indels under a "no-SNV" mutation class. Many datasets have few DBS mutations so this is not a big issue. As DBS mutations commonly occur as a result of UV-exposure, the current dataset has many such mutations and these need to be accounted for in the modeling.

DBS mutations are much more numerous than indels. I will run a separate negative binomial model for each mutation class. This involves running dNdScv twice, once excluding indels and once excluding DBS mutations. The results for single-base substitutions (SBSs) are unaffected. The P-values from each model can then be combined using Fisher's method.

```
## Format the data a little
all_mutations$Chr <- gsub("chr", "", all_mutations$Chr)
dbs_only <- all_mutations[all_mutations$type!="Indel",c("MicrobiopsyID", "Chr", "Pos_hg38", "Ref", "Alt")]
indel_only <- all_mutations[all_mutations$type!="DBS",c("MicrobiopsyID", "Chr", "Pos_hg38", "Ref", "Alt")]
colnames(dbs_only) = colnames(indel_only) = c("SampleID", "Chr", "Pos", "Ref", "Alt")

## Read in the covariates for dNdScv
covs = "/lustre/scratch126/humgen/projects/psoriasis/resources/covariates_20pc_GRCh37-38.altogether_withoutepiout
.Rdat"
load(covs) # it loads an object called scores

refcds_38 = "/lustre/scratch126/humgen/projects/psoriasis/resources/refcds_GRCh38-GencodeV18+Appris.rda"
```

There are a few samples that truly have a very high mutation burden. I will overwrite the default parameters for `max_muts_per_gene_per_sample` and `max_coding_muts_per_sample` to include all mutations. Below, I show that the inclusion of hypermutators does not affect the results.

```
d38_dbs_only = dndscv(dbs_only, refdb=refcds_38, cv=scores,max_muts_per_gene_per_sample = Inf, max_coding_muts_pe
r_sample = Inf)
d38_indel_only = dndscv(indel_only, refdb=refcds_38, cv=scores,max_muts_per_gene_per_sample = Inf, max_coding_mut
s_per_sample = Inf)


sel_cv_dbs = d38_dbs_only$sel_cv
sel_cv_indel = d38_indel_only$sel_cv
# Results are sorted by significance - sort the dataframes in the same way:
sel_cv_indel = sel_cv_indel[order(sel_cv_indel$gene_name),]
sel_cv_dbs = sel_cv_dbs[order(sel_cv_dbs$gene_name),]

# Fisher combined p-values (single base substitutions, double base substitutions and indels)
p_global <- 1-pchisq(-2 * (log(sel_cv_indel$pallsubs_cv) + log(sel_cv_indel$pind_cv) + log(sel_cv_dbs$pind_cv)),
df = 6)
q_global <- p.adjust(p_global, method="BH")

sel_cv_indel$pdbs_cv <- sel_cv_dbs$pind_cv
sel_cv_indel$wdbs_cv <- sel_cv_dbs$wind_cv
sel_cv_indel$pglobal_cv <- p_global
sel_cv_indel$qglobal_cv <- q_global
sel_cv_indel$n_dbs <- sel_cv_dbs$n_ind

sel_cv_indel = sel_cv_indel[order(sel_cv_indel$pglobal_cv, sel_cv_indel$pallsubs_cv, sel_cv_indel$pmis_cv, sel_cv
_indel$ptrunc_cv, -sel_cv_indel$wmis_cv),] # Sorting genes in the output file

signif_genes <- sel_cv_indel[sel_cv_indel$qglobal_cv < 0.05,]
signif_genes
```

```
##       gene_name n_syn n_mis n_non n_spl n_ind  wmis_cv   wnon_cv   wspl_cv
## 10643    NOTCH1    14   108    33    28    14 7.079368 54.401364 54.401364
## 5513       FAT1    14    53    48    15    31 1.729644 31.579216 31.579216
## 12602     PPM1D     1     0    14     0     2 0.000000 54.060797 54.060797
## 17180      TP53     1    23     5     0     3 8.888560 19.577054 19.577054
## 10644    NOTCH2     7    33     8     2     6 3.140914  9.542485  9.542485
## 2999      CHEK2     2    15     3     3     0 5.782366 16.531743 16.531743
## 6776     GXYLT1     0     7     5     2     1 3.927487 26.464145 26.464145
## 18571   ZFP36L2     6    14     1     0     4 4.244585 12.068969 12.068969
## 4736     EEF1A1     4    11     1     1     1 4.556871  8.637449  8.637449
##         wind_cv      pmis_cv    ptrunc_cv  pallsubs_cv      pind_cv
## 10643 19.795603 0.000000e+00 0.000000e+00 0.000000e+00 1.731752e-08
## 5513  22.792296 1.926672e-02 0.000000e+00 0.000000e+00 4.599866e-13
## 12602  8.771611 3.468602e-02 0.000000e+00 0.000000e+00 2.791332e-02
## 17180 20.968633 2.600571e-10 1.531432e-05 5.375256e-12 9.622535e-04
## 10644 10.378967 4.881740e-05 1.150806e-06 7.853484e-08 4.733263e-04
## 2999   0.000000 7.143712e-06 6.336257e-06 2.075037e-08 1.000000e+00
## 6776   5.703350 8.533551e-03 6.659677e-08 7.329042e-08 1.560063e-01
## 18571 16.934491 1.504069e-04 7.826796e-02 2.304162e-04 4.381137e-04
## 4736   5.472396 3.756794e-04 2.688436e-02 2.879537e-04 1.618246e-01
##            qmis_cv    qtrunc_cv  qallsubs_cv   pglobal_cv   qglobal_cv
## 10643 0.000000e+00 0.0000000000 0.000000e+00 0.000000e+00 0.000000e+00
## 5513  7.862315e-01 0.0000000000 0.000000e+00 0.000000e+00 0.000000e+00
## 12602 7.862315e-01 0.0000000000 0.000000e+00 0.000000e+00 0.000000e+00
## 17180 2.498628e-06 0.0403587599 2.582273e-08 0.000000e+00 0.000000e+00
## 10644 1.595753e-01 0.0044227774 2.155894e-04 4.004908e-12 1.539166e-08
## 2999  4.575785e-02 0.0202929194 7.974781e-05 1.390695e-07 4.453933e-04
## 6776  7.862315e-01 0.0003199309 2.155894e-04 2.132264e-06 5.853369e-03
## 18571 3.211355e-01 0.8213585315 2.951785e-01 3.188282e-06 7.658252e-03
## 4736  5.508156e-01 0.8213585315 3.254894e-01 1.207730e-05 2.578639e-02
##            pdbs_cv   wdbs_cv n_dbs
## 10643 6.807617e-11  7.058805    36
## 5513  2.362045e-04  3.346236    23
## 12602 4.740368e-04  9.110472     5
## 17180 1.545748e-10 17.314951    13
## 10644 1.909303e-04  5.289073    10
## 2999  2.688338e-02  4.978598     3
## 6776  1.000000e+00  0.000000     0
## 18571 1.775103e-01  2.650694     2
## 4736  1.718188e-03  8.876233     4
```

```
# No need to re-write the file every time the document is knitted.
#write.table(sel_cv_indel, file="/nfs/users/nfs_s/so11/phd/psoriasis/bsub_jupyter_lab/psoriasis/manuscript_data_and_figures/sel_cv_dNdS_results.txt", sep="\t", quote=F, row.names = F)
```

Test the effect of excluding the covariates

It's worth checking what effect the covariates have on the above analysis. This is done simply by setting the cv parameter to NULL in dndscv(). The covariates have been extensively tested in TCGA and they generally give superior results. I report the selection analysis *with* covariates in the paper.

```
dbs_noCov = dndscv(dbs_only, refdb=refcds_38, cv=NULL,max_muts_per_gene_per_sample = Inf, max_coding_muts_per_sam
ple = Inf)
indel_noCov = dndscv(indel_only, refdb=refcds_38, cv=NULL,max_muts_per_gene_per_sample = Inf, max_coding_muts_per
_sample = Inf)

sel_cv_dbs_noCov = dbs_noCov$sel_cv
sel_cv_indel_noCov = indel_noCov$sel_cv
# Sort the dataframes in the same way
sel_cv_indel_noCov = sel_cv_indel_noCov[order(sel_cv_indel_noCov$gene_name),]
sel_cv_dbs_noCov = sel_cv_dbs_noCov[order(sel_cv_dbs_noCov$gene_name),]

p_global <- 1-pchisq(-2 * (log(sel_cv_indel_noCov$pallsubs_cv) + log(sel_cv_indel_noCov$pind_cv) + log(sel_cv_dbs
_noCov$pind_cv)), df = 6)
q_global <- p.adjust(p_global, method="BH")

sel_cv_indel_noCov$pdbs_cv <- sel_cv_dbs_noCov$pind_cv
sel_cv_indel_noCov$wdbs_cv <- sel_cv_dbs_noCov$wind_cv
sel_cv_indel_noCov$pglobal_cv <- p_global
sel_cv_indel_noCov$qglobal_cv <- q_global
sel_cv_indel_noCov$n_dbs <- sel_cv_dbs_noCov$n_ind

sel_cv_indel_noCov = sel_cv_indel_noCov[order(sel_cv_indel_noCov$pglobal_cv, sel_cv_indel_noCov$pallsubs_cv, sel_
cv_indel_noCov$pmis_cv, sel_cv_indel_noCov$ptrunc_cv, -sel_cv_indel_noCov$wmis_cv),] # Sorting genes in the outpu
t file

signif_genes_noCov <- sel_cv_indel_noCov[sel_cv_indel_noCov$qglobal_cv < 0.05,]
# Print out genes that are significant in either analysis.
sel_cv_indel_noCov[sel_cv_indel_noCov$gene_name %in% unique(c(signif_genes$gene_name),signif_genes_noCov$gene_nam
e),]
```

```
##       gene_name n_syn n_mis n_non n_spl n_ind   wmis_cv   wnon_cv   wspl_cv
## 10643    NOTCH1    14   108    33    28    14  5.882644 45.205142 45.205142
## 5513       FAT1    14    53    48    15    31  1.941906 35.454616 35.454616
## 12602     PPM1D     1     0    14     0     2  0.000000 42.958067 42.958067
## 17180      TP53     1    23     5     0     3 10.523903 23.178898 23.178898
## 10644    NOTCH2     7    33     8     2     6  2.478091  7.528747  7.528747
## 6776     GXYLT1     0     7     5     2     1  3.846491 25.918381 25.918381
## 18571   ZFP36L2     6    14     1     0     4  3.005842  8.546752  8.546752
## 2999      CHEK2     2    15     3     3     0  4.918302 14.061392 14.061392
## 4736     EEF1A1     4    11     1     1     1  3.523855  6.679390  6.679390
##         wind_cv      pmis_cv    ptrunc_cv  pallsubs_cv      pind_cv
## 10643 18.086176 0.000000e+00 0.000000e+00 0.000000e+00 1.066297e-07
## 5513  22.306077 8.792866e-03 0.000000e+00 0.000000e+00 5.120150e-12
## 12602 10.897753 2.492705e-02 8.437695e-15 0.000000e+00 1.956446e-02
## 17180 25.142277 1.802082e-09 1.261253e-05 8.893708e-11 6.374008e-04
## 10644  8.014609 3.891458e-03 1.721945e-05 2.899751e-05 1.884998e-03
## 6776   7.487571 1.913497e-02 2.983930e-07 9.874783e-07 1.216245e-01
## 18571 26.682982 6.000522e-03 1.167973e-01 9.690697e-03 9.891691e-05
## 2999   0.000000 2.119866e-04 3.105269e-05 3.455001e-06 1.000000e+00
## 4736   7.131791 5.137191e-03 4.782408e-02 5.643228e-03 1.271187e-01
##            qmis_cv    qtrunc_cv  qallsubs_cv   pglobal_cv   qglobal_cv
## 10643 0.0000000000 0.000000e+00 0.000000e+00 0.000000e+00 0.0000000000
## 5513  0.7610978496 0.000000e+00 0.000000e+00 0.000000e+00 0.0000000000
## 12602 0.7695392807 5.404625e-11 0.000000e+00 0.000000e+00 0.0000000000
## 17180 0.0000173144 4.847246e-02 4.272537e-07 0.000000e+00 0.0000000000
## 10644 0.6083139263 5.514816e-02 6.965202e-02 1.645490e-07 0.0006323949
## 6776  0.7695392807 1.433480e-03 3.795077e-03 1.728211e-05 0.0394374564
## 18571 0.7030855804 8.124141e-01 9.291924e-01 1.873004e-05 0.0394374564
## 2999  0.2017609298 7.458856e-02 1.106522e-02 2.052324e-05 0.0394374564
## 4736  0.6969008351 8.124141e-01 8.421045e-01 3.776899e-04 0.5200043269
##            pdbs_cv   wdbs_cv n_dbs
## 10643 1.790345e-11 11.113408    36
## 5513  5.228709e-04  3.954717    23
## 12602 3.173290e-03  6.510330     5
## 17180 3.630909e-11 26.034710    13
## 10644 1.227763e-02  3.191958    10
## 6776  1.000000e+00  0.000000     0
## 18571 1.372586e-01  3.188089     2
## 2999  4.224463e-02  4.351390     3
## 4736  5.847797e-03  6.816864     4
```

Test the effect of hypermutators

Hypermutators can have an effect on selection analyses as they contribute many passengers but relatively few drivers. I want to test the effect of excluding hypermutators. This involves doing two things: First, remove the samples from P34H, which have an incredibly high mutation burden. Second, account for other possible hypermutators by using the default settings for max_muts_per_gene and sample.

```
dbs_noHype <- dbs_only[grep("P34H", dbs_only$SampleID, invert=T),]
indel_noHype <- indel_only[grep("P34H", indel_only$SampleID, invert=T),]

dbs_hype = dndscv(dbs_noHype, refdb=refcds_38, cv=scores)
indel_hype = dndscv(indel_noHype, refdb=refcds_38, cv=scores)

sel_cv_dbs_hype = dbs_hype$sel_cv
sel_cv_indel_hype = indel_hype$sel_cv

sel_cv_indel_hype = sel_cv_indel_hype[order(sel_cv_indel_hype$gene_name),]
sel_cv_dbs_hype = sel_cv_dbs_hype[order(sel_cv_dbs_hype$gene_name),]

p_global <- 1-pchisq(-2 * (log(sel_cv_indel_hype$pallsubs_cv) + log(sel_cv_indel_hype$pind_cv) + log(sel_cv_dbs_h
ype$pind_cv)), df = 6)
q_global <- p.adjust(p_global, method="BH")

sel_cv_indel_hype$pdbs_cv <- sel_cv_dbs_hype$pind_cv
sel_cv_indel_hype$wdbs_cv <- sel_cv_dbs_hype$wind_cv
sel_cv_indel_hype$pglobal_cv <- p_global
sel_cv_indel_hype$qglobal_cv <- q_global
sel_cv_indel_hype$n_dbs <- sel_cv_dbs_hype$n_ind

sel_cv_indel_hype = sel_cv_indel_hype[order(sel_cv_indel_hype$pglobal_cv, sel_cv_indel_hype$pallsubs_cv, sel_cv_i
ndel_hype$pmis_cv, sel_cv_indel_hype$ptrunc_cv, -sel_cv_indel_hype$wmis_cv),] # Sorting genes in the output file

# Print out genes that are significant in either analysis.
signif_genes_hype <- sel_cv_indel_hype[sel_cv_indel_hype$qglobal_cv<0.05,]
sel_cv_indel_hype[sel_cv_indel_hype$gene_name %in% unique(c(signif_genes$gene_name),signif_genes_hype$gene_name),
]
```

```
##       gene_name n_syn n_mis n_non n_spl n_ind  wmis_cv   wnon_cv   wspl_cv
## 10643    NOTCH1    14   108    32    28    14 7.120927 54.003918 54.003918
## 5513       FAT1    14    50    48    14    31 1.650095 31.557558 31.557558
## 12602     PPM1D     1     0    13     0     2 0.000000 50.994060 50.994060
## 17180      TP53     1    23     5     0     3 8.954853 19.897862 19.897862
## 10644    NOTCH2     7    33     8     2     6 3.171931  9.709661  9.709661
## 2999      CHEK2     2    15     3     3     0 5.878076 17.067400 17.067400
## 6776     GXYLT1     0     7     5     2     1 4.016087 27.125340 27.125340
## 18571   ZFP36L2     6    14     1     0     4 4.304309 12.275375 12.275375
## 4736     EEF1A1     4    11     1     1     1 4.625159  8.836645  8.836645
##         wind_cv      pmis_cv    ptrunc_cv  pallsubs_cv      pind_cv
## 10643 19.924305 0.000000e+00 0.000000e+00 0.000000e+00 1.524526e-08
## 5513  22.959330 3.482043e-02 0.000000e+00 0.000000e+00 3.484356e-13
## 12602  8.849431 3.559521e-02 2.220446e-16 0.000000e+00 2.743666e-02
## 17180 21.413216 2.260823e-10 1.412898e-05 4.376943e-12 9.023988e-04
## 10644 10.470245 4.183010e-05 9.855678e-07 6.065391e-08 4.472788e-04
## 2999   0.000000 5.943054e-06 5.267843e-06 1.486805e-08 1.000000e+00
## 6776   5.755292 7.607241e-03 5.615772e-08 5.716063e-08 1.547987e-01
## 18571 16.992137 1.312106e-04 7.679084e-02 1.999518e-04 4.286992e-04
## 4736   5.503878 3.329499e-04 2.567056e-02 2.483905e-04 1.610536e-01
##            qmis_cv    qtrunc_cv  qallsubs_cv   pglobal_cv   qglobal_cv
## 10643 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## 5513  7.874679e-01 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## 12602 7.874679e-01 1.422270e-12 0.000000e+00 0.000000e+00 0.000000e+00
## 17180 2.172198e-06 3.807097e-02 2.102684e-08 0.000000e+00 0.000000e+00
## 10644 1.607615e-01 3.787734e-03 1.665037e-04 2.923328e-12 1.123494e-08
## 2999  3.806724e-02 1.687115e-02 5.714088e-05 1.022645e-07 3.275192e-04
## 6776  7.874679e-01 2.697817e-04 1.665037e-04 1.694158e-06 4.650704e-03
## 18571 2.801492e-01 8.234364e-01 2.708333e-01 2.757423e-06 6.623330e-03
## 4736  4.921512e-01 8.234364e-01 2.983170e-01 1.051551e-05 2.245179e-02
##            pdbs_cv   wdbs_cv n_dbs
## 10643 6.551400e-11  7.059927    36
## 5513  2.293538e-04  3.352013    23
## 12602 4.731561e-04  9.110320     5
## 17180 1.516628e-10 17.324706    13
## 10644 1.872143e-04  5.299507    10
## 2999  2.677542e-02  4.985733     3
## 6776  1.000000e+00  0.000000     0
## 18571 1.777303e-01  2.648388     2
## 4736  1.711674e-03  8.882732     4
```

Restricting the analysis to lesional skin

We may wonder if pooling samples from lesional and non-lesional skin impacts the analysis. The following analysis shows that no recurrently mutated genes are missed by pooling samples but as the sample size is smaller and there are fewer mutations, some genes no longer reach significance.

```
dbs_lesional <- dbs_only[!grepl("H", dbs_only$SampleID),]
d38_dbs_lesional = dndscv(dbs_lesional, refdb=refcds_38, cv=scores,max_muts_per_gene_per_sample = Inf, max_coding
_muts_per_sample = Inf)
sel_cv_dbs = d38_dbs_lesional$sel_cv

indel_lesional <- indel_only[!grepl("H", indel_only$SampleID),]
d38_indel_lesional = dndscv(indel_lesional, refdb=refcds_38, cv=scores,max_muts_per_gene_per_sample = Inf, max_co
ding_muts_per_sample = Inf)
sel_cv_indel_lesional = d38_indel_lesional$sel_cv


sel_cv_indel_lesional = sel_cv_indel_lesional[order(sel_cv_indel_lesional$gene_name),]
sel_cv_dbs = sel_cv_dbs[order(sel_cv_dbs$gene_name),]

p_global <- 1-pchisq(-2 * (log(sel_cv_indel_lesional$pallsubs_cv) + log(sel_cv_indel_lesional$pind_cv) + log(sel_
cv_dbs$pind_cv)), df = 6)
q_global <- p.adjust(p_global, method="BH")

sel_cv_indel_lesional$pdbs_cv <- sel_cv_dbs$pind_cv
sel_cv_indel_lesional$wdbs_cv <- sel_cv_dbs$wind_cv
sel_cv_indel_lesional$pglobal_cv <- p_global
sel_cv_indel_lesional$qglobal_cv <- q_global
sel_cv_indel_lesional$n_dbs <- sel_cv_dbs$n_ind

sel_cv_indel_lesional = sel_cv_indel_lesional[order(sel_cv_indel_lesional$pglobal_cv, sel_cv_indel_lesional$palls
ubs_cv, sel_cv_indel_lesional$pmis_cv, sel_cv_indel_lesional$ptrunc_cv, -sel_cv_indel_lesional$wmis_cv),] # Sorti
ng genes in the output file

signif_genes_lesional <- sel_cv_indel_lesional[sel_cv_indel_lesional$qglobal_cv < 0.05,]

sel_cv_indel_lesional[sel_cv_indel_lesional$gene_name %in% unique(c(signif_genes$gene_name),signif_genes_lesional
$gene_name),]
```

```
##       gene_name n_syn n_mis n_non n_spl n_ind  wmis_cv  wnon_cv  wspl_cv
## 10643    NOTCH1    13    84    27    18    11 7.019856 51.51734 51.51734
## 5513       FAT1     9    36    38    12    19 1.690002 36.58948 36.58948
## 17180      TP53     1    16     4     0     1 8.099085 20.79898 20.79898
## 12602     PPM1D     1     0    10     0     2 0.000000 51.12939 51.12939
## 10644    NOTCH2     5    22     6     2     4 2.949814 10.89947 10.89947
## 2999      CHEK2     2    12     3     3     0 6.078629 22.15425 22.15425
## 4736     EEF1A1     3     8     1     1     1 4.581142 12.10417 12.10417
## 18571   ZFP36L2     4    10     1     0     3 4.393832 17.65646 17.65646
## 6776     GXYLT1     0     4     3     2     1 2.989310 25.40303 25.40303
##         wind_cv      pmis_cv    ptrunc_cv  pallsubs_cv      pind_cv
## 10643 20.663841 0.000000e+00 0.000000e+00 0.000000e+00 2.979685e-08
## 5513  17.795820 5.082614e-02 0.000000e+00 0.000000e+00 3.882490e-10
## 17180  9.683031 8.124996e-08 7.504527e-05 1.903884e-09 9.673791e-02
## 12602 11.805702 6.162730e-02 2.944311e-13 1.342260e-13 1.559047e-02
## 10644  8.947100 7.827596e-04 4.379900e-06 1.152512e-06 3.101438e-03
## 2999   0.000000 2.127829e-05 1.116898e-06 8.017294e-09 1.000000e+00
## 4736   7.534266 1.755221e-03 1.355400e-02 6.004712e-04 1.220910e-01
## 18571 18.396571 7.680201e-04 5.118730e-02 7.126999e-04 1.184621e-03
## 6776   7.661010 8.596480e-02 4.364987e-06 1.159419e-05 1.202329e-01
##            qmis_cv    qtrunc_cv  qallsubs_cv   pglobal_cv   qglobal_cv
## 10643 0.0000000000 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## 5513  0.8027364762 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## 17180 0.0007806496 1.442070e-01 9.146260e-06 1.332268e-15 8.533618e-12
## 12602 0.8027364762 1.885930e-09 8.597620e-10 1.023626e-13 4.917498e-10
## 10644 0.6267295131 1.402736e-02 3.691112e-03 2.628387e-09 1.010142e-05
## 2999  0.0798764075 5.365578e-03 3.081206e-05 1.618688e-07 5.184117e-04
## 4736  0.7847006786 8.459965e-01 5.357220e-01 5.848092e-06 1.605385e-02
## 18571 0.6267295131 8.459965e-01 5.629391e-01 4.214596e-05 8.998630e-02
## 6776  0.8027364762 1.402736e-02 3.182770e-02 1.469048e-04 2.566293e-01
##            pdbs_cv   wdbs_cv n_dbs
## 10643 3.983118e-11  8.026094    31
## 5513  8.123731e-04  3.320814    17
## 17180 8.442670e-09 17.785926    10
## 12602 6.963941e-02  4.770846     2
## 10644 2.071016e-03  4.762052     7
## 2999  8.220890e-02  4.321640     2
## 4736  4.824407e-04 12.667308     4
## 18571 3.923593e-01  1.962829     1
## 6776  1.000000e+00  0.000000     0
```

Restricting the analysis to non-lesional skin

We find positive selection in a few genes that have not been previously reported for normal skin. The question we want to answer is if there is evidence for mutations in these genes being positively selected in non-lesional skin. If there is then that is evidence that these mutations have little to do with psoriasis itself.

```
dbs_nonLes <- dbs_only[grepl("H", dbs_only$SampleID),]
d38_dbs_nonLes = dndscv(dbs_nonLes, refdb=refcds_38, cv=scores,max_muts_per_gene_per_sample = Inf, max_coding_mut
s_per_sample = Inf)
sel_cv_dbs = d38_dbs_nonLes$sel_cv

indel_nonLes <- indel_only[grepl("H", indel_only$SampleID),]
d38_indel_nonLes = dndscv(indel_nonLes, refdb=refcds_38, cv=scores,max_muts_per_gene_per_sample = Inf, max_coding
_muts_per_sample = Inf)
sel_cv_indel_nonLes = d38_indel_nonLes$sel_cv


sel_cv_indel_nonLes = sel_cv_indel_nonLes[order(sel_cv_indel_nonLes$gene_name),]
sel_cv_dbs = sel_cv_dbs[order(sel_cv_dbs$gene_name),]

p_global <- 1-pchisq(-2 * (log(sel_cv_indel_nonLes$pallsubs_cv) + log(sel_cv_indel_nonLes$pind_cv) + log(sel_cv_d
bs$pind_cv)), df = 6)
q_global <- p.adjust(p_global, method="BH")

sel_cv_indel_nonLes$pdbs_cv <- sel_cv_dbs$pind_cv
sel_cv_indel_nonLes$wdbs_cv <- sel_cv_dbs$wind_cv
sel_cv_indel_nonLes$pglobal_cv <- p_global
sel_cv_indel_nonLes$qglobal_cv <- q_global
sel_cv_indel_nonLes$n_dbs <- sel_cv_dbs$n_ind

sel_cv_indel_nonLes = sel_cv_indel_nonLes[order(sel_cv_indel_nonLes$pglobal_cv, sel_cv_indel_nonLes$pallsubs_cv,
sel_cv_indel_nonLes$pmis_cv, sel_cv_indel_nonLes$ptrunc_cv, -sel_cv_indel_nonLes$wmis_cv),] # Sorting genes in th
e output file

signif_genes_nonLes <- sel_cv_indel_nonLes[sel_cv_indel_nonLes$qglobal_cv < 0.05,]

sel_cv_indel_nonLes[sel_cv_indel_nonLes$gene_name %in% c("CHEK2", "GXYLT1", "ZFP36L2", "EEF1A1"),]
```

```
##       gene_name n_syn n_mis n_non n_spl n_ind  wmis_cv wnon_cv wspl_cv  wind_cv
## 18571   ZFP36L2     2     4     0     0     1 5.727376  0.0000  0.0000 13.69636
## 6776     GXYLT1     0     3     2     0     0 6.131248 26.8669 26.8669  0.00000
## 2999      CHEK2     0     3     0     0     0 5.137455  0.0000  0.0000  0.00000
## 4736     EEF1A1     1     3     0     0     0 6.081441  0.0000  0.0000  0.00000
##          pmis_cv   ptrunc_cv pallsubs_cv    pind_cv   qmis_cv qtrunc_cv
## 18571 0.01088410 0.849444047 0.038185310 0.06792245 0.7711319 0.9213464
## 6776  0.02221642 0.002823397 0.001278334 1.00000000 0.7711319 0.9213464
## 2999  0.03578235 0.678473337 0.099061031 1.00000000 0.7711319 0.9213464
## 4736  0.02198161 0.754013666 0.068235954 1.00000000 0.7711319 0.9213464
##       qallsubs_cv pglobal_cv qglobal_cv   pdbs_cv  wdbs_cv n_dbs
## 18571   0.9396631 0.02054961   0.999999 0.2172717 4.044568     1
## 6776    0.9396631 0.03816420   0.999999 1.0000000 0.000000     0
## 2999    0.9396631 0.19297844   0.999999 0.1322319 7.013037     1
## 4736    0.9396631 0.49735927   0.999999 1.0000000 0.000000     0
```

Restricted hypothesis testing

Not all genes are equally likely to be under positive selection in the skin. The above analysis assumes we know nothing of what to expect, but in reality we do. Past studies of skin and oesophagous have identified a number of recurrently mutated genes and we may be interested in seeing if there is evidence of positive selection of mutations in those genes in the current dataset.

I have compiled a list of genes from Fowler et a. https://doi.org/10.1158/2159-8290.CD-20-1092 (https://doi.org/10.1158/2159-8290.CD-20-1092), Martincorena et al https://doi.org/10.1126/science.aau3879 (https://doi.org/10.1126/science.aau3879) and Yokoyama et al https://doi.org/10.1038/s41586-018-0811-x (https://doi.org/10.1038/s41586-018-0811-x).

The restricted hypothesis testing basically just involves only changing the q-values to reflect less multiple testing. Even in with much reduced burden of multiple testing, no evidence for selection in these genes is found.

```
rht_genes <- read.table("/nfs/users/nfs_s/so11/phd/psoriasis/bsub_jupyter_lab/psoriasis/08_selection_analyses/rht
_genes.txt")

sel_cv_rht <- sel_cv_indel[sel_cv_indel$gene_name %in% rht_genes$V1, ]
sel_cv_rht$qrht <- p.adjust(sel_cv_rht$pglobal_cv, method = "BH")
sel_cv_rht <- sel_cv_rht[order(sel_cv_rht$qrht),]
sel_cv_rht
```

```
##       gene_name n_syn n_mis n_non n_spl n_ind   wmis_cv   wnon_cv   wspl_cv
## 8302      KMT2D    17    30     7     3     0 0.8112327 3.8393209 3.8393209
## 1004      ARID2     1    12     5     0     1 1.5816826 6.8840196 6.8840196
```

```
## 13047     PTCH1     4    11     1     1     1 1.5675698 3.0423208 3.0423208
## 17197      TP63     1    12     0     2     0 3.5057851 5.2269680 5.2269680
## 10649    NOTCH3     6    25     3     3     1 1.7705010 5.5652371 5.5652371
## 3583     CREBBP     1    18     1     0     2 2.1309025 1.1990655 1.1990655
## 8089      KDM6A     1     2     0     0     2 0.4237191 0.0000000 0.0000000
## 12140    PIK3CA     4    14     0     0     0 2.2702234 0.0000000 0.0000000
## 1002     ARID1A     6     8     0     0     1 0.6832892 0.0000000 0.0000000
## 2487      CCND1     0     0     0     0     0 0.0000000 0.0000000 0.0000000
## 5001      EP300     6    21     1     0     1 1.7900089 0.8194030 0.8194030
## 12371    PLXNB2    10    10     0     0     2 1.1013011 0.0000000 0.0000000
## 4299     DICER1     2    15     0     0     0 1.7617321 0.0000000 0.0000000
## 10650    NOTCH4     4    13     0     0     1 1.1851198 0.0000000 0.0000000
## 490       AJUBA     1     5     1     0     0 1.9671710 6.6530024 6.6530024
## 16569    TGFBR2     3     1     0     1     0 0.3610332 3.8449175 3.8449175
## 19070    ZNF750     1     5     0     0     0 1.3716952 0.0000000 0.0000000
## 3822       CUL3     2     5     0     0     0 1.4922922 0.0000000 0.0000000
## 5596      FBXW7     0     2     0     0     0 0.7533226 0.0000000 0.0000000
## 10793      NSD1    11    11     1     0     0 0.5969390 0.6359945 0.6359945
## 13431       RB1     0     5     1     1     0 1.3947360 3.8404894 3.8404894
## 10448    NFE2L2     0     2     1     0     0 0.9031943 5.9928554 5.9928554
## 11693      PAX9     0     3     0     0     0 1.9836328 0.0000000 0.0000000
##        wind_cv     pmis_cv   ptrunc_cv pallsubs_cv    pind_cv   qmis_cv
## 8302  0.000000 0.423747277 0.001199005 0.001157984 1.00000000 0.8687378
## 1004  1.423183 0.257761127 0.002043402 0.007654828 0.46265517 0.8078884
## 13047 2.160820 0.260966230 0.197869932 0.278123848 0.34656500 0.8078884
## 17197 0.000000 0.002647366 0.074156242 0.004005285 1.00000000 0.6909816
## 10649 1.132877 0.063313739 0.001983061 0.004174063 0.53174567 0.7862315
## 3583  2.716761 0.037013443 0.864723967 0.111506158 0.17703278 0.7862315
## 8089  4.576859 0.212106319 0.266448198 0.271713827 0.08261487 0.8077863
## 12140 0.000000 0.028949064 0.164511967 0.024510496 1.00000000 0.7862315
## 1002  1.328253 0.361009861 0.188262736 0.306972773 0.48327636 0.8404194
## 2487  0.000000 0.167781695 0.687645375 0.358486997 1.00000000 0.8001252
## 5001  1.126011 0.070785906 0.841741268 0.172128507 0.53361609 0.7862315
## 12371 2.693266 0.801855841 0.203629160 0.419686169 0.17910967 0.9753646
## 4299  0.000000 0.130536871 0.185651615 0.101434727 1.00000000 0.7862315
## 10650 1.844645 0.651615512 0.162683133 0.312238304 0.38851229 0.9467775
## 490   0.000000 0.224889286 0.153632455 0.194397853 1.00000000 0.8077863
## 16569 0.000000 0.251218596 0.277827268 0.265169546 1.00000000 0.8077863
## 19070 0.000000 0.558988410 0.564925155 0.704452851 1.00000000 0.9170269
## 3822  0.000000 0.458455231 0.369679601 0.488034486 1.00000000 0.8832641
## 5596  0.000000 0.701960137 0.388635400 0.651215126 1.00000000 0.9589213
## 10793 0.000000 0.149819003 0.634541079 0.342711928 1.00000000 0.7864592
## 13431 0.000000 0.543112939 0.133456394 0.298083414 1.00000000 0.9130652
## 10448 0.000000 0.892865787 0.172885461 0.384900735 1.00000000 0.9916753
## 11693 0.000000 0.323956512 0.654613139 0.550002454 1.00000000 0.8265948
##       qtrunc_cv qallsubs_cv pglobal_cv qglobal_cv    pdbs_cv   wdbs_cv n_dbs
## 8302  0.8213585   0.7178008 0.01021164          1 0.19822676 1.3771606    16
## 1004  0.8213585   0.9297071 0.02758551          1 0.23416021 1.8222291     3
## 13047 0.8213585   0.9297071 0.03379491          1 0.01126764 3.3536965     7
## 17197 0.8213585   0.9297071 0.04614439          1 0.41265083 1.8363818     1
## 10649 0.8213585   0.9297071 0.04771649          1 0.77942867 0.6645653     3
## 3583  0.8824501   0.9297071 0.07193081          1 0.15463631 1.7608003     6
## 8089  0.8213585   0.9297071 0.11508058          1 0.26682758 1.9842640     2
## 12140 0.8213585   0.9297071 0.28397151          1 1.00000000 0.0000000     0
## 1002  0.8213585   0.9297071 0.31735009          1 0.19972340 1.6754952     5
## 2487  0.8213585   0.9297071 0.32699631          1 0.08706286 4.1714187     2
## 5001  0.8653844   0.9297071 0.48469877          1 0.70537923 0.7617713     2
## 12371 0.8213585   0.9297071 0.51350321          1 0.96883266 0.2473016     1
## 4299  0.8213585   0.9297071 0.55880531          1 0.85766380 0.4709775     1
## 10650 0.8213585   0.9297071 0.62488659          1 0.92082930 0.3527044     1
## 490   0.8213585   0.9297071 0.62698658          1 0.57913019 1.1128618     1
## 16569 0.8213585   0.9297071 0.62807571          1 0.42629012 1.7569310     1
## 19070 0.8213585   0.9297071 0.63630646          1 0.16545076 2.2013777     3
## 3822  0.8213585   0.9297071 0.81817569          1 0.47462644 1.5108812     1
## 5596  0.8213585   0.9297071 0.83875448          1 0.38703941 2.0002260     1
## 10793 0.8213585   0.9297071 0.86603150          1 0.82697257 0.5711352     2
## 13431 0.8213585   0.9297071 0.87722792          1 1.00000000 0.0000000     0
## 10448 0.8213585   0.9297071 0.92782728          1 1.00000000 0.0000000     0
## 11693 0.8213585   0.9297071 0.97709829          1 1.00000000 0.0000000     0
##           qrht
## 8302  0.2194959
## 1004  0.2194959
## 13047 0.2194959
## 17197 0.2194959
## 10649 0.2194959
## 3583  0.2757348
## 8089  0.3781219
## 12140 0.7520915
## 1002  0.7520915
```

```
## 2487  0.7520915
## 5001  0.8608852
## 12371 0.8608852
## 4299  0.8608852
## 10650 0.8608852
## 490   0.8608852
## 16569 0.8608852
## 19070 0.8608852
## 3822  0.9607734
## 5596  0.9607734
## 10793 0.9607734
## 13431 0.9607734
## 10448 0.9700012
## 11693 0.9770983
```

Site- dN/dS

I want to run site-wise dN/dS to identify mutation hotspots that may be found in genes that do not reach significance on the gene level. The following analysis shows that only NOTCH1 E455K reaches significance.

```
d38_sites = dndscv(dbs_only, refdb=refcds_38, cv=scores,max_muts_per_gene_per_sample = Inf, max_coding_muts_per_s
ample = Inf, outmats=T)
sites=sitednds(d38_sites)
head(sites$recursites)
```

```
##   chr       pos ref mut      gene aachange     impact ref3_cod mut3_cod freq
## 1   9 136517830   C   T    NOTCH1    E455K   Missense      CGA      CAA    7
## 2   9 136518238   G   A    NOTCH1    S385F   Missense      TCC      TTC    5
## 3  20  33663730   G   A C20orf144    E109K   Missense      AGA      AAA    4
## 4  12 120738875   G   A     ACADS    R330H   Missense      CGC      CAC    3
## 5   2 179483369   C   T   ZNF385B    T206T Synonymous      CGA      CAA    5
## 6  12  85980280   C   T    MGAT4C    R149H   Missense      CGT      CAT    4
##            mu      dnds         pval         qval
## 1 0.010021673  698.4861 1.344934e-13 1.376329e-05
## 2 0.012524985  399.2021 1.508268e-09 6.454599e-02
## 3 0.004917534  813.4159 1.892209e-09 6.454599e-02
## 4 0.001163337 2578.7883 3.302796e-09 7.516383e-02
## 5 0.015244463  327.9879 3.941409e-09 7.516383e-02
## 6 0.006086329  657.2106 4.406956e-09 7.516383e-02
```

Pathway-level dN/dS

In our recent manuscript "Somatic evolution in non-neoplastic IBD-affected colon" https://doi.org/10.1016/j.cell.2020.06.036 (https://doi.org/10.1016/j.cell.2020.06.036), we found that genes in the IL-17 and TLR pathways were enriched in somatic mutations in the colonic mucosa, even though no individual genes reached significance. We are interested in seeing if we can see evidence of positive selection on the pathway-level.

The mutation spectra for both UV-light and psoralen exposure extend beyond the trinucleotide model. This doesn't make too much of a difference on a gene level but it makes sense to implement a pentanucleotide model when considering genes in aggregate.

Please see the R-script pathway_dnds_pentanuc_model.r for details on how this was done. The script can be found in the Github repository accompanying this manuscript (see main text).

```
geneLists <- read.table("/lustre/scratch126/humgen/projects/psoriasis/selection_analyses/pathway_dNdS_geneLists.t
xt")

df <- data.frame()
for(geneL in geneLists$V1) {
  results <- read.table(paste("/lustre/scratch126/humgen/projects/psoriasis/selection_analyses/pathway_pentamodel
/", geneL, "_Full3075_1x2w_model_dNdSvals.txt", sep=""),h=T)
  results$pathway <- geneL
  df <- rbind(df, results)
}

df <- df[df$omega %in% c("wmis_driv", "wnon_driv"),]
df <- df[df$omega!="r_drivpass",]
df$q <- p.adjust(df$P, method="BH")
df
```

```
##        omega       MLEs      lowbd     highbd            P         pathway
## 1  wmis_driv  2.2122135  1.7260880  2.8352486 3.574584e-10 Normal_skin_pos
## 2  wnon_driv 15.1478793 11.4118277 20.1070551 6.369498e-79 Normal_skin_pos
## 8  wmis_driv  1.3521823  0.9481951  1.9282920 9.567306e-02             BCC
## 9  wnon_driv  2.0523151  1.0579729  3.9811958 3.344857e-02             BCC
## 15 wmis_driv  1.3442398  1.0276268  1.7584015 3.086267e-02  GWAS_psoriasis
## 16 wnon_driv  0.9190429  0.4436065  1.9040295 8.202964e-01  GWAS_psoriasis
## 22 wmis_driv  0.9146156  0.7195190  1.1626124 4.659323e-01            IL17
## 23 wnon_driv  0.7706908  0.4032517  1.4729367 4.306066e-01            IL17
## 29 wmis_driv  1.0720342  0.7882336  1.4580160 6.575299e-01         IL12_23
## 30 wnon_driv  1.2761316  0.6541418  2.4895396 4.745208e-01         IL12_23
## 36 wmis_driv  0.9657801  0.7262132  1.2843767 8.108145e-01             TNF
## 37 wnon_driv  0.9268954  0.4638567  1.8521561 8.298205e-01             TNF
## 43 wmis_driv  0.7897253  0.6364000  0.9799907 3.207389e-02      IL36_MyD88
## 44 wnon_driv  0.6972013  0.3866221  1.2572733 2.305544e-01      IL36_MyD88
## 50 wmis_driv  1.1340048  0.9251650  1.3899865 2.259126e-01            IFNg
## 51 wnon_driv  0.9961984  0.6026683  1.6466954 9.881488e-01            IFNg
## 57 wmis_driv  1.0128463  0.9191600  1.1160817 7.965931e-01      MHC_classI
## 58 wnon_driv  1.2348756  1.0040527  1.5187627 4.568710e-02      MHC_classI
## 64 wmis_driv  0.9012028  0.7783051  1.0435065 1.643321e-01             TLR
## 65 wnon_driv  0.6657081  0.4388242  1.0098971 5.566522e-02             TLR
## 71 wmis_driv  1.2560263  0.8493946  1.8573253 2.534041e-01      IBD_mucosa
## 72 wnon_driv  0.9814744  0.3492549  2.7581346 9.717045e-01      IBD_mucosa
##               q
## 1  3.932042e-09
## 2  1.401290e-77
## 8  2.631009e-01
## 9  1.471737e-01
## 15 1.471737e-01
## 16 9.128025e-01
## 22 6.959638e-01
## 23 6.959638e-01
## 29 9.041037e-01
## 30 6.959638e-01
## 36 9.128025e-01
## 37 9.128025e-01
## 43 1.471737e-01
## 44 4.611087e-01
## 50 4.611087e-01
## 51 9.881488e-01
## 57 9.128025e-01
## 58 1.675194e-01
## 64 4.017008e-01
## 65 1.749478e-01
## 71 4.645741e-01
## 72 9.881488e-01
```

Effects of Psoralen exposure

There is some literature (https://www.cell.com/cell-stem-cell/pdfExtended/S1934-5909(18)30402-8) out there to suggest that the selection of TP53 in particular is affected by UV-light exposure. We find no evidence that TP53 or any other gene is particulary selected for in the Psoralen-exposed skin. Please remember that the absence of evidence is not evidence of absence and we may simply lack power to detect differences in selection between psoralen-exposed and non-exposed skin.

```
cluster_burden <- read.table("/nfs/users/nfs_s/so11/phd/psoriasis/bsub_jupyter_lab/psoriasis/manuscript_data_and_
figures/Supplementary_material/Supplementary_Table3_clone_mutationBurden.txt", h=T)

## Mutations in samples with high PUVA exposure
puva_muts <- all_mutations[all_mutations$ClusterID %in% cluster_burden$CloneID[cluster_burden$PUVA>100],c("Microb
iopsyID", "Chr", "Pos_hg38", "Ref", "Alt")]

colnames(puva_muts) = c("SampleID", "Chr", "Pos", "Ref", "Alt")

puva_dnds = dndscv(puva_muts, refdb=refcds_38, cv=scores,max_muts_per_gene_per_sample = Inf, max_coding_muts_per_
sample = Inf)

sel_cv_puva <- puva_dnds$sel_cv
sel_cv_puva[sel_cv_puva$gene_name=="TP53",]
```

```
##       gene_name n_syn n_mis n_non n_spl  wmis_cv wnon_cv wspl_cv   pmis_cv
## 17180      TP53     0     2     0     0 10.46806       0       0 0.0182516
##       ptrunc_cv pallsubs_cv   qmis_cv qtrunc_cv qallsubs_cv
## 17180 0.8279757  0.06006747 0.8080215 0.9628665   0.9805273
```

```
sel_cv_indel[sel_cv_indel$gene_name=="TP53",]
```

```
##       gene_name n_syn n_mis n_non n_spl n_ind wmis_cv  wnon_cv  wspl_cv
## 17180      TP53     1    23     5     0     3 8.88856 19.57705 19.57705
##        wind_cv      pmis_cv    ptrunc_cv  pallsubs_cv      pind_cv      qmis_cv
## 17180 20.96863 2.600571e-10 1.531432e-05 5.375256e-12 0.0009622535 2.498628e-06
##        qtrunc_cv  qallsubs_cv pglobal_cv qglobal_cv      pdbs_cv  wdbs_cv n_dbs
## 17180 0.04035876 2.582273e-08          0          0 1.545748e-10 17.31495    13
```